Analyzing the Impact of Acoustic Features on Music Recommendation for Children Across Age Groups

Erkin Basol
Supervisor: Sole Pera, Robin Ungruh
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Erkin Basol
Final project course: CSE3000 Research Project
Thesis committee: Sole Pera, Robin Ungruh, Masoud Mansoury

Abstract

Children are generally underrepresented in music recommender system research, despite having distinct preferences and developmental needs that set them apart from adult audiences. Traditional recommender approaches, designed primarily for adults, often fail to capture the unique listening behaviors of younger users and may fail to serve them effectively. At the same time, acoustic features play a significant role in shaping children's music preferences, yet their potential to enhance and provide optimal recommendations for children remains largely unexplored.

To address this gap, our study examines whether extending a standard collaborative filtering recommender with individual acoustic features can yield more age-appropriate music suggestions for children. We integrate content-based attributes, such as acousticness, danceability, energy, instrumentalness, liveness, loudness, mode, speechiness, tempo, and valence, into an item-based collaborative filtering algorithm and evaluate its performance on users aged 15 through 18. By comparing accuracy and diversity metrics before and after the inclusion of each feature, we identify which acoustic feature improves recommendation quality for each age group and results in the highest performance.

Our findings emphasize the significance of acoustic features, including mode, instrumentalness, and acousticness, in improving performance metrics for distinct age groups. By identifying these age-specific features, our research contributes to the development of age-appropriate and child-centric music recommender systems.

1 Introduction

Music plays a vital role in people's lives, serving purposes such as education, motivation, and entertainment. Unlike other media forms, songs are typically short and often consumed repeatedly or in sessions [1]. These distinct consumption patterns distinguish music from different types of content, creating a unique kind of user interaction. As a result, users tend to listen to familiar tracks while constantly seeking new music that aligns or expands their tastes. In this context, music recommender systems (MRS) are essential tools, as they help users widen and develop their music preferences and discover new content.

Collaborative filtering (CF) is a widely used technique in music recommendation [1], relying on user interaction data to predict preferences. On the other hand, the significance of content-based filtering (CBF), particularly the role of acoustic and descriptive song features, is especially important in music domains, primarily due to the relative sparsity of explicit feedback on music platforms [1]. This limitation makes it more challenging to apply CF effectively on its own [2]. Given these constraints, incorporating content features into CF-based algorithms presents a promising direction [2], as it combines the robustness of collaborative techniques with the contextual information of content-based (CB) approaches to enhance recommendation quality.

Despite advances in recommendation technologies, most existing systems are designed for general audiences [3], particularly adults [4], and often neglect the distinct preferences and consumption patterns of children [4]. Children represent a unique user group with different musical interests [4] and developmental needs such as language acquisition, emotional regulation, and identity formation, where music plays a critical supportive role [5]. Yet the difference in the consumption patterns is rarely considered in the design or evaluation of recommender systems [3, 4]. Moreover, there is limited research on how age influences music preferences [6], especially during childhood. Some studies indicate a strong association between age and genre preferences among children [6, 7], further emphasizing the need to consider age as a critical factor in recommendation models [8]. This shortcoming can result in suboptimal user experiences, where recommendations may fail to meet children's unique tastes or developmental needs. While some studies acknowledge age-related differences in music preference in MRS, they are often not targeted at children. This highlights a

critical research gap in our understanding of how to design MRS that effectively caters to the diverse needs, preferences, and listening behaviors of children.

In this context, acoustic features emerge as a valuable resource for improving music recommendations for children [9]. Prior research highlights the role of acoustic similarity in shaping musical preferences [10]. While children's genre preferences often vary with age [6, 7], both preferences and genres can often be described by acoustic properties such as instrumentation, tempo, and sound textures [2]. These features not only help define genres [2] but also influence how young listeners engage with music [9]. However, relying on genre labels alone can be limiting, as genre boundaries are often ambiguous and inconsistent [11]. In contrast, acoustic features provide more consistent and structured information for recommendations, which is vital given the diverse and evolving listening habits of children. This highlights a crucial gap in current research as MRS frequently overlooks age-specific preferences, especially those of children. Additionally, acoustic features can identify specific musical characteristics that are particularly relevant to younger audiences. This presents an opportunity to enhance the quality of recommendations by designing systems that are more tailored to meet these needs. However, there is currently a lack of research on the impact of individual acoustic features on the performance of MRS for children.

This gap highlights the need for research on how the incorporation of song acoustic features influences the performance of recommender systems across different age groups. To address this, we pose the following research question: **"To what degree can the incorporation of different acoustic features improve the performance of a CF-based recommender system for children?"**

To investigate this question, we examine how incorporating individual content features into collaborative filtering can enhance music recommendations for children. Using the LFM-2b dataset [3] for age-related user interactions and the LFM-BeyMS dataset [12] for acoustic song features, we focus on age groups 15 to 18. Our analysis evaluates the impact of various content features, such as energy and mode, on recommendation performance when used as an extension to CF, considering both accuracy and diversity metrics. We aim to identify which feature, when integrated into the system, yields the best results for each age group. Additionally, we explore whether performance trends remain consistent or vary significantly across these groups. To support reproducibility, all code and materials will be published in a public repository.[1]

The outcomes of this work provide empirical evidence on the performance of acoustic content features in age-specific MRS, revealing the acoustic features that are most relevant for different age groups in MRS design. This research extends the CF approaches by identifying the features that can improve music recommenders for children.

2 Related Work

In this section, we review the related literature on children's music perception, followed by an overview of MRS, with a focus on CBF and CF algorithms and how incorporating content features can benefit CF-based algorithms. Finally, we discuss acoustic features and their relationship to musical preferences.

2.1 Children in Music Recommender Systems

Music plays a significant role in children's lives, often more so than for adults, serving both as entertainment and a means of socialization and skill development [9]. However, most MRS are designed for general or adult audiences, often overlooking the specific needs of younger users [3, 4].

[1] https://github.com/Protestak/RP_Project

Consequently, the lack of focus on children limits the effectiveness of the MRS algorithms even though they aim to enhance user satisfaction.

Understanding children's listening behavior is key to developing more inclusive systems. Studies identify diverse preference clusters among children based on audio features like energy, danceability, acousticness, and instrumentalness [4, 13]. While these findings show the variation in children's musical preferences, their implications for improving recommender systems remain unclear. This suggests a need to explore how the diversification in audio features can be leveraged to design systems that better serve young listeners.

Adolescence is a formative period for shaping musical preferences [7]; therefore, making accurate recommendations for children is especially important. Research using the LFM-1B dataset shows genre preferences vary widely across child age groups, emphasizing the need for age-aware recommender systems [6]. Furthermore, [14] suggests that users in adolescence and young adulthood exhibit the greatest variability in music listening behavior, whereas this variation declines significantly as listeners reach middle adulthood [7, 14]. Consequently, designing an MRS that takes into account the developmental stages is important.

2.2 Music Recommender Systems

Music recommendation systems employ various algorithms to provide personalized suggestions, with CF being one of the most common approaches [1]. In music, CB features like acoustic descriptors are more influential than in other domains [1]. However, many systems overlook contextual and personal factors that can be identified by CB features [15].

CF recommends items based on the preferences of similar users, but it can suffer from cold start issues when user data is limited [16]. On the other hand, CBF, which relies on item characteristics, can help mitigate these problems but may struggle to capture highly diverse user preferences. While CF generally outperforms CB filtering in tasks like playlist generation [17], CF systems that integrate acoustic features from CB methods can effectively complement CF [2].

Research indicates that CF systems trained on the LFM dataset have better performance for children than adults [6], suggesting that CF may be especially beneficial for users under 18. Consequently, combining CF with CB features such as acoustic descriptors could further improve recommendations for younger listeners. Furthermore, hybrid systems that integrate CF and CB methods offer a promising solution by utilizing their complementary strengths [16, 18]. However, most existing systems still overlook the unique needs of children, highlighting the need for age-aware MRS design.

