Improving Music Recommender Systems For Youngsters

Using the listening history of youngsters to predict the features of the perfect song

Borislav Semerdzhiev[1]

Supervisor(s): Dr. Maria Soledad Pera[1]**, Robin Ungruh**[1]

[1]**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 23, 2024

Name of the student: Borislav Semerdzhiev
Final project course: CSE3000 Research Project
Thesis committee: Dr. Maria Soledad Pera , Robin Ungruh, Julian Urbano Merino

Abstract

Music plays a crucial role in children's development by helping them express their identity, teaching them to belong to a culture, and developing their cognitive well-being and inner self-worth [4, 8]. Most music nowadays is consumed through online streaming websites like Spotify [3], which make use of recommendation systems to suggest new tracks. These recommendation algorithms internally make use of user and song features that aid in making predictions. However, as most of the research literature focuses on making recommendations for the adult group, few studies explore what makes the recommended songs appealing to children. In this paper, we perform user modelling techniques on the listening history of the children combined with high-level descriptors of the songs in order to capture their music preferences. By constructing user profiles, we aim to identify the characteristics of music that resonate most with the listener. In our research, we focus on children belonging to the age group of 6-17 years. The goal is to aid in the design of future recommender systems that operate with greater transparency[5], allowing the impact of the user choices to be clearly observed by the consumer.

1 Introduction

Music is an essential part of our lives [7]. Nowadays, most of it is consumed through online streaming services [3] that make use of music recommendation systems (**MRS**) that aid the user in selecting new songs out of the millions available to him [14, 13]. However, existing MRS are designed to cater mainly to the majority of listeners—the adults. Therefore, designing a MRS specifically tailored for the youngster group has the potential to outperform the current ones, as there is a significant difference in the listening behaviour between a child and an adult user [9].

Considering that song feature preferences can be used for recommendation [2], it is worth answering whether they can also easily be used in recommendation systems for young listeners. The impact of music features on the listening behaviour of children has been explored by Spear et al.[15], revealing that children's music preferences are influenced by different traits as they age. This suggests that a "one size fits all" recommendation strategy does not exist for children across various age groups. Although works such as the aforementioned have examined how music traits affect the listening behaviour of children, we believe that research has only begun to explore how they could be leveraged to enhance recommendation systems for children.

Schedl et al. [12, 11] observe that user preferences have been largely neglected in the realm of music recommendation. They also emphasize that user modelling techniques remain insufficiently explored in depth. In light of this, we believe that user modelling techniques, which focus on the individual user, have significant potential to capture children's music preferences accurately.

To advance knowledge in MRS for children, this paper proposes a method for modelling user profiles for young listeners. Those profiles are built upon children's listening history and enriched with features that describe the songs. The goal is to capture the unique music preferences of young users through high-level music features.

The possible uses of the user profiles include making MRS more transparent to the user. The captured user preferences can be used to explain why a certain track has been suggested to a user.

The remainder of this paper is organised as follows. Section 2 discusses related work, and Section 3 presents the dataset, features, and user models. Section 4 details the experimental setup and Section 5 presents the results of our study, which are then discussed in Section 6 where we further go over the potential implications and applications of the user-profiles. Section 7 discusses the responsible research aspect of our study. Section 8 concludes the paper and discusses future work and limitations.

2 Related work

Schedl et al. [12, 11] note in their work that the user's preferences have been mainly neglected when it comes to music recommendation tasks. One of the problems they discuss in their studies is the traditional approach of assuming the existence of an objective "ground truth" against which different MRS are evaluated. The main issue here is that this "ground truth" might be an ill-defined concept, as had happened before with genre classification experiments[1]. Finally, they state that user modelling techniques have not yet been explored and evaluated in depth in the context of MRS, and there is potential worth in doing so as they avoid some of the pitfalls that MRS currently suffers from.

For musical preferences of users, Pitch et al. [6] found in a large-scale study of Spotify users that users listen to different types of music, which they also store in different playlists based on the types. Those types can furthermore be observed using k-means clustering based on the track descriptors. The authors highlight the importance of comprehensive user modelling techniques that thoroughly capture unique user preferences. Those findings lead us to the conclusion that k-means could be an effective strategy for modelling the users.

In the context of our target group - children between the ages of 6 and 17, Spear et al.[15] have found that children from different age groups are interested in different aspects of the songs as determined by differences in the high-level features of the tracks. Their findings suggest that even though there is a "stereotypical" audience among the children, there is enough difference between young listeners of different age groups, such that current recommendation strategies designed for adults might not work on them. This further reinforces our belief that MRS for children could greatly benefit from user modelling techniques.

Along these lines, we propose a user modelling technique that would capture the unique preferences of young listeners.

