Spotify Song Predictions 2023

THANEESHA K A
MSc Computer Science (Data Analytics)
Department of Computer Science
Rajagiri College of Scocial Sciences, Kalamassery

Abstract - The project utilizes unsupervised machine learning to group Spotify 2023 songs using the audio features.We preprocessed the data to contain only numeric values.There was normalization of data and missing values treated by taking into consideration the median value.There was K-Means and Hierarchical clustering carried out, and the Elbow Method and Silhouette Coefficient helped us decide on the number of clusters.The results show that the data naturally separate into two large clusters with great subsets of songs sharing highly similar musical features. The research confirms that cluster algorithms are useful to further our knowledge of music composition and to make recommendation systems more efficient.Spotify provides several musical analytics capabilities, including cluster algorithms like K-Means and Hierarchical clustering, and the Silhouette Method used to measure the quality of clusters.

Keywords -Spotify, clustering, K-Means, Hierarchical clustering, Silhouette Method, music analytics.

I. Introduction

Spotify and similar websites have evolved into having enormous databases which can be used to store collections of music features for millions of songs. Dissecting features like danceability, energy, loudness, and tempo into their components, you can see underlying trends in music. Supervised learning may be employed to forecast the popularity of a track, and unsupervised learning, in this case, clustering, can identify natural groups of tracks without any labeled data. We use the 2023 Spotify dataset in our experiment to test whether or not clustering can tell us interesting groups of music. Employing Hierarchical clustering and K-Means, we attempt to uncover natural groups and quantify how intense they are.

II. Literature Review

With increasingly more streaming platforms like Spotify prescribing how the world listens to music in the present, there is increasingly more fascination with the ability to predict what songs will be of interest to individuals.
Machine learning, clustering, and behavior analysis have been applied over the past decade or so in studies on variables that lead to the success of a song. Five such notable studies are covered in this section.

[1] Searched through 114,000 Spotify tracks over the past two decades to determine whether machine learning can identify the popularity of a track. We employed three models: Simple Linear Regression, Random Forest Regressor, and Gradient Boosting Machines. Random Forest performed best in detecting non-linear relationships between the features. Dancing ability, level of enthusiasm, and volume were the top three features. The study also showed that marketing and social media need to be taken into account since sound cannot rightly represent the popularity potential of a track.

[2] Contrasted 30,000 tracks' popularity on Spotify with regression, decision trees, random forests, and neural networks in an attempt to forecast popularity for every track. Their work showed that sonic attributes such as danceability, loudness, and tempo were crucial. Their work also illustrated how streaming companies and record labels could utilize the findings in a practical way, for example, through improving playlist suggestions and helping new artists produce good music. This research showed that machine learning algorithms can be incredibly useful to practitioners if supported by good feature engineering.

[3] Tried forecasting hit charts based on audio and streaming interaction features in the 14,639 distinct song Spotify Top 200 Daily Charts data. They employed four different models: XGBoost, Random Forest, K-Nearest Neighbors, and Logistic Regression. Tree-based models like Random Forest and XGBoost worked very well with macro F1-scores of about 0.95 and accuracy of about 97%. Their findings indicated that models learned from audio features alone could still make reasonable predictions, even without metadata like stream numbers. The technique is useful in scouting and sampling artists and repertoire prior to release.

[4] Utilized the K-Means algorithm for top 2023 Spotify songs clustering and then calculated the result using Silhouette Coefficient. They did it through an unsupervised learning approach. Two well-separated clusters were obtained by the study with a value of silhouette = 0.81 to affirm good separation. The study uncovered insights of music trends and listening interests of the audience by structuring data in the format of artist and song frequency data. The results show that clustering algorithms have the ability of improving recommendation systems by personalizing content to a group of people according to their requirements.

[5] Looked at Spotify users from another perspective by taking into consideration the way they act as well as music streaming. Twenty subjects were recruited to interview them regarding their background in streaming. These numbers accounted for three listening styles: emotional lean-back, attentive listening, and emotional lean-forward.
The research reflected the co-existence of algorithmic recommendation and user reflexivity, and the manner in which users consciously switch between both their own requirements and default recommendations.