2.3 The Role of Acoustic Features in Music Recommendation

Acoustic features are used in music classification and recommendation. Studies have demonstrated that genres can be effectively predicted using machine learning models trained on acoustic features such as acousticness, energy, speechiness, loudness, valence, instrumentalness, liveness, tempo, and danceability [19, 20]. For example, [20] identified which acoustic features are most predictive for genre classification tasks. Additionally, [21] showed that features like speechiness, danceability, and loudness significantly influence secondary genre preferences. These findings suggest a connection between acoustic features and user-specific music preferences, which can be incorporated into MRS. A related study [9] trained a recommender system using a complete set of acoustic features for children aged 6 to 17. Although their model outperformed the baseline, their results were not significant, suggesting that acoustic features can be promising but require further investigation. .

3 Experimental Setup

We adopt an offline evaluation approach to assess the performance of our recommender systems, which generate top-10 music recommendations for the age groups 15, 16, 17, and 18. Our choice of offline experimentation is driven by ethical constraints associated with conducting live experiments on children and collecting real-time user data. As a result, we rely on historical interaction data for model training and evaluation.

3.1 Data

Our research utilizes two **datasets**. The first is the publicly available **LFM-BeyMS** [12] dataset, which comprises 4,148 users and 1,084,922 tracks—each with acoustic features—along with 110,898 artists and 16,687,363 listening events. We use the LFM-BeyMS dataset due to its public availability and its inclusion of acoustic music features that align with the objectives of our study. The dataset provides ten key acoustic features commonly used in music information retrieval research [9, 21, 22]. Table 1 presents the audio features and their short descriptions.

Feature	**Description**	**Range**
Mode	Binary (1 = major, 0 = minor)	$\{0, 1\}$
Acousticness	Probability a track is acoustic	$[0, 1]$
Danceability	Suitability for dancing	$[0, 1]$
Energy	Perceived intensity of a track	$[0, 1]$
Instrumentalness	Likelihood of instrumental-only composition	$[0, 1]$
Liveness	Probability of live performance	$[0, 1]$
Loudness	Loudness in decibels (dB)	$[-60, 0]$
Speechiness	Presence of spoken words	$[0, 1]$
Tempo	Speed in beats per minute	$[0, 294]$
Valence	Emotional tone or mood of the track	$[0, 1]$

Table 1: Description of Audio Features

The second dataset used is **LFM-2B** [3], which contains 2,014,164,872 listening events from 120,322 users across 50,813,373 songs, collected between February 2005 and March 2020 [3]. Each listening event in this dataset is tagged with the user's age at the time of listening, enabling the formation of age-based groupings. This dataset is chosen as it represents the most extensive and up-to-date resource in the domain [23], and it includes age labels, which are central to the aims of this research.

To maintain consistent input for our extended recommender system, we eliminate all listening interactions involving tracks that lack any required acoustic features. This filtering step ensures that only songs with complete content information are used, enabling balanced comparisons across different models that rely on various feature types. To standardize the interaction data, we binarize listening events by considering each user interaction with a track as a positive interaction. Since the dataset does not contain explicit ratings, we employ an implicit feedback approach. In this method, the presence of an interaction is regarded as a positive signal, while the absence of interaction is viewed as negative. Additionally, we exclude users under **12** due to social media policies [3], and those over **18** to maintain a focus on children. As a result, we retain only the listening interactions of users aged 12 to 18 that are associated with tracks containing full content feature information.

To ensure the generalizability of the dataset, we limit our analysis to interactions from **2012**, the year with the highest number of retained listening events after applying the filtering steps described above. Focusing on a single calendar year allows us to preserve stable user age information

and avoids shifts in musical preferences over time. This decision supports the ecological validity of our research by aligning data selection with realistic listening patterns [24]. We then apply a global temporal split to divide the dataset into *training (January–May), validation (June–July), and test (August–October) sets*, addressing temporal inconsistencies present in random or order-aware splits [24]. This split is chosen to maximize interaction volume in the training set while ensuring meaningful overlap across all periods. We exclude *November* and **December** due to low user activity and minimal user overlap with previous months. To further improve consistency, we only retain users who are active in all three periods. This ensures that the model can be trained, validated, and tested on the same users, improving evaluation quality.

We apply a user-level k-filtering strategy to ensure that each user contributes sufficient data for effective model training and evaluation. Specifically, we retain users with more than ***25*** *interactions* in the *training set* and more than ***7*** *interactions* in both the *validation* and *test sets*. We select these thresholds to retain the majority of listening events and maximize user inclusion. In particular, we observe that the number of user interactions drops significantly after **25**, providing a trade-off between sparsity and retained user interactions. We also chose the value of **7** interactions for the validation and test sets to represent approximately **20%** of the total interactions. After applying k-core filtering, the average number of interactions per age group in the validation and test sets ranges from **10 to 11** interactions. This is illustrated in Table 2, which aligns well with our *top-10* recommendation setup and supports a balanced performance evaluation. While previous work recommends a **70%/30%** train/evaluation split [24], we adopt a **65%/17.5%/17.5%** split. This provides slightly more training data while having enough interactions in the evaluation phases to make top-10 recommendations feasible and meaningful.

Following user-level filtering, we apply a second round of k-filtering at the song level. We retain only tracks that have at least **10** total interactions across the full dataset, at least **5** in the training set, and at least **1** interaction in both the validation and test sets. This helps reduce the sparsity of the interaction matrix and ensures that all retained songs are relevant in each data split.

Lastly, to ensure consistent and standardized comparisons, we balance the number of users in each age group. We achieve this by *randomly downsampling* the age groups with the most users to match the size of the *smallest group*. Specifically, we focus on the **15-, 16-, 17-, and 18-year-old group*s***, as they have sufficient users meeting the above criteria. Users aged **12 to 14** are excluded due to insufficient user data. This final step ensures that no age group dominates the results due to size, allowing for more reliable analysis of age-based patterns in music recommendation.

Age	Test Set			Validation Set			Training Set		
	Total Int.	**Avg. Int.**	**Users**	**Total Int.**	**Avg. Int.**	**Users**	**Total Int.**	**Avg. Int.**	**Users**
15	3219	11.54	279	3158	11.32	279	10958	39.28	279
16	3233	11.59	279	2854	10.23	279	10642	38.14	279
17	3190	11.43	279	2958	10.60	279	10850	38.89	279
18	3321	11.90	279	2753	9.87	279	10660	38.21	279
Total	**12963**	**11.62**	**1116**	**11723**	**10.50**	**1116**	**43110**	**38.63**	**1116**

Table 2: Interaction Data by Age Group

Table 2 shows the distribution of user interactions across the training, validation, and test sets for each age group. Each set contains **279** users per age group, totaling **1116** users overall. The *training set* includes approximately **10,600- 11,000** interactions per age group, with an average of around **38** interactions per user. The *validation set* contains between **2,700 and 3,100** interactions per age group, averaging about **10 to 11** interactions per user. Similarly, the *test set* has roughly **3,200** interactions per age group, with average interactions per user ranging from **11.4 to 11.9**. This shows that the interaction counts are relatively consistent and balanced across all age groups.

3.2 Algorithms and Recommender System Design

We use the *aiolli* [2] implementation of the *item-based k-Nearest Neighbors (item-kNN) as the baseline CF algorithm*, training it on the *training set* and tuning it using the *validation set*. ***Item-kNN*** is widely recognized in the literature as a robust and commonly used *memory-based method*. A comparative study [25] also showed its strong performance on the **LFM dataset**, reinforcing its suitability as a reliable baseline for our evaluation.

To evaluate the contribution of individual acoustic features in music recommendation, we *extend* the traditional item-based k-Nearest Neighbors (item-kNN) algorithm by incorporating acoustic features. In doing so, we aim to enhance the performance of the widely used item-kNN algorithm. In its original form, item-kNN operates on a *user-item interaction matrix* with *binary values*, where a value of **1** indicates that a user has interacted with an item, and **0** indicates no interaction. Similarity between items is then computed based on item interaction data using a similarity metric.