3 Methodology

In this section, we first present the dataset and song descriptors we will be using, and then we follow with a method of creating user profiles which aims to capture the listening preferences of the children. The method we use for building the profiles is motivated by the modelling strategy presented by Pich et al. [16], and the feature selection is driven by prior studies on children's music preferences [15, 4].

3.1 Dataset and features

We use the LFM-2b[10] dataset, which contains $2,014,164,872$ real listening events (LE) of which $49,423,141$ are from children in the target age group (6-17 years old), as the primary basis of our study. For each LE, the dataset includes information about the track, the user, and the demographics. Besides the information provided in the LFM-2b dataset, we rely on Spotify API[1] to extract the following descriptions for each track(definitions are taken from Spotify API documentation):

1. *Danceability* describes how suitable a track is for dancing based on a combination of musical elements, including tempo, rhythm stability, beat strength, and overall regularity. The value is a decimal number in the range of [0, 1].
2. *Energy* measures the perceived intensity and activity of a track. This feature is based on the dynamic range, perceived loudness, timbre, onset rate and general entropy of a track. The value is a decimal number in the range of [0, 1].
3. *Instrumentalness* predicts whether a track contains no vocals. The value is a decimal number in the range of [0, 1].
4. *Acousticness* is a confidence measure of whether the track is acoustic. The value is a decimal number in the range of [0, 1].
5. *Tempo* is the overall estimated tempo of a track in beats per minute (BPM). The value is a positive decimal number.
6. *Liveness* detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live. The value is a decimal number in the range of [0, 1].
7. *Speechiness* detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audiobook, poetry), the closer to 1.0 the attribute value. The value is a decimal number in the range of [0, 1].
8. *Valence* measures from 0.0 to 1.0, describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

[1] https://developer.spotify.com/documentation/web-api

Item	Value
Listening Events(LE)	49,423,141
Users	3,350
Tracks distinct	1,011,435
Min. LE per User	10
Q_1 LE per User	1,103
Median LE per User	7,160
Q_3 LE per User	18,157
Max. LE per User	339,215
Avg. LE per User	14,753.18

Table 1: Dataset statistics computed only using LE from users in our target group (children between the ages of 6-17).

Pich et al.[16] have achieved significant results in making recommendations by utilizing user profiles that use the aforementioned features. With that in mind, we chose to use the same high-level descriptors for our user profiles.

We present descriptive statistics about the dataset in Table 1. On average, users have made $14,753.18$ LE, and the minimum amount of LE a user has is 10.

Feature extraction

To obtain these features for all tracks of the dataset, we get the `track_id` for a song and find the corresponding Spotify URI, which we then query using the Spotify API. Multiple track_ids might be related to the same Spotify URI, which needs to be accounted for, as this might affect the clustering part of the algorithm. In order to query the $1,011,435$ songs we have in our dataset, some important factors should be noted. First, Spotify API states that they use a time window of 30 seconds to determine if a user is making too many requests, and if they determine that this is the case, he receives a 24-hour time out. Therefore, they suggest that after every query you should check the 'Retry after' header and send a new request after said seconds. However, our experience of using the API was much different. The aforementioned header did not contain any value, and from our observations, we concluded that no time window of 30 seconds existed, but rather, every account that a user registers at Spotify API receives $200,000$ requests per 24 hours. Our belief is that the documentation does not reflect the current state of their system. Therefore, we decided that the best course of action was to create multiple accounts, put their credentials in a list, and continuously query from one account until we get a response that the current account has been blocked. Afterwards, we continue with the next one. In that manner, we managed to extract all the track features in the time span of 24 hours.

3.2 Feature Space

Before using our feature space for any computations, we normalize all the feature values to be between 0 and 1. Optimally, we would cluster all of the songs into different groups based on the feature similarity between the songs inside of them. However, the large number of tracks we are working with (around 1 million) makes the task of performing the clustering operation computationally expensive. Therefore, we first perform a dimensionality reduction on the tracks' features. To achieve this, we decided to test both PCA and UMAP and

reduce the feature space from the original 8 to 2. Although reducing the dimensionality comes at the cost of losing information about the relationship between different songs, a feature space of 2 allows us to visualize the data points while still preserving enough data to assist in the subsequent clustering phase. This procedure allows us to be more efficient as we perform the clustering while also providing us with an informative visualisation of the proximity of the tracks that can assist us during the creation of the user profiles.

3.3 User Profiles

To create the user profiles we utilize the membership of users in each cluster. We perform the clustering step by using the k-means clustering algorithm. To find which cluster of tracks most closely represents the user's preference, we count the number of songs the user has listened to in each cluster and choose the cluster with the highest amount. Once we have determined which cluster we use to represent the user's preferences, we can take the average of the non-normalized features of the tracks that both the user has listened to and also belong to the chosen cluster in order to approximate the features of the song the user will enjoy the most. The main advantage of using clusters is that outliers (songs that the user has listened to a small number of times and have features that differ significantly from the rest) do not impact the overall user preference severely.