This research teaches us about the individual parameters which are controlling the way people are engaging with music, something essential to improve prediction models. This research reflects how impossible it is to forecast song success. Random Forest, Gradient Boosting, and XGBoost are all machine learning algorithms best suited to providing

numeric output. Clustering, however, provides recommendations that are more personalized to you. While qualitative study of the way individuals respond to music reveals that more than the quality of sound for a song is at issue in creating a song into popularity, it also relies on many social, situational, and reflexive factors. Future systems will require the integration of the computational and behavior paradigms so that music prediction will be robust and sensitive to context.

III. Materials and Methods

- Dataset: Spotify 2023 dataset (~n rows, ~m columns). Only numeric features were retained.
- Preprocessing:
  - ➢ Removed non-numeric columns.
  - ➢ Imputed missing values with median imputation.
  - ➢ Standardized features using z-score scaling.
- Clustering Models:
  1. K-Means Clustering - To find number of clusters K for K-Means Clustering, we can utilize the Elbow Method and Silhouette Analysis.
  2. Hierarchical Clustering - Ward's method with Euclidean distance. Dendogram used to visualize clusters.
- Checking the quality of the clusters: The silhouette score was utilized to find out how well the groups fit together and how well they fit apart.

IV. Data Visualization and Interpretation

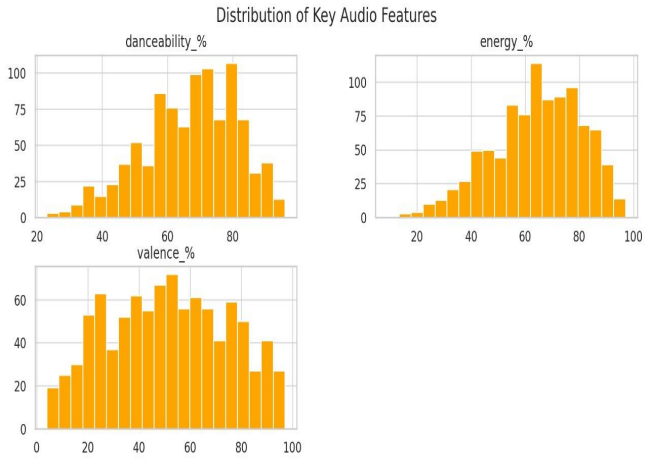


Fig: Histograms of key audio features

Shows how three main audio features-danceability, energy, and valence—are spread out in the Spotify dataset. These features are shown as percentages, from 0 to 100, and they show important parts of a track's acoustic profile.

Danceability: The histogram shows that most of the tracks are between 60% and 80%. This means that most of the songs in the dataset are moderately to very danceable, which is in line with what most people like in music that gets them moving.

Energy: The energy is mostly between 50% and 80%, and the distribution is tilted to the right. This pattern shows that popular songs in 2023 are more likely to feature a combination of loud and soft sounds than very loud or very soft sounds.

Valence: Valence, which shows how positive or emotionally bright a track is, is spread out across the spectrum, but there are peaks around 40–60%. This means that Spotify's 2023 catalog has songs that are both emotionally neutral and moderately positive, showing that there is a range of moods and musical styles.

In general, the histograms show that danceability and energy have clustered distributions around higher mid-range values, while valence has a wider spread. This means that the emotional tone of popular songs is very different, even though the rhythmic and energetic qualities are pretty much the same. These kinds of insights are helpful for clustering analysis because they point out things that are likely to affect how songs naturally group together.