User \ Item	**Song A**	**Song B**	**Song C**
Standard ItemKNN	1 (listened)	0 (not listened)	1 (listened)
Hybrid (Danceability)	0.7 (danceability)	0 (no interaction)	0.4 (danceability)

Table 3: Comparison of User-Item Interaction Matrix Representations

In our approach, we **replace the binary values in the user-item interaction matrix** with corresponding **acoustic feature values**. Traditionally, in the item-kNN algorithm, a value of **1** is assigned to an entry (u,i) if a user u listened to a track i, and **0** otherwise. However, we enhance the item-kNN algorithm by substituting the positive interactions with the respective acoustic feature values while keeping the negative interactions unchanged.

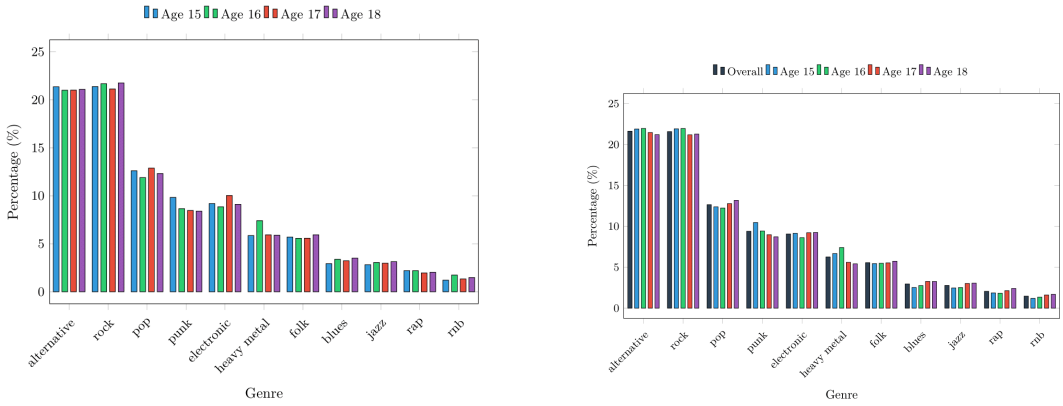
Specifically, for each positive interaction in the training set, when a user listens to a song, we replace the binary value of 1 with the content value of a selected *acoustic feature* associated with the item. For example, if a user listened to a track with a *danceability* score of 0.4, the corresponding matrix entry becomes 0.4 instead of 1. Negative or missing interactions remain **0**, as in the original binary interaction matrix. This modification allows the model to capture more nuanced user preferences by leveraging item-level content features, as illustrated in Table 3.

This modification is motivated by the need to capture *content similarity* between items, not just co-occurrence frequency. While standard item-kNN relies solely on co-listening behavior, this combination enables the model to account for the *musical characteristics* users are likely responding to. This is particularly useful in music recommendation, where audio content (e.g., tempo, mood, or danceability) often aligns more closely with user taste than pure interaction frequency. We employ *cosine similarity* to normalize these feature-weighted interactions to ensure that differences in feature scale do not dominate the similarity computation.

3.3 Acoustic Features and Genre Distribution

After preprocessing, **4,700** songs remained for analysis. To ensure reliability, we examine the distributions of acoustic features and genres. Although genre is not directly used, it is included to verify that the preprocessing preserved the natural music catalog. As shown in Figure 1, the genre distribution closely matches previous LFM-based studies [6], indicating that the dataset's diversity and balance are maintained. Similarly, the distribution of acoustic features aligns with patterns reported in prior work [4], supporting the applicability of existing insights to our context.

[2] https://github.com/sisinflab/elliot/blob/master/elliot/recommender/knn/item_knn/aiolli_ferrari.py

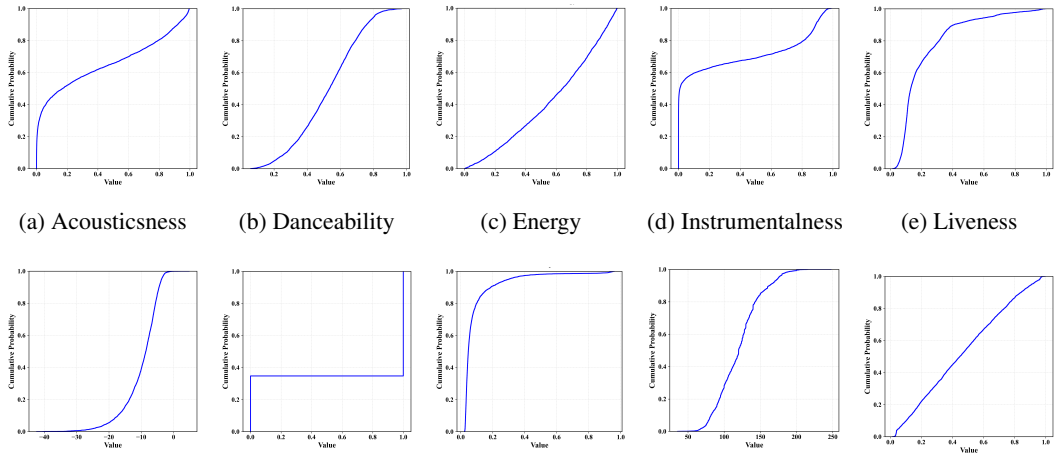


(a) Genre Distribution of the Test Age Groups

(b) Genre Distribution of the Training Age Groups

Figure 1: Genre Distributions of Test and Training Age Groups

This analysis serves two main purposes: first, *to verify that our preprocessing steps did not introduce any imbalances* that could distort the recommendation process; and second, *to gain insight into how acoustic features are distributed across the dataset*. This is particularly important because tightly clustered or skewed features may lead to high performance, as the model could rely on feature similarity rather than capturing meaningful user preferences. To identify such cases, we use **ECDF plots** (Figure 2) to visualize feature distributions. If two features have similar distributions but lead to different model outcomes, this may highlight the relative importance of one feature over another. Conversely, if *features* with similar distributions yield similar performance, it could indicate redundancy or the potential for feature grouping. In both cases, analyzing these patterns can help us better understand the role of individual features in the recommendation process and inform future model design choices.



(a) Acousticsness (b) Danceability (c) Energy (d) Instrumentalness (e) Liveness

(f) Loudness (g) Mode (h) Speechiness (i) Tempo (j) Valence

Figure 2: ECDF Plots of Audio Features

For most of the acoustic feature values, the ECDF plots indicate a near-linear distribution, suggesting that these features are approximately uniformly distributed across the dataset. However, for *liveness*, *acousticness*, and *speechiness*, there is a high concentration of values in the **0–0.2 range**, indicating that most songs exhibit low levels of these characteristics. Additionally, the distribution of the *mode* feature shows a notable inclination toward *major mode* songs. This presents, overall, a potential imbalance in the dataset toward more mainstream musical attributes, which could influence the generalizability.

3.4 Evaluation

The primary goal of recommender systems, both in research and industry, is to enhance user satisfaction [23]. As a result, evaluation metrics are typically focused on ranking and accuracy. However, when designing systems for children, additional care is necessary, as they have unique developmental needs and are more susceptible to the effects of digital interactions.

According to a literature survey [25], **Hit Rate (HR)***,* **Mean Reciprocal Rank (MRR)***, and* **Normalized Discounted Cumulative Gain (NDCG)** are among the most commonly used evaluation metrics. Each serves a distinct purpose: MRR measures how early the first relevant item appears in the ranked list, reflecting the system's ability to present useful content quickly. HR, on the other hand, assesses whether at least one relevant item is included in the top-K recommendations, which is crucial to ensure the recommender provides some value to the user. NDCG evaluates the ranking quality by taking the positions of all relevant items into account, giving higher importance to those appearing higher in the list, and thus rewards well-ordered recommendation lists.

In the context of child-oriented systems, HR is crucial to ensure at least one relevant recommendation, while MRR assesses how well a system ranks the most appropriate items. Additionally, beyond accuracy-focused metrics, diversity and coverage should also be considered to promote a richer and more engaging user experience [26]. **Intra-list diversity** refers to the similarity among songs recommended in the top-N list, where we use all ten content features to measure it. Moreover, **coverage** is defined as the percentage of the catalog recommended to users in the top-N recommendations.