3.4 Number of clusters

Choosing the correct number of clusters is a crucial aspect that directly dictates how accurate our user profiles will be. A too-small number of clusters means that there is a small variety in the user profiles, which might not represent the specific aspects of the user's preference, while a too-large number eventually leads to a cluster that will contain just a few songs. In order to find a good balance between overgenerality and being too specific, we employ silhouette analysis. This method provides a way to assess the number of clusters visually and is a measure of how close each point in one cluster is to points in the neighbouring clusters. The silhouette score is in the range of $[-1, +1]$, and a value close to $+1$ means that the sample is far away from neighbouring clusters, a value of 0 means that the sample is very close to a decision boundary, and a negative score means that the sample could be assigned to the wrong cluster.

4 Experimental Setup

We model the evaluation of the proposed user profiling method as a task of estimating the features of the most listened-to song by the child. For this purpose, we have made the assumption that the song that has been replayed the most amount of times represents the music preference of the child the best.

Since our research focuses on children between the ages of 6 and 17, all the LE of users who do not fall in this category are filtered out, as are users who have not provided their age and therefore have -1 in the column. Additionally, we ignore all of the songs that have been played less than 5 times by users in our target group. This step ensures that we won't be

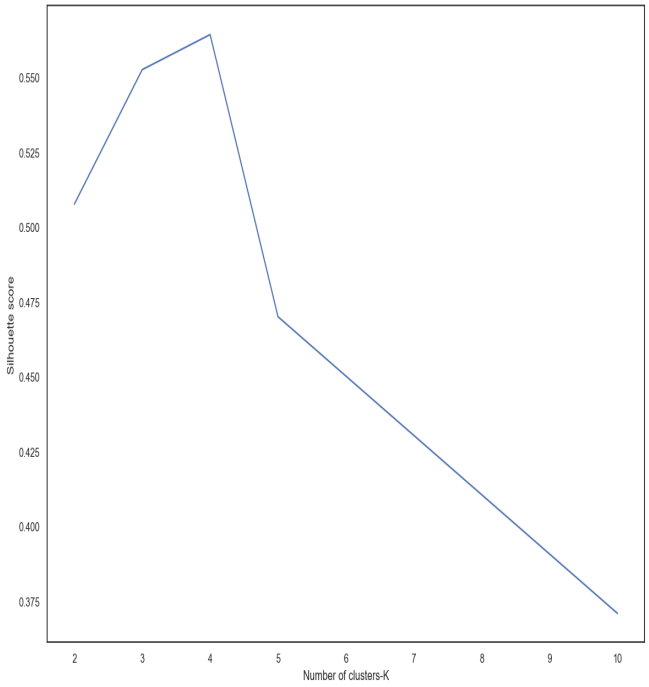

Figure 1: Silhouette score for the different number of clusters. The score gradually increases until it reaches its peak at 4 clusters, after which it drastically declines.

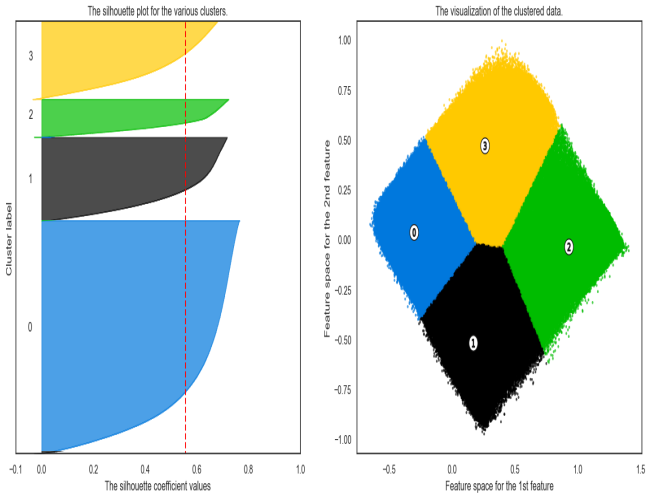

Figure 2: Silhouette analysis of the clustering. The left diagram showcases the silhouette coefficients for each cluster, while the right diagram displays each cluster and its centre.

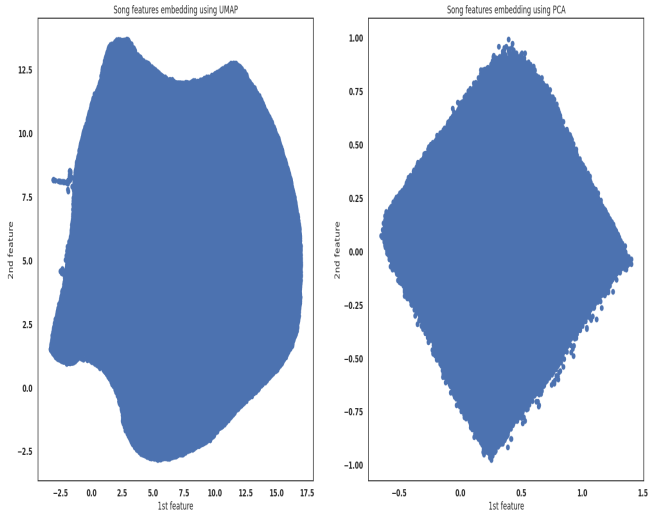

Figure 3: Song feature embedding with UMAP (left diagram) and PCA (right diagram).

forming unnecessary clusters during the k-means clustering stage.