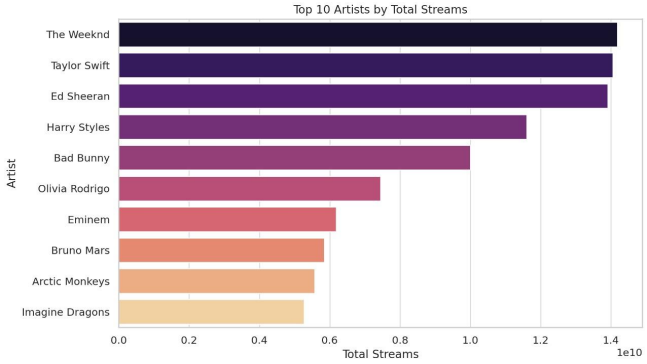


Fig: Top 10 Artists by Total Streams

Lists the top 10 artists by total streams. The Weeknd and Taylor Swift are at the top of the chart, followed by Ed Sheeran and Harry Styles. This shows how popular they are around the world. The rise of Latin music is shown by Bad Bunny's presence, while well-known artists like Eminem and Imagine Dragons still have a lot of power. This means that only a few artists get most of the streaming activity.

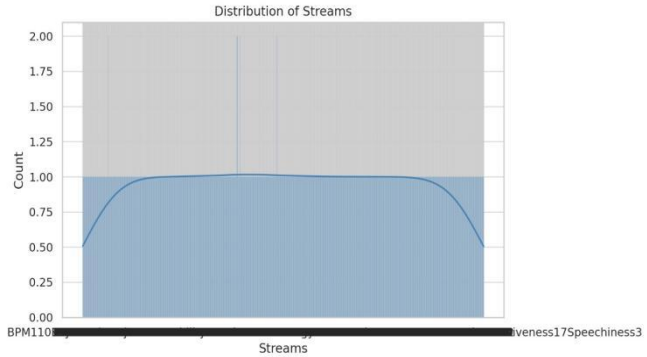


Fig: Distribution of Streams

Shows how streams are spread out across tracks. The plot shows a very uneven distribution, with most songs getting only a few streams and a very small number getting a lot of streams. This proves that there is a long-tail effect in music streaming, where only a small number of songs are listened to the most around the world.

V. Preprocessing and Feature Selection

There were varying numbers of audio features in Spotify 2023 songs in the data.We took only numeric features into account as clustering is efficient only with numeric data.

1. Feature Selection

- We only used features that are numerics, but not artist names, song names, and category metadata.
- We only used numeric values for audio features like danceability, energy, loudness, pace, and valence while creating our results.

2. Handling Missing Values

- AWe used median imputation to fill in any missing numeric values in the data set. The process is protected against factors that are unsuitable.

3. Scaling

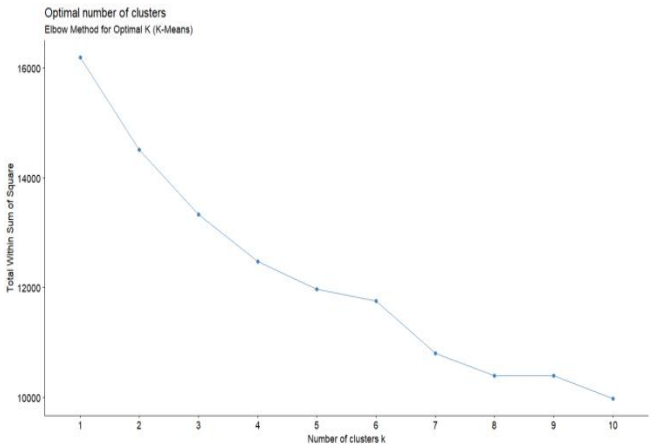
- We applied z-score scaling so that all the features would make the same contribution to distance calculations.
- This was due to the fact that audio features are measured in different scales, for example, beat per minute for tempo and dB for loudness.

Following preprocessing, the data possessed the same number of rows as before but now contained numeric features standardized and ready for clustering.