3.5 Experiments

The goal of our experiments is to assess the contribution of individual acoustic features to the CF-based recommender system, specifically for users aged **15 to 18 years old**. To achieve this, *we extend the item-kNN algorithm* by modifying the interaction matrix to incorporate the values of one acoustic feature at a time, rather than combining multiple features. This setup allows us to **compare the performance of these feature-specific extended recommenders across age groups** and **identify which features most enhance recommendation quality and diversity**, offering insights for designing systems tailored to young users.

We conduct offline *top-10* recommendation experiments, evaluating each model on age-specific test sets using the historical interactions of **279** users per group as ground truth. In total, we train **11** models: a baseline using binary interactions and **10** extended models, each utilizing a single acoustic feature, substituting the binary values in the baseline algorithm. Each model is evaluated independently for each age group.

Given that the test set includes approximately **12** items per user, we focus on the *top-10* recommendations to ensure evaluations remain meaningful and metrics are reliable. *Grid search* is used to conduct hyperparameter tuning for optimal configurations. *Hyperparameters* are adjusted with a validation set and applied consistently across all models to ensure comparable results.

To maintain internal validity [24], we hold control variables constant and vary only the independent variable of the acoustic feature. Multiple experimental runs ensure repeatability, while statistical testing addresses generalizability.

To evaluate whether each extended model significantly differs from the baseline, we use the *Wilcoxon signed-rank test*, following the methodology in [27]. This non-parametric test is chosen because it does not assume normality, as our *Shapiro-Wilk* scores indicate **0.0000** for each feature in MRR and NDCG metrics visible in Table 9 in the appendix. We apply a two-tailed hypothesis with a significance threshold of $\boldsymbol{p < 0.05}$. Paired comparisons are conducted using MRR and NDCG for each of the **279** users in each age group, comparing the baseline model to its content-extended version. For HR, we use *McNemar's test*, which is appropriate for paired binary outcomes. This test has been used in CF and list-based recommendation studies [27, 28, 29] and statistical significance is assessed at $\boldsymbol{p < 0.05}$.

4 Results

We present the results separately for each age group (15, 16, 17, and 18) in Table 4. Detailed p-values for the accuracy metrics are provided in the appendix in the Tables 5, 6, 7, and 8 for reference. Following the individual results, we examine patterns across age groups to highlight common trends and notable differences, enabling both within-group and cross-group analysis. For significance, all the extended models are compared to the baseline model, and the importance refers to that throughout the section.

Recommender Algorithm	**Age 15**					**Age 16**				
	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10
Item-KNN + Acousticness	0.1326	**0.0402**	0.0200	0.3448	0.9950	0.1183	0.0306	0.0190	0.3567	0.9949
Item-KNN + Danceability	0.1362	0.0372	0.0194	0.3501	0.9951	0.1147	0.0314	0.0183	0.3574	0.9950
Item-KNN + Energy	0.1326	0.0366	0.0188	0.3510	0.9952	0.1147	0.0325	0.0183	0.3576	0.9950
Item-KNN + Instrumentalness	0.1577	0.0416	**0.0226**	0.3280	0.9945	0.1183	0.0382	0.0180	0.3202	0.9946
Item-KNN + Liveness	0.1326	0.0360	0.0185	0.3478	0.9952	0.1147	0.0322	0.0185	0.3569	0.9951
Item-KNN + Loudness	0.1470	**0.0390**	**0.0200**	0.3503	0.9953	0.1183	0.0318	0.0180	0.3531	0.9951
Item-KNN + Mode	0.1613	0.0418	0.0219	0.3076	0.9952	0.1039	0.0324	0.0162	0.3127	0.9947
Item-KNN + Speechiness	0.1326	0.0353	0.0185	0.3484	0.9951	0.1147	0.0319	0.0182	0.3552	0.9951
Item-KNN + Tempo	0.1398	0.0369	0.0195	0.3486	0.9953	0.1147	0.0320	0.0188	0.3561	0.9950
Item-KNN + Valence	0.1326	0.0368	0.0189	0.3493	0.9952	0.1147	0.0318	0.0184	0.3578	0.9951
Item-KNN (Baseline)	0.1326	0.0361	0.0187	0.3503	0.9952	0.1147	0.0320	0.0182	0.3569	0.9950
Recommender Algorithm	**Age 17**					**Age 18**				
	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10
Item-KNN + Acousticness	0.1039	0.0203	**0.0116**	0.3773	0.9950	0.1219	0.0315	0.0168	0.3695	0.9953
Item-KNN + Danceability	0.1039	**0.0192**	0.0106	0.3739	0.9952	0.1111	0.0306	0.0169	0.3714	0.9954
Item-KNN + Energy	0.1004	0.0184	0.0101	0.3737	0.9952	0.1075	0.0306	0.0163	0.3714	0.9955
Item-KNN + Instrumentalness	0.1111	0.0276	0.0123	0.3416	0.9946	0.1183	0.0315	0.0167	0.3446	0.9952
Item-KNN + Liveness	0.0968	0.0182	0.0099	0.3759	0.9951	0.1183	0.0313	0.0172	0.3688	0.9955
Item-KNN + Loudness	0.1004	0.0187	0.0102	0.3782	0.9952	0.1183	0.0290	0.0174	0.3701	0.9955
Item-KNN + Mode	**0.1434**	**0.0390**	**0.0169**	0.3208	0.9951	0.1398	0.0387	0.0196	0.3263	0.9954
Item-KNN + Speechiness	0.1039	0.0190	0.0106	0.3741	0.9952	0.1219	0.0292	0.0172	0.3722	0.9955
Item-KNN + Tempo	0.0932	0.0178	0.0096	0.3724	0.9952	0.1147	0.0286	0.0169	0.3697	0.9955
Item-KNN + Valence	0.1004	0.0187	0.0102	0.3735	0.9952	0.1111	0.0280	0.0158	0.3731	0.9955
Item-KNN (Baseline)	0.0968	0.0181	0.0099	0.3756	0.9951	0.1075	0.0302	0.0163	0.3714	0.9955

Table 4: Combined Performance of Recommender Algorithms Across Ages 15 to 18. Metrics include HitRate@10, MRR@10, NDCG@10, Coverage@10, and Diversity@10. Significant improvements over the baseline are bolded.

Among all acoustic features, **mode** emerges as the most robust and consistent attribute across multiple age groups, particularly for users aged **15, 17, and 18**. At age **15**, the recommender extended with mode achieves the highest performance across all ranking metrics. Similarly, for age **17**, the inclusion of **mode** yields substantial and statistically significant improvements over the baseline in HR, MRR, and NDCG. Age 18 also benefits from mode, showing improved recommendation quality, although without statistical significance, similar to the age group 15. These patterns suggest that mode, as a musical attribute, plays a crucial role in enhancing the quality of recommendations for older teens, particularly in terms of ranking precision and top-item placement.

However, this trend does not hold uniformly across all age groups. For **age 16**, mode performs the worst compared to all features and baseline, showing reduced effectiveness across ranking metrics. This contrast indicates that mode, while generally a strong contributor to recommendation performance, may have an age-dependent effect and might not align with the preferences of specific age groups. It is also important to note that no feature performed significantly better than the baseline for the age 16 group, suggesting that the results for this group may lack descriptive power or reveal less distinct patterns in feature effectiveness.

Beyond mode, **instrumentalness** also shows notable performance improvements across various age groups to a lesser extent. Instrumentalness demonstrates a significant increase in performance for users aged **15** and a non-significant improvement for users aged **17**. For users aged **16** and **18**, the improvements are near baseline, with the MRR being highest for the **16-year-old group**, although this difference is not statistically significant. These findings suggest the potential benefit of incorporating instrumentalness into recommender systems targeted at children as a broader group, rather than focusing solely on age-specific significance. However, further investigation is necessary to draw definitive conclusions.