In order to embed our feature space into 2D, we need to normalize all of our feature values. Since most of the song features are already normalized in the range of [0, 1], using Min-max scaling for the rest of them appears to be more suitable compared to other normalization techniques, such as Standard Scaling.

A comparison of the results of applying PCA and UMAP was performed. With both approaches, we reduced the feature space from 8 features to 2, and a slight difference can be observed between both embeddings in Figure 3. The difference between both approaches will be discussed in Section 5. However, our main evaluation will be based on PCA moving forward.

We rely on silhouette analysis to select the most appropriate number of clusters for our tracks and perform the clustering operation using k-means.

In order to evaluate how well the user profiles are able to approximate the track features of a song that will be frequently listened to by a given user, we will build the user profile without the child's most frequently listened track. We have made the assumption that the most listened-to track of a child is the one that most accurately represents his preference. Afterwards, we compare how similar the predicted features are to the actual features of the song the user has replayed the most. To calculate the difference between the predicted user preferences and the features of a track, we use cosine similarity, as all of the values are in the range between $[0, 1]$.

Due to the significant difference in LE between the Q1 user$(1, 103)$ and the median user$(7, 160)$, we repeat the test excluding the users with less LE than the Q1 user and also less than the Q2 user. This test shows us how much the average cosine similarity improves when users with fewer LE are not included in the mean result. This metric is important because it reveals the susceptibility of our user modelling method to users with less data.

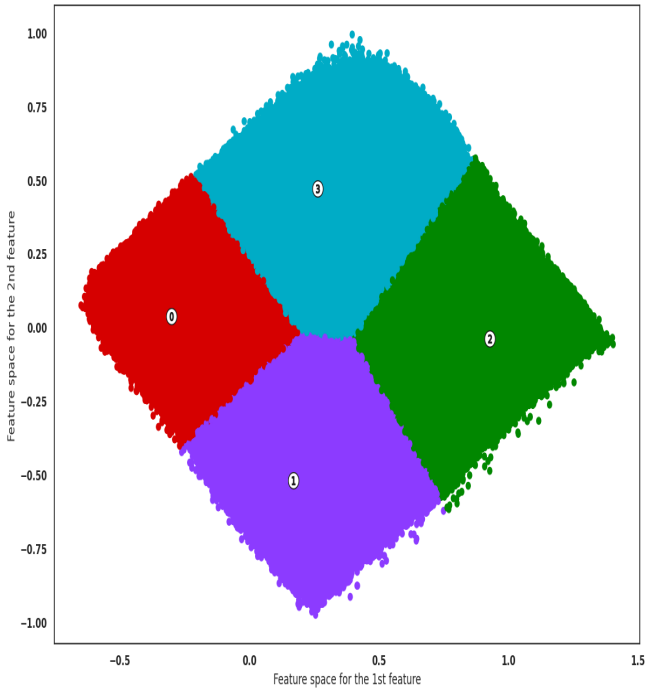

Figure 4: Clustering of the songs using k-means and the feature embedding produced by PCA. The centres of each cluster are showcased and numbered on the diagram.

To evaluate the importance of our cluster-picking strategy, we repeat the aforementioned cosine similarity test but this time picking the $2_{nd}, 3_{th}$ and 4_{th} cluster based on the number of songs the user has listened to in each of them. This test gives us insight into how much the cosine similarity decreases as we take less optimal choices for the cluster that represents the user's preferences.

Lastly, we conduct a test to evaluate the accuracy of our component prediction. Specifically, we assess how often the component we predict to be optimal for the user is indeed the one that maximizes the cosine similarity score. To perform this evaluation, we build a separate profile for the current user from each cluster and determine which cluster yields the highest cosine similarity score. Furthermore, we note how much the cosine similarity score is impacted by each choice of component compared to picking the most optimal one.

5 Results

The results of performing the clustering are displayed in Figure 4. When choosing the number of clusters, we performed the silhouette tests whose results can be observed in Figure 1. From the figure we can see that the score is the highest with K equal to 4, and as K grows higher than 4 the silhouette score decreases. A more in-depth silhouette analysis can be seen in Figure 2. The left part of the graph illustrates the silhouette coefficient for each cluster. We can see that each component has a silhouette coefficient between 0.7 and 0.8, which serves as an indicator of optimal clustering. Furthermore, the y-axis of the plot(the thickness of each component) showcases the amount of points inside each cluster. From the figure, we see

Figure 5: Confusion matrix showing the amount of times we have picked the i_{th} component when the j_{th} component is the most optimal. The x-axis represents the component we have picked, while the y-axis represents the optimal component.