VI. Model Construction

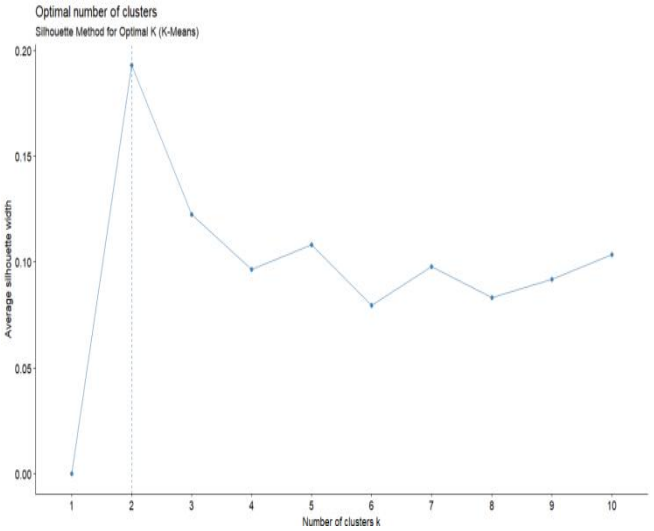
1. Elbow Method Plot

- We plotted a graph of the Within-Cluster Sum of Squares (WSS) against the number of clusters.
- For K = 2, there was an apparent "elbow," indicating that two clusters were the optimal number to have for keeping the model basic but effective in variance reduction.



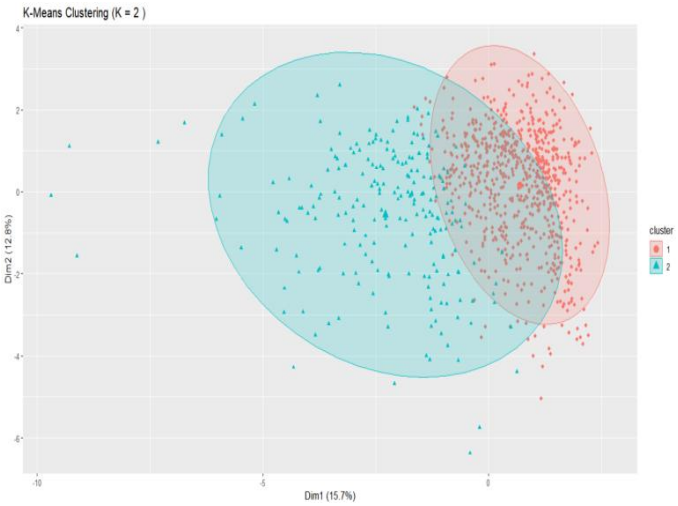
2. Silhouette Analysis Plot

- We calculated the average silhouette width for K ranging from 2 to 10.
- The silhouette value was greatest when K = 2, i.e., two clusters are optimal to be separated.
- But the values were not ridiculously high and they showed that there were clusters, but not the best.



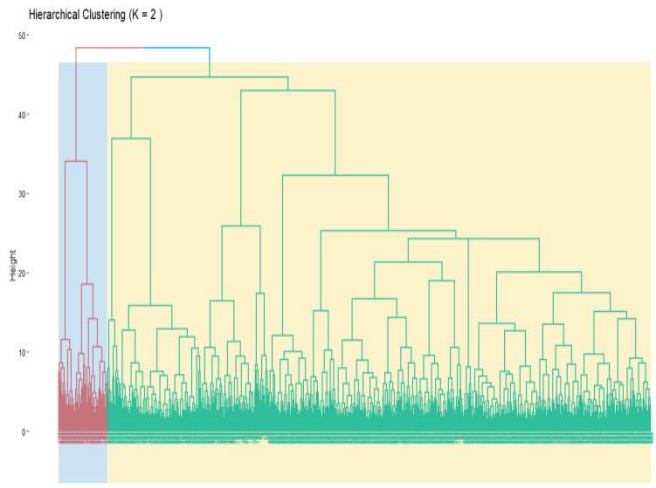
Two unsupervised clustering algorithms were applied to the preprocessed data:

1. K-Means Clustering



- We employed K-Means since it is efficient and can cluster large datasets into effective clusters.
- We employed Elbow Method and Silhouette Analysis for determining the optimal number of clusters (K).
- For additional reliability of the outcome, the process was initiated with 50 random starts (nstart = 50) initially.
- The K-Means final model placed each song into some clusters, which were plotted in scatterplots.