Moreover, **acousticness** emerges as a prominent feature that performs well for users **aged 15 and 17.** Although the 16-year-old age group does not show a clear improvement in recommendations based on this feature, its performance remains stable. For users aged 18, the recommender extended with acousticness slightly improves recommendation quality. These patterns suggest that acousticness, as a song attribute, can generally contribute positively to enhancing recommendation systems, particularly for younger teen users. While the improvement in **acousticness** is significant in the MRR and NDCG metrics for the 15- and 17-year-old age groups, the HR shows only a slight increase. This suggests that, even though there is no strong evidence of an increase in new hits, the improvement in ranking metrics over the baseline indicates that for users to whom acousticness mattered, this extension helped rank the first relevant item higher and improved the overall ranking quality. This further suggests that users who enjoy acoustic songs or have previously interacted with acoustic content may benefit more from recommendations based on this feature during these ages.

Furthermore, **acousticness** and **instrumentalness** showed statistical significance in different quality metrics: acousticness in MRR and instrumentalness in NDCG for users aged 15. Both features showed relatively similar performance improvements in other metrics, indicating consistent performance across evaluation criteria. This consistency might be due to their similar value distributions, as illustrated in Table 2. Given their comparable distributions and complementary performance in ranking metrics, it may be reasonable to group these features for further exploration. For the other age groups, this pattern was not observed, suggesting that a unique and valuable relationship may exist specifically for age 15, indicating a promising direction.

It is worth noting that **loudness** also demonstrated significance in MRR and NDCG, further supporting the idea that it may be a particularly suitable feature for users in the age group of 15. Although the generalizability of these findings has not yet been established, the current results indicate that **loudness** can contribute meaningfully to recommendation quality for the age group of 15. The relevance of the loudness could be attributed to younger children listening to more energetic

and upbeat music [4]. This suggests that this feature stands out for the 15-year-old age group, with potential for more performative recommendations.

These findings demonstrate that CB features, such as *mode, instrumentalness, acousticness, and loudness*, can enhance recommendations, particularly when aligned with user demographics. However, the weak performance of the **mode** for age 16 underscores the importance of tailoring features to age-specific preferences. **Coverage** trends reveal that features such as loudness and acousticness broaden item reach across age groups, while mode and instrumentalness reduce it. This suggests a trade-off between ranking accuracy and diversity of item coverage. **The intra-list diversity** remains high across all features and age groups, driven by the itemKNN algorithm. This suggests that incorporating content features does not introduce additional diversity.

5 Discussion

Across our experimental analysis, the features that established statistical significance are ***acousticness***, ***instrumentalness***, and ***loudness*** for the age group **15**; ***acousticness***, ***danceability***, and ***mode*** for the age group **17**. Among these, ***mode*** emerges as the most impactful feature, consistently contributing to improved performance across all age groups, except for age **16**, where it shows the lowest effect. Furthermore, the age groups 16 and 18 have no hybrid recommenders that establish significance.

The prominence of **mode** as a key feature in our analysis can be attributed to its strong influence on emotional responses to music, which likely enhances the effectiveness of personalized recommendations. Mode distinguishes between major and minor scales, with major modes evoking happy, uplifting emotions and minor modes conveying sadder, more melancholic tones [30]. Research highlights mode as a critical factor in shaping listeners' emotional reactions, often more influential than other musical elements [30]. This emotional resonance explains the mode's significant contribution to recommendation performance for age groups 15 and 17, where preferences may align with distinct emotional profiles. The lower impact observed in the age group 16 could reflect unique musical tastes or a reduced sensitivity to mode-driven emotional responses, possibly due to developmental or contextual factors.

These findings highlight the significant role of mood in tailoring music recommendations to individual emotional preferences across most age groups of children. The mode values of the age groups were further analyzed based on their listening events in the test and training sets, leading to the results shown in Figure 3.

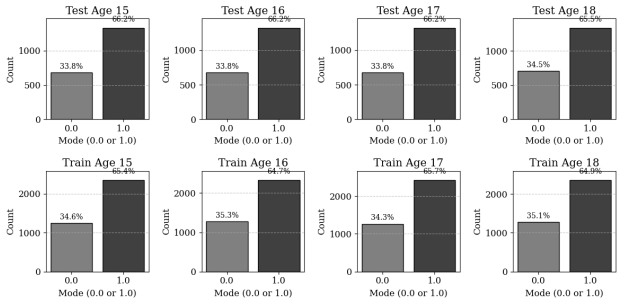


Figure 3: Mode Value Distribution of Test and Training Sets per Age Group

There appears to be no prominent difference in the distribution of mode values, reducing concerns about over-representation or under-representation of these values across age groups, which could impact the performance of the recommender system. This suggests that mode values can be effectively utilized for age-centric music recommendations, particularly for the 15 and 17-year-old age groups.

An important point to highlight, as explained in [4], is that as users mature (from age 15 to 17), the cluster of users who deviate from mainstream trends becomes more prominent. This trend is also visible in our study: we observe a decreasing performance of the baseline algorithm concerning age, with age 15 achieving the best results across metrics such as HR, MRR, and NDCG, followed by ages 16, 18, and 17. This suggests that as adolescents mature, they develop more individualized musical preferences, leading to increased diversity in their listening habits. As a result, recommender algorithms perform relatively worse, particularly those based on CF, which rely on shared patterns across users. We can make this claim because we balance the age groups in terms of song interactions in the training, validation, and test sets, and each group is represented approximately equally. The stronger performance among younger users may be attributed to the fact that they tend to follow more common trends. This aligns with findings from developmental psychology, which suggest that music often serves as a tool for social bonding and peer interaction during adolescence [6].

This study also extends the research by [9], which showed that including audio features improved MRR, HR, and NDCG metrics in top-10 recommendation systems for high school students, despite the lack of statistical evidence. Unlike [9], which combines all audio features, our work analyzes each feature separately to gain more precise insights. Our findings suggest that specific audio features, such as instrumentalness, loudness, and acousticness for age 15, as well as danceability and particularly mode for age 17, can be practically valuable in refining music recommendation strategies for adolescent users. These insights can inform the design of more targeted, age-aware recommender systems for children. More broadly, this research encourages a shift away from general recommender systems toward more personalized, age-aware recommenders that utilize these acoustic features that reflect the emotional and social roles music plays in adolescent life.

Additionally, our research can be extended in a backward-looking manner. Due to the age group balancing in our dataset, we were unable to assess the performance of the recommender system for younger users, specifically those aged 12, 13, and 14. Previous studies [4, 9] on the LFM-2B dataset analyze children's musical preferences grouping by their educational stages—Grade School (GS 6–11), Middle School (MS 12–14), and High School (HS 15–17). Our research focuses only on age groups corresponding to the HS segment.

Having identified performance patterns for HS users aged 15–17 (with 18 as an exception), we can potentially extrapolate insights about recommender system behavior in GS and MS users. This can be done by drawing on earlier findings about age- or education-level-related similarities and differences in music listening behavior. Therefore, while we do not directly evaluate younger age groups, our analysis contributes to understanding how recommender systems might perform for them, given observed developmental and behavioral trends.

To contextualize what other research has done in terms of comparing educational levels, [9] indicates that GS students exhibit distinctly different listening patterns compared to both MS and HS students, particularly across all measured audio features. Moreover, it states that the musical preferences of MS and HS users are much more aligned. This suggests a closer similarity in audio feature preferences between MS and HS students, with GS users standing out as a more unique group in terms of musical behavior. Based on these findings, we might expect features such as *acousticness*, *loudness*, *mode*, and *instrumentalness*—which performed well for HS users—to also be effective for MS users, due to their comparable listening behavior. This implies that future research could explore applying the features identified here to recommender systems targeting the MS group.

In contrast, given the significant differences observed in GS users' preferences, it is difficult to draw clear conclusions about how recommenders trained on specific features might perform for this group. Furthermore, [4] found that GS users tend to prefer music that is more "happy" or "upbeat" than that preferred by MS and HS students, which aligns with the idea that this age group is generally more energetic and expressive. From this, we can suggest that features such as *energy*, *tempo*, *danceability*, *valence*, or the use of *major* modes may be more appropriate for recommender systems designed for GS users.