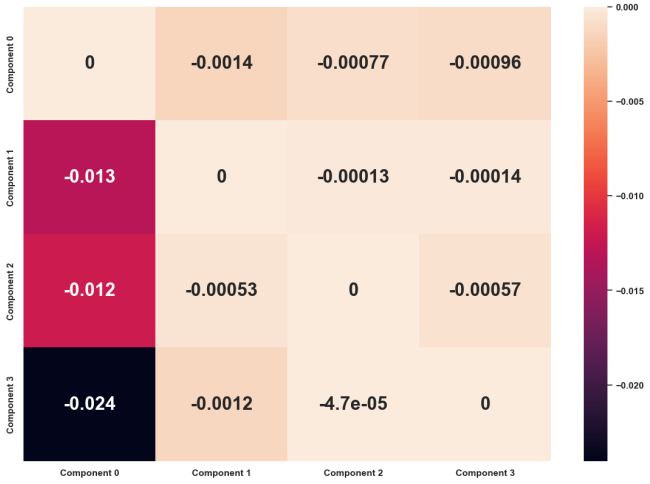

Figure 6: Total mean cosine similarity loss caused by picking the i_{th} component when the j_{th} component is the most optimal. The x-axis represents the component we have picked, while the y-axis represents the optimal component.

that cluster 0 contains the highest number of songs.

Figure 10 offers us deeper insight into what features separate the different clusters. The values in each diagram were computed by taking the mean of every feature inside the clusters. Cluster 0 has the highest energy score out of all of them, while cluster 1 comes in second place but has much higher instrumentalness. Cluster 2 heavily focuses on acousticness and instrumentalness, and finally, cluster 3 has comparable acousticness but practically no instrumentalness. The features that they all have nearly the same values on are tempo, liveness and speechiness.

We present the results of our cosine similarity evaluation in Figure 7. Cosine similarity measures how similar two vectors are, and in our case, we compare the predicted user preferences with the features of their most listened-to song. Cosine similarity close to $+1$ is an indicator that the two vectors are similar, while a value close to 0 is a sign that they are completely different. The mean cosine similarity of all the users is around 0.9, and 50% of all users have a cosine distance even higher than 0.93. Both PCA and UMAP achieve similar results despite the difference in the song feature embedding, which leads to different shapes in the 2D space. However, even though both methods achieve similar results, there is a significant difference in the computational time it takes to run both of them (embedding 1M tracks with 8 features each — PCA: 0.5 min, UMAP: 30 min). Due to this huge difference in the execution time and the fact that with both methods, we reach similar outcomes, we advise using PCA.

The long tails in both the UMAP and PCA boxplots (Figure 7) warrant further discussion. Based on the assumption that they are caused by users with fewer LE, we decide to filter the listeners that have less listening events than the Q1 user (the user that has more listening events than 25% of all the people), and also listeners that have less LE than the Q2 user. However, after recomputing the cosine similarity score following the filtering of those users, we find that the length of the tail still remains unchanged, suggesting our assumption is wrong. This is an indicator that the low scores for some users are not caused by the lack of LE, but rather some other unexplored reason. The result of this test can be seen in Figure 9.

The validity of our method for choosing the cluster which we use to build the user profile is tested. We perform this test by choosing the $2_{nd}, 3_{th}$ and 4_{th} most optimal component for each user. The results of this experiment are presented in Figure 8. We can clearly observe a decline in the cosine similarity score as we select components with fewer songs the user has listened to. The variance increases as we pick less optimal clusters, which can be explained by more low accuracy scores emerging for more users. This suggests that some users will have their cosine similarity score impacted much more severely by choosing a less optimal component.

We present the results of our **Confusion Matrix Test** in Figure 5. This test gives us insight into how often the component we have predicted to be the most representative of the user's preference is, in fact, a sub-optimal choice. The results show that component 0 is selected most often, which could be explained by the fact that the 0_{th} component contains the highest number of songs, as indicated in Figure 2. Moreover,