2. Hierarchical Clustering



- We conducted agglomerative hierarchical clustering with Ward's linkage and Euclidean distance.
- We built a dendrogram to display the cluster structure.
- Silhouette analysis indicated that the best number would be K = 2, and the tree was trimmed at that level. 3–5 sub-clusters were also identified nested within the main cluster for further comprehension.

VII. Performance and Evaluation of Models

We both checked the models' numbers and performances of the clustering by observing the difference how well the clusters were separated.

1. Silhouette Coefficient:

- Silhouette Coefficient was utilized as a significant measure to confirm the cohesion and separation of the clusters.
- It is a point of measure in a cluster (cohesiveness) similarity and between different clusters (separation) dissimilarity.
- When K = 2, the silhouette value will be highest, indicating two clusters best fit to explain the data structure. That was the case when K = 2 to 10.
- The silhouette scores showed that there was moderate clustering present and confirmed the existence of well-defined clusters and again confirmed that there was considerable overlap between them.

2. Elbow Method (Within-Cluster Sum of Squares):

- We applied the Elbow Method to find where there is more growth in the number of clusters that only gives us a small reduction in variance.
- We experienced the "elbow" at K = 2, which verified the signal from the silhouette.

3. Visual Evaluation:

- When the K-Means clustering was also graphed in lower dimensions, using scatterplots two groups were very clearly illustrated which were very easily distinguishable, albeit with some overlap.
- Hierarchical dendrogram gave some interesting information by showing two large branches and possible sub-clusters (3–5) in the large group.

Both techniques always revealed that Spotify 2023 dataset is indeed encompassed by two broad clusters. Low Silhouette values, however, mean that the dataset is poorly separated into clusters. This is very typical as musical content of songs overlaps naturally.

VIII. Results and Discussion

Silhouette Coefficient as well as Elbow Method suggested that the best number of clusters (K) of Spotify 2023 dataset was 2.

K-Means separated the data into two broad categories.
The silhouette score was such that the categories were somewhat different from each other, i.e., that they might possibly be separated out individually, although there was a bit of overlap between the two.
The image depicted this notion by showing clear high-level groupings, but imprecise boundaries.
Hierarchical clustering supported this finding and found two broad clusters.
The dendrogram gave us more structural information, though. It signaled the possibility of having 3–5 groups within the entire group, all of the same musical style or listener liking.

The findings show that 2023 Spotify trends would presumably consist of two broader categories.
There was one group consisting of compositions of more alike-sounding music, and the other group consisting of compositions of more different musical traits.
Even though the division is not perfect, the outcomes are beneficial for playlist and recommendation system building where even small groups can be beneficial towards more specific ends.

The highly graded silhouette scores also correspond to the reality that musical attributes do overlap freely.

It is questionable whether songs might easily be divided up since they overlap across several styles and moods.

It is also how human beings communicate in daily life, so that style differences harmoniously mix.

The research points out that the application of clustering can be used to reveal hidden patterns in music data and states that the only dependence on sound features in seeking accurate categorization has its own weaknesses.

IX. Conclusion

This research proved unsupervised clustering methods for audio data interest in Spotify. Hierarchical clustering and K-Means always reflected that there were two major groups of music. Results indicate that clustering is applicable to discover new music, construct playlists, and create recommendation systems. We are able to utilize dimensionality reduction (PCA) in the future such that we are able to clearly differentiate between the clusters. We can add non-audio metadata like listening history by a user or popularity information to learn more.

X. References

[1] N. Rohman and A. Wibowo, "Clustering of popular Spotify songs in 2023 using K-Means method and Silhouette Coefficient," *Pilar*, vol. 20, no. 1