One study [10] found that acoustic similarity plays a significant role in both playlist construction and music recommendation, using both quantitative and qualitative analysis. The researchers also examined whether the perceived similarity was related to the acoustic content features used in our study. Although they found no statistically significant evidence for acoustic music similarity, they identified features such as valence, speechiness, tempo, liveness, and energy as having p-values close to significance. While this study did not focus on children or assess recommender algorithms, its findings offer a useful perspective for interpreting our results.

This insight may be particularly relevant when considering children, whose music preferences and listening behaviors are still developing. It raises the question of whether acoustic similarity holds the same importance for younger users as it does for adults. Furthermore, it challenges the generalizability of such findings, as listeners of different ages may interpret acoustic similarity in distinct ways. It also highlights the importance of considering the characteristics of the data samples used to analyze recommenders and listening events, which is particularly critical when establishing statistical evidence. These factors further support the idea that music preferences and behaviors are highly diverse across individuals.

Despite improved ranking metrics and HR, the recommender extended with mode as a feature has the lowest coverage across all age groups, which recommends fewer unique songs. Instrumentalness, although beneficial for performance, also led to lower coverage. In contrast, features like acousticness and loudness improved performance without compromising coverage, suggesting they are more suitable for balanced recommenders. For younger users, who are often exploring their musical preferences, this reduced coverage could limit exposure to a broad range of songs, potentially decreasing engagement with the system despite improved performance. Additionally, mode and instrumentalness tend to create mainstreamness due to lower coverage, thus limiting users' opportunities to discover diverse genres and songs.

Finally, diversity scores indicate low intra-list diversity for all recommenders. This can be explained by the fact that the item-KNN algorithm serves as the foundation for the extended recommender, resulting in recommendations that are highly similar to one another.

6 Responsible Research

Studying children in real-world settings presents significant ethical challenges, particularly regarding privacy, consent, and data handling. To address these concerns and avoid the need for formal university ethics approval for online evaluation, our research exclusively relies on historical datasets that do not contain any personally identifiable information. Specifically, our study utilizes the publicly available **LFM-BeyMS** [12] dataset and the **LFM-2B** [3] dataset, which is not publicly distributed at the moment but has been used in many previous research studies.

The decision to use these pre-existing datasets ensures that no direct interaction with minors or data collection from children was required. This approach supports ethical standards in child-centered research while still enabling meaningful analysis of listening behavior across different age groups.

This study provides valuable insights into age-specific recommendation patterns and has practical implications beyond academic research. For instance, music producers and digital platforms could leverage these findings to design content and recommendation systems that are better aligned with the preferences and emotional needs of specific age groups. This not only enhances user engagement but also supports the development of safer and more age-appropriate digital experiences. However, if misused, such insights could potentially be exploited to steer users in undesirable directions or expose them to inappropriate content.

In terms of reproducibility, all code developed during this study will be made publicly available through a dedicated GitHub repository. A comprehensive README file outlining the necessary steps to replicate the full pipeline, including preprocessing and evaluation, is also published, ensuring transparency and encouraging future work. We also disclose all hyperparameter choices for our item-KNN model, including neighborhood size, similarity measure, shrinkage, and normalization method. We also follow in the reproducibility best practices outlined in [31] in our README file to ensure that our results can be reliably reproduced.

7 Conclusions, Limitations, and Future Work

In this work, we investigated how the incorporation of music acoustic features influences the performance of a CF-based recommender for children across different age groups. By isolating acoustic features and analyzing their effects individually, we uncovered age-specific patterns in recommender performance. Our results show that *mode*, among all features, consistently contributed to improved performance, especially for ages 15 and 17. Importantly, we also observed that while certain features enhanced ranking metrics, they came at the cost of coverage, raising concerns about reinforcing mainstream consumption.

However, these findings have some limitations and require further exploration in the future. Our preprocessed dataset comprises only 279 users from each age group, achieved through random down-sampling, treating age as a controlled variable. However, this method may have excluded users who are less responsive to certain features. Additionally, for the further generalizability of our findings, more users are required. Another limitation is the dataset's age, collected in 2012. Over the past 13 years, music culture and listening habits have evolved significantly due to social media, diverse listening contexts, and accessible streaming platforms like Spotify and YouTube [1]. These shifts suggest that today's youth preferences likely differ from those in the dataset, meaning our study's findings may not fully reflect current music consumption trends. A further limitation involves the use of implicit data. Because we binarized listening events, it is unclear whether users fully listened to the songs, skipped them after a few seconds, or actively selected the tracks versus receiving them through recommendations [1]. This limits our ability to interpret user intent.

Finally, a limitation of our algorithm arises from replacing binary interaction values with acoustic features in recommenders extended with the mode feature. Specifically, incorporating the mode feature as 0 for minor-mode tracks causes these tracks to resemble non-interacted items in similarity calculations after substitution. This results in zero similarity scores during cosine comparisons, reducing the recommendation frequency of minor-mode tracks. Consequently, the system favors major-mode content, and users who prefer minor-mode music may receive recommendations that do not align with their tastes. This issue is likely the primary reason for the low coverage of the recommender when incorporating the mode feature.

Building on these insights, future work could explore how the identified acoustic features, such as *mode*, *instrumentalness*, and *loudness*, can be integrated into real-world recommendation systems to assess their practical impact. This would involve deploying the system in a live environment and collecting user interaction data to evaluate the effectiveness, user satisfaction, and generalizability

of the feature-based models. Additionally, a similar study could be conducted using a new dataset or by collecting data directly from children to assess the recommender's performance based on these features. Finally, zero feature values indicating no interaction, such as in the case of a minor mode, could be mitigated with improved algorithm design.

References

[1] M. Schedl, P. Knees, B. McFee, and D. Bogdanov, *Music Recommendation Systems: Techniques, Use Cases, and Challenges*. New York, NY: Springer US, 2022, pp. 927–971. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_24

[2] B. Shao, D. Wang, T. Li, and M. Ogihara, "Music recommendation based on acoustic features and user access patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1602–1611, 2009.

[3] R. Ungruh, A. Bellogín, and M. S. Pera, "The impact ofÂ mainstream-driven algorithms onÂ recommendations forÂ children," in *Advances in Information Retrieval*, C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, and N. Tonellotto, Eds. Cham: Springer Nature Switzerland, 2025, pp. 67–84.

[4] L. Spear, A. Milton, G. Allen, A. Raj, M. Green, M. D. Ekstrand, and M. S. Pera, "Baby shark to barracuda: Analyzing children's music listening behavior," in *Proceedings of the 15th ACM Conference on Recommender Systems*, ser. RecSys '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 639–644. [Online]. Available: https://doi.org/10.1145/3460231.3478856

[5] T. Alimi Selmani, "The influence of music on the development of a child: Perspectives on the influence of music on child development," *EIKI Journal of Effective Teaching Methods*, vol. 2, pp. 191–201, 04 2024.

[6] M. Schedl and C. Bauer, "Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young," *CoRR*, vol. abs/1912.11564, 2019. [Online]. Available: http://arxiv.org/abs/1912.11564

[7] T. Bogt, M. Delsing, M. Zalk, P. Christenson, and W. Meeus, "Intergenerational continuity of taste: Parental and adolescent music preferences," *Social Forces - SOC FORCES*, vol. 90, pp. 297–319, 09 2011.

[8] A. Laplante, "Improving music recommender systems: What can we learn from research on music tastes?" in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*. Taipei, Taiwan: ISMIR, 2014, pp. 451–456, licensed under Creative Commons Attribution 4.0 International License (CC BY 4.0).