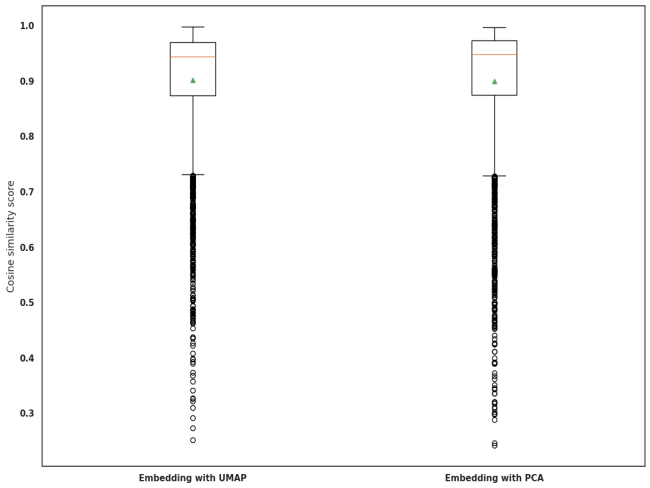

Figure 7: Cosine similarity score of all the users using UMAP embedding (left boxplot) and PCA (right plot).

for 30% of the user, we do not select the most optimal cluster, raising the question of how much accuracy is sacrificed as a result. Figure 6 depicts the total loss in cosine similarity score incurred from choosing the less optimal components.

In conclusion, the high average cosine similarity score suggests that k-means clustering is an effective approach to user modelling that allows us to accurately capture the children's music preferences. Furthermore, our choice for track descriptors, backed up by previous studies[15, 16], is confirmed to be an appropriate one for constructing user profiles for children.

6 Discussion

Music positively benefits children's development[8, 7] with music recommendation systems acting as a tool in navigating through the vast amount of songs available nowadays. However, there is still work ahead of us before the recommendation systems are ready to serve the young audience.

Building upon previous studies[15, 12, 16, 11], we have confirmed that high-level descriptors of the tracks can be used to model user profiles that capture the unique preferences of children. Although we found that the songs were best grouped in 4 clusters during the k-means clustering, Pichl et al. [16] found that the most optimal number of clustering of the tracks from the same dataset was 5 in their study. This difference could be due to the fact that we are working solely with songs that have been listened to at least 5 times by the children between the age of $6 - 17$, and the songs that have been filtered out could have contributed to another cluster being formed.

To help us gain more insight into what separates the different components, Figure 10 visualizes the average feature values within each cluster in a radar chart. This allows us to draw the following conclusions:

- **Cluster 0** is characterized by high energy and high danceability songs. As shown in Figure 2, this cluster contains the highest number of songs. The music genres that most closely match these features are Hip-Hop and Pop.

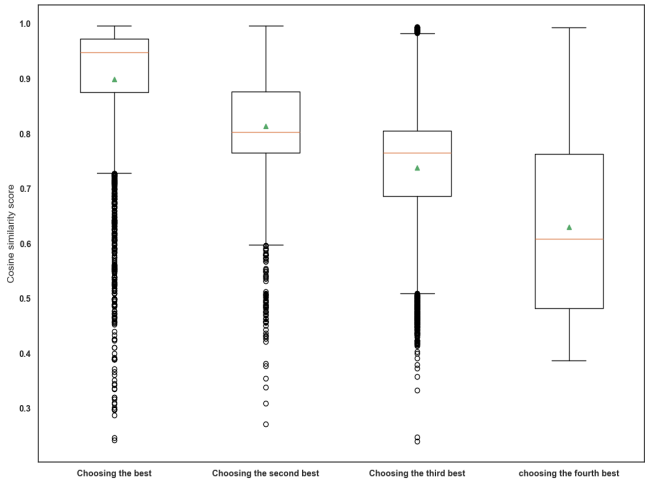

Figure 8: Cosine similarity score declining when choosing less optimal components. Left to right, the box plots present the cosine similarity as we pick the 1_{st}, 2_{nd}, 3_{th} and 4_{th} best component for each user.

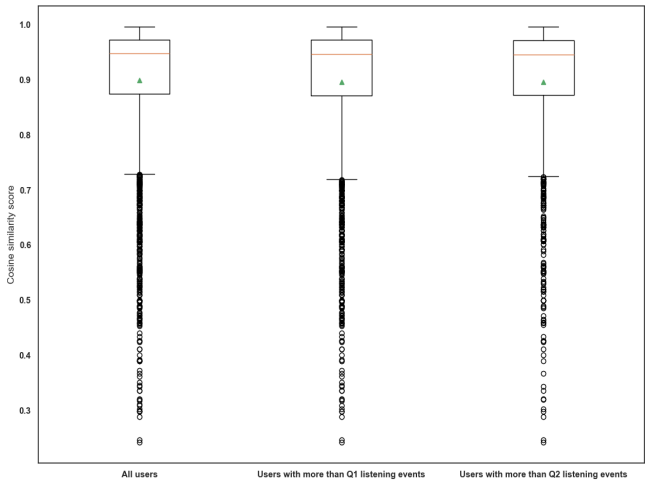

Figure 9: Cosine similarity score computed after filtering users with fewer than Q1 user's LE (middle box plot) and fewer than Q2 user's LE (right most box plot). The left-most box plot contains all the users and serves as a reference.

- **Cluster 1** features songs with high energy and high instrumentalness, suggesting it represents the Rock and Metal genre.
- **Cluster 2** contains songs high in instrumentalness and acousticness, but low on energy, characteristics closely related to classical music.
- **Cluster 3** contains songs with high acousticness and low instrumentalness, which could represent the Acoustic Pop Genre.

Our method of building user models combined with our feature selection yield a cosine similarity of 90%, which is consistent with multiple studies advocating for the necessity of user modeling in MRS and the consideration of diversity among children.

The results from the **Confusion Matrix Test** (Figure 5) warrant further discussion. Cluster 0 is the one most often chosen to represent children's preferences and often results to sub-optimal outcomes. However, we believe that the cause of these sub-optimal results is not our method of cluster selection but rather the characteristics of the songs and potentially the feature embeddings we use. As shown in Figure 2, Cluster 0 contains the highest number of songs, nearly as many as all of the other clusters combined. Therefore, it is unsurprising that the cluster with the highest amount of tracks also contains the highest number of songs a user has listened to. Another possible explanation is that users mainly listen to similar mainstream songs (the songs in Cluster 0), but their most listened-to song belongs to a different genre with features corresponding to another cluster. In this case, using a different evaluation strategy, rather than selecting the most listened-to song, might lead to more accurate results. Overall, while our current strategy produces sub-optimal results when creating user-profiles for some listeners, the total cosine similarity is not severely impacted by this. Figure 6 depicts the cosine similarity loss we have incurred due to our sub-optimal choices for components, and we see that the harshest we have been penalized is due to choosing Component 0 on rows 1, 2 and 3. In total, we lose around 0.05 similarity score, suggesting that our strategy is still an effective one.

Interpreting recommendations is a highly researched area in the field of recommendation systems [5] as most systems work in a black-box manner, providing little to no feedback on their decision-making process. A possible application of our study is by using the profiles we create for the children to enhance the interpretability of MRS by making use of the calculated preferences. To our knowledge, no research has yet focused on making recommendations more interpretable for children. Therefore, we believe our study could serve as a crucial first step in this direction.

7 Responsible Research

In our experiment, we have ensured easy reproducibility by using a well-known dataset[10] and a well-known public API for extracting the features. The features we use to describe the music content are high-level descriptors of the tracks, and have been used in multiple other researches [15, 16].

Since our study focuses on children between 6 and 17, we have purposefully removed all LE from users outside this group. Furthermore, we provide the dataset containing the collected track features using the Spotify API to facilitate the experiment's reproducibility.

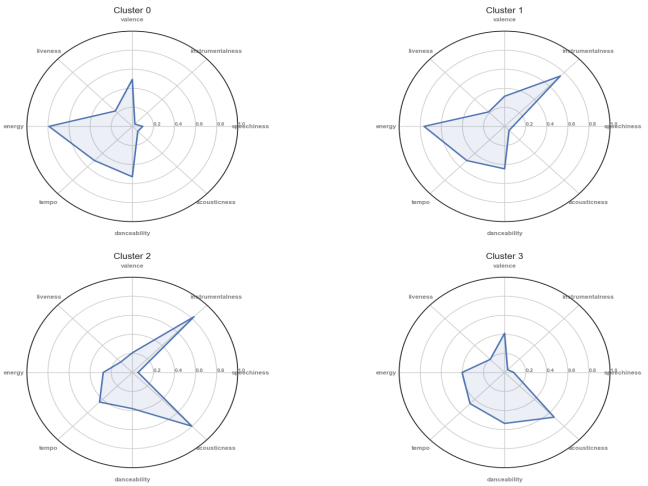

Figure 10: Radar charts displaying the mean feature values of the 4 clusters

The LFM-2b dataset additionally provides other demographic information about the users — nationality and gender, which we have decided are aspects that we do not focus on. Furthermore, we do not consider any of the LE from the users who have decided not to provide their age and, therefore, have a value of -1 in the age column.

The anonymity of the users is completely kept as the dataset does not contain any personally identifiable information such as names, email, addresses, or any other information that could be used to identify individuals.

Lastly, the Last.fm dataset is considered to be derivative work and their Terms of Service grants us a license to use their data.

8 Conclusions, Limitations, and Future Work

We proposed a method for capturing children's music preferences using their listening history enriched with high-level track descriptors. We find that clustering songs based on their features embedded into 2D latent space captures the intrinsic similarity between different songs. Furthermore, we find that picking the one cluster that contains the most songs the child has listened to is an effective approach when choosing which cluster represents the child's preferences most accurately.

This study extends previous research indicating the necessity of a more user-centred approach in MRS[12, 11]. Our user modelling strategy, as proposed by Pichl et al.[16], along with the chosen track descriptors studied by Spear et al.[15], has proven effective in capturing children's music preferences. Future work can leverage our findings to provide children with more in-depth explanations of why certain songs are recommended to them. The computed user profiles can serve as guidelines for aligning the features of recommended songs with the preferences of young listeners.