[9] I. Papadimitriou, "Leveraging children's music preferences to enhance the recommendation process," Master's thesis, Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, Netherlands, 2024, master's thesis. Chair: Sole Pera. Supervisor: Sole Pera. Daily Supervisor: Sole Pera. External Examiner: Maliheh Izadi. Project Duration: May 2024 – December 2024. Student No. 5847079. [Online]. Available: https://repository.tudelft.nl/file/File_a38a6a7a-2fd3-4a89-b58e-e905a4ec3a8f

[10] D. Cheng, T. Joachims, and D. Turnbull, "Exploring acoustic similarity and preference for novel music recommendation," *International Symposium on Music Information Retrieval*, 2020. [Online]. Available: https://par.nsf.gov/biblio/10290593

[11] G. Cerati, "Difficult to define, easy to understand: the use of genre categories while talking about music," *SN Social Sciences*, vol. 1, no. 12, pp. 1–20, 2021. [Online]. Available: https://doi.org/10.1007/s43545-021-00296-2

[12] P. MÃŒllner, D. Kowald, M. Schedl, C. Bauer, E. Zangerle, and E. Lex, "Lfm-beyms," May 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3784765

[13] B. Semerdzhiev, "Improving music recommender systems for youngsters: Using the listening history of youngsters to predict the features of the perfect song," pp. 1–8, 2024, supervised by Maria Soledad Pera (Boise State University) and Robin Ungruh (TU Delft - Web Information Systems). [Online]. Available: https://resolver.tudelft.nl/70258914-4432-4024-993c-308ca136949c

[14] B. Ferwerda, M. Tkalcic, and M. Schedl, "Personality traits and music genre preferences: How music taste varies over age groups," ser. CEUR Workshop Proceedings. ACM, 01 2017, pp. 16–20.

[15] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *International Journal of Multimedia Information Retrieval*, vol. 7, no. 2, pp. 95–116, 2018. [Online]. Available: https://doi.org/10.1007/s13735-018-0154-2

[16] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)*, Queen Mary University of London, UK, Jun. 2012, pp. 395–410, 19–22 June 2012.

[17] L. Barrington, R. Oda, and G. R. G. Lanckriet, "Smarter than genius? human evaluation of music recommender systems." in *ISMIR*, K. Hirata, G. Tzanetakis, and K. Yoshii, Eds. International Society for Music Information Retrieval, 2009, pp. 357–362. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismir/ismir2009.html#BarringtonOL09

[18] D. Perera, M. Rajaratne, S. Arunathilake, K. Karunanayaka, and B. Liyanage, "A critical analysis of music recommendation systems and new perspectives," in *Human Interaction, Emerging Technologies and Future Applications II*, T. Ahram, R. Taiar, V. Gremeaux-Bader, and K. Aminian, Eds. Cham: Springer International Publishing, 2020, pp. 82–87.

[19] R. Singhal, S. Srivatsan, and P. Panda, "Classification of music genres using feature selection and hyperparameter tuning," *Journal of Artificial Intelligence and Capsule Networks*, vol. 4, pp. 167–178, 08 2022.

[20] R. Stetler, "Exploring music genres: A study of optimal differentiation by feature," https://scholarworks.bgsu.edu/honorsprojects/788, 2022, honors Projects. 788, Bowling Green State University.

[21] M. D. Barone, J. Bansal, and M. H. Woolhouse, "Acoustic features influence musical choices across multiple genres," *Frontiers in Psychology*, vol. Volume 8 - 2017, 2017. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2017.00931

[22] D. Kowald, P. Müllner, E. Zangerle, C. Bauer, M. Schedl, and E. Lex, "Support the underground: Characteristics of beyond-mainstream music listeners," *CoRR*, vol. abs/2102.12188, 2021. [Online]. Available: https://arxiv.org/abs/2102.12188

[23] E. Zangerle and C. Bauer, "Evaluating recommender systems: Survey and framework," *ACM Comput. Surv.*, vol. 55, no. 8, Dec. 2022. [Online]. Available: https://doi.org/10.1145/3556536

[24] L. Michiels, "Methodologies to evaluate recommender systems," Doctoral thesis, University of Antwerp, Faculty of Science, Department of Computer Science, Antwerp, 2024, supervisor: Bart Goethals. Research supported by the Flanders Artificial Intelligence Research program (FAIR) - second cycle, and the Serendipity Engine project. [Online]. Available: https://repository.uantwerpen.be/docstore/d:irua:25195

[25] Z. Sun, H. Fang, J. Yang, X. Qu, H. Liu, D. Yu, Y.-S. Ong, and J. Zhang, "Daisyrec 2.0: Benchmarking recommendation for rigorous evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8206–8226, 2023.

[26] E. G. Gutiérrez, V. Charisi, and S. Chaudron, "Evaluating recommender systems with and for children: Towards a multi-perspective framework," in *CEUR Workshop Proceedings*, vol. 2955. CEUR-WS.org, 2021. [Online]. Available: http://ceur-ws.org/Vol-2955/

[27] G. van Capelleveen, "Industrial symbiosis recommender systems," Ph.D. dissertation, University of Twente, May 2020, phD thesis.

[28] S. C. Anyosa, J. a. Vinagre, and A. M. Jorge, "Incremental matrix co-factorization for recommender systems with implicit feedback," in *Companion Proceedings of the The Web Conference 2018*, ser. WWW '18. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2018, pp. 1413–1418. [Online]. Available: https://doi.org/10.1145/3184558.3191585

[29] S. Y. Sert, Y. Ar, and G. E. Bostancı, "Evolutionary approaches for weight optimization in collaborative filtering-based recommender systems," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, no. 3, pp. 1982–1996, 2019. [Online]. Available: https://journals.tubitak.gov.tr/elektrik/vol27/iss3/40

[30] G. Carraturo, V. Pando-Naude, M. Costa, P. Vuust, L. Bonetti, and E. Brattico, "The major–minor mode dichotomy in music perception: A systematic review on its behavioural, physiological, and clinical correlates," *bioRxiv*, 2023. [Online]. Available: https://www.biorxiv.org/content/early/2023/03/18/2023.03.16.532764

[31] M. Ferrari Dacrema, S. Boglio, P. Cremonesi, and D. Jannach, "A troubling analysis of reproducibility and progress in recommender systems research," *ACM Trans. Inf. Syst.*, vol. 39, no. 2, Jan. 2021. [Online]. Available: https://doi.org/10.1145/3434185

Appendix

7.1 Results Tables with Corresponding p-values

Recommender Algorithm	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10	HitRate p-val	MRR p-val	NDCG p-val
Item-KNN + Acousticness	0.1326	0.0402	0.0200	0.3448	0.9950	1.0000	**0.0447**	0.0584
Item-KNN + Danceability	0.1362	0.0372	0.0194	0.3501	0.9951	1.0000	0.1386	0.1420
Item-KNN + Energy	0.1326	0.0366	0.0188	0.3510	0.9952	1.0000	0.1834	0.3313
Item-KNN + Instrumentalness	0.1577	0.0416	0.0226	0.3280	0.9945	0.2295	0.0894	**0.0391**
Item-KNN + Liveness	0.1326	0.0360	0.0185	0.3478	0.9952	1.0000	0.5281	0.6784
Item-KNN + Loudness	0.1470	0.0390	0.0200	0.3503	0.9953	0.1250	**0.0031**	**0.0037**
Item-KNN + Mode	0.1613	0.0418	0.0219	0.3076	0.9952	0.2153	0.1874	0.2184
Item-KNN + Speechiness	0.1326	0.0353	0.0185	0.3484	0.9951	1.0000	0.2249	0.2249
Item-KNN + Tempo	0.1398	0.0369	0.0195	0.3486	0.9953	0.5000	0.4007	0.2326
Item-KNN + Valence	0.1326	0.0368	0.0189	0.3493	0.9952	1.0000	0.4004	0.6002
Item-KNN (Baseline)	0.1326	0.0361	0.0187	0.3503	0.9952	-	-	-

Table 5: Results for Age 15: Performance metrics and Wilcoxon p-values for MRR@10, and NDCG@10; McNemar p-values for HitRate@10. Statistically significant p-values compared to the baseline ($p < 0.05$) are highlighted in bold.

Recommender Algorithm	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10	HitRate p-val	MRR p-val	NDCG p-val
Item-KNN + Acousticness	0.1183	0.0306	0.0190	0.3567	0.9949	1.0000	0.8766	0.6873
Item-KNN + Danceability	0.1147	0.0314	0.0183	0.3574	0.9950	1.0000	0.3985	0.2845
Item-KNN + Energy	0.1147	0.0325	0.0183	0.3576	0.9950	1.0000	0.8923	0.7532
Item-KNN + Instrumentalness	0.1183	0.0382	0.0180	0.3202	0.9946	1.0000	0.2678	0.6374
Item-KNN + Liveness	0.1147	0.0322	0.0185	0.3569	0.9951	1.0000	0.4615	0.2249
Item-KNN + Loudness	0.1183	0.0318	0.0180	0.3531	0.9951	1.0000	0.8335	0.8334
Item-KNN + Mode	0.1039	0.0324	0.0162	0.3127	0.9947	0.6900	0.7512	0.5015
Item-KNN + Speechiness	0.1147	0.0319	0.0182	0.3552	0.9951	1.0000	0.4615	0.6002
Item-KNN + Tempo	0.1147	0.0320	0.0188	0.3561	0.9950	1.0000	0.8880	0.5406
Item-KNN + Valence	0.1147	0.0318	0.0184	0.3578	0.9951	1.0000	0.2164	0.4838
Item-KNN (Baseline)	0.1147	0.0320	0.0182	0.3569	0.9950	-	-	-

Table 6: Results for Age 16: Performance metrics and Wilcoxon p-values for MRR@10, and NDCG@10; McNemar p-values for HitRate@10. Statistically significant p-values compared to the baseline ($p < 0.05$) are highlighted in bold.

Recommender Algorithm	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10	HitRate p-val	MRR p-val	NDCG p-val
Item-KNN + Acousticness	0.1039	0.0203	0.0116	0.3773	0.9950	0.6875	0.1649	**0.0499**
Item-KNN + Danceability	0.1039	0.0192	0.0106	0.3739	0.9952	0.5000	**0.0343**	0.0910
Item-KNN + Energy	0.1004	0.0184	0.0101	0.3737	0.9952	1.0000	0.7150	0.7525
Item-KNN + Instrumentalness	0.1111	0.0276	0.0123	0.3416	0.9946	0.6076	0.4405	0.8213
Item-KNN + Liveness	0.0968	0.0182	0.0099	0.3759	0.9951	1.0000	0.9165	0.8885
Item-KNN + Loudness	0.1004	0.0187	0.0102	0.3782	0.9952	1.0000	0.5743	0.7533
Item-KNN + Mode	0.1434	0.0390	0.0169	0.3208	0.9951	**0.0294**	**0.0011**	**0.0067**
Item-KNN + Speechiness	0.1039	0.0190	0.0106	0.3741	0.9952	0.5000	0.0769	0.0796
Item-KNN + Tempo	0.0932	0.0178	0.0096	0.3724	0.9952	1.0000	0.5754	0.4838
Item-KNN + Valence	0.1004	0.0187	0.0102	0.3735	0.9952	1.0000	0.1041	0.1380
Item-KNN (Baseline)	0.0968	0.0181	0.0099	0.3756	0.9951	-	-	-

Table 7: Results for Age 17: Performance metrics and Wilcoxon p-values for MRR@10, and NDCG@10; McNemar p-values for HitRate@10. Statistically significant p-values compared to the baseline ($p < 0.05$) are highlighted in bold.

Recommender Algorithm	HitRate @10	MRR @10	NDCG @10	Coverage @10	Diversity @10	HitRate p-val	MRR p-val	NDCG p-val
Item-KNN + Acousticness	0.1219	0.0315	0.0168	0.3695	0.9953	0.2891	0.1767	0.2789
Item-KNN + Danceability	0.1111	0.0306	0.0169	0.3714	0.9954	1.0000	0.5534	0.4406
Item-KNN + Energy	0.1075	0.0306	0.0163	0.3714	0.9955	1.0000	0.1088	0.2249
Item-KNN + Instrumentalness	0.1183	0.0315	0.0167	0.3446	0.9952	0.6900	0.4717	0.7267
Item-KNN + Liveness	0.1183	0.0313	0.0172	0.3688	0.9955	0.2500	0.1207	0.2353
Item-KNN + Loudness	0.1183	0.0290	0.0174	0.3701	0.9955	0.2500	0.1806	0.1404
Item-KNN + Mode	0.1398	0.0387	0.0196	0.3263	0.9954	0.1628	0.1060	0.2651
Item-KNN + Speechiness	0.1219	0.0292	0.0172	0.3722	0.9955	0.1250	0.2334	0.1677
Item-KNN + Tempo	0.1147	0.0286	0.0169	0.3697	0.9955	0.5000	0.2204	0.1621
Item-KNN + Valence	0.1111	0.0280	0.0158	0.3731	0.9955	1.0000	0.8927	0.6744
Item-KNN (Baseline)	0.1075	0.0302	0.0163	0.3714	0.9955	-	-	-

Table 8: Results for Age 18: Performance metrics and Wilcoxon p-values for MRR@10, and NDCG@10; McNemar p-values for HitRate@10. Statistically significant p-values compared to the baseline ($p < 0.05$) are highlighted in bold.

Feature	Age 15 MRR/NDCG	Age 16 MRR/NDCG	Age 17 MRR/NDCG	Age 18 MRR/NDCG
Acoust.	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Dance.	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Energy	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Instrumentallness	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Live.	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Loud.	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Mode	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Speech.	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Tempo	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0
Val.	0.0/0.0	0.0/0.0	0.0/0.0	0.0/0.0

Table 9: Shapiro-Wilk p-values for MRR/NDCG by Age

7.2 LLM Usage

We used ChatGPT and Grok to generate Tables 1, 2, 4, 5, 6, 7, 8, 9 and Figures 1, 2. After generating the tables and figures, we carefully verified all values to ensure accuracy by cross-checking them against the original data.

Furthermore we used Grammarly's generative AI solely for grammar checking and proofreading. It was not used to extend the text. Only specific unclear sentences were revised for grammatical correctness. The resulting texts were not used directly; instead, they were manually paraphrased before use.

Prompts Used for LLM-Based Table and Figure Generation

We utilized Large Language Models only for table creation and figure layout. The following bullet points show the example prompts that we used, and Figures 4 and 5 are an example of precise prompts that we used.

- "Here are the raw results from my Python script. Can you convert them into a LaTeX table?"
- "The LaTeX table does not fit within the document margins. Can you adjust it accordingly?"
- "Can you create a LaTeX histogram using the following data?"
- "Can you create a LaTeX table from the descriptive text provided below?"
- "Can you align these images in two rows with five columns each? Here is the directory containing the image files."

```
Test Set Data
Age Group       Total Interactions      Average Interaction     User Count
15      3221     11,54  279
16      3233     11,59  279
17      3192     11,44  279
18      3322     11,91  279
Total   12968    11,62  1116
can you create the latex version of this table
```

Figure 4: Prompt for Table Creation

```
ge Group: 18
============================

age 18

Feature           HitRate@10     MRR@10 Srecall@10    NDCG@10 Coverage@10 Entropy@10 Diversity@10
key                   0.1354     0.0339     0.3878     0.0185     0.3586     0.9725     0.9955
acousticness          0.1146     0.0327     0.3814     0.0168     0.3707     0.9743     0.9953
danceability          0.1181     0.0324     0.3807     0.0181     0.3737     0.9740     0.9955
energy                0.1215     0.0345     0.3805     0.0186     0.3754     0.9742     0.9955
instrumentalness      0.1250     0.0358     0.3946     0.0178     0.3478     0.9686     0.9951
liveness              0.1250     0.0356     0.3810     0.0189     0.3737     0.9739     0.9954
loudness              0.1215     0.0319     0.3792     0.0182     0.3735     0.9739     0.9956
mode                  0.1424     0.0532     0.4105     0.0220     0.3293     0.9652     0.9954
speechiness           0.1181     0.0345     0.3794     0.0185     0.3744     0.9740     0.9955
tempo                 0.1215     0.0322     0.3819     0.0182     0.3727     0.9734     0.9955
valence               0.1181     0.0319     0.3810     0.0178     0.3731     0.9741     0.9955
pure                  0.1181     0.0325     0.3788     0.0181     0.3746     0.9741     0.9955

and this is the overall can you convert this to a latex table
```

Figure 5: Prompt for Table Creation for Age Group 18