Our study focuses on children in the age group of 6-17 in our dataset. Potential limitations of our work include the fol-

lowing: the children may be influenced by the music preferences of the adults in their lives. Therefore, it is possible that we are not capturing their own listening behaviour but a projection of the one they are exposed to in their environment. Similarly, we can not ensure that a LE has been manually triggered since auto-play functionality exists, nor can we determine if a user likes a song due to the lack of provided ratings. Therefore, it is possible that the user profile we model does not represent the user's preferences but rather reflects the one Last.fm's MRS uses for them. Lastly, we cannot guarantee that the ages provided by users accurately reflect their real ages. As a result, our study may include "contaminated" listening events from users outside our target group. The impact of this issue depends on the validity of our data. However, given that we are using a well-established dataset employed in multiple other studies, such as [10, 16], this should not significantly affect our findings.

In future work, it might be worth trying to use more low-level descriptors of the tracks, such as the temporal, spectral, cepstra and perceptual audio descriptors, and seeing how well they capture the child's preference. This could potentially allow us to capture more unique aspects of children's tastes and result in a deeper understanding of their listening behaviour. Furthermore, it might be worth considering a better option for the "ground truth" we choose to represent the user's preferences(the most listened-to song by the user) while performing our evaluation. In doing so, the evaluation of our strategy will become a more accurate reflection of the true accuracy.

References

[1] Jean-Julien Aucouturier and Francois Pachet. Representing musical genre: A state of the art. *Journal of new music research*, 32(1):83–93, 2003.

[2] Dmitry Bogdanov, Martín Haro Berois, Ferdinand Fuhrmann, Emilia Gómez Gutiérrez, Herrera Boyer, et al. Content-based music recommendation based on user preference examples. In *Anglade A, Baccigalupo C, Casagrande N, Celma Ò, Lamere P, editors. Workshop on Music Recommendation and Discovery 2010 (WOMRAD 2010); 2010 Sep 26; Barcelona, Spain. Aachen: CEUR Workshop Proceedings; 2010. p. 33-8*. CEUR Workshop Proceedings, 2010.

[3] Michael Bull. *Investigating the Culture of Mobile Listening: From Walkman to iPod*, pages 131–149. Springer Netherlands, Dordrecht, 2006.

[4] Susan Hallam. The power of music: Its impact on the intellectual, social and personal development of children and young people. *International Journal of Music Education*, 28(3):269–289, 2010.

[5] Millecamp Martijn, Cristina Conati, and Katrien Verbert. "knowing me, knowing you": personalized explanations for a music recommender system. *User Modeling and User-Adapted Interaction*, 32(1):215–252, 2022.

[6] Martin Pichl, Eva Zangerle, and Günther Specht. Understanding playlist creation on music streaming platforms. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 475–480, 2016.

[7] Peter J Rentfrow. The role of music in everyday life: Current directions in the social psychology of music. *Social and personality psychology compass*, 6(5):402–416, 2012.

[8] Sarah J. Wilson Rudi Črnčec and Margot Prior. The cognitive and academic benefits of music to children: Facts and fiction. *Educational Psychology*, 26(4):579–594, 2006.

[9] Markus Schedl and Christine Bauer. Online music listening culture of kids and adolescents: Listening analysis and music recommendation tailored to the young. *CoRR*, abs/1912.11564, 2019.

[10] Markus Schedl, Stefan Brandl, Oleg Lesota, Emilia Parada-Cabaleiro, David Penz, and Navid Rekabsaz. Lfm-2b: A dataset of enriched music listening events for recommender systems research and fairness analysis. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, pages 337–341, 2022.

[11] Markus Schedl and Arthur Flexer. Putting the user in the center of music information retrieval. In *ISMIR*, pages 385–390, 2012.

[12] Markus Schedl, Arthur Flexer, and Julián Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.

[13] Markus Schedl, Peter Knees, Brian McFee, and Dmitry Bogdanov. *Music Recommendation Systems: Techniques, Use Cases, and Challenges*, pages 927–971. Springer US, New York, NY, 2022.

[14] Yading Song, Simon Dixon, and Marcus Pearce. A survey of music recommendation systems and future perspectives. In *9th international symposium on computer music modeling and retrieval*, volume 4, pages 395–410. Citeseer, 2012.

[15] Lawrence Spear, Ashlee Milton, Garrett Allen, Amifa Raj, Michael Green, Michael D Ekstrand, and Maria Soledad Pera. Baby shark to barracuda: Analyzing children's music listening behavior. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 639–644, New York, NY, USA, 2021. Association for Computing Machinery.

[16] Eva Zangerle and Martin Pichl. The Many Faces of Users: Modeling Musical Preference. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pages 709–716. ISMIR, November 2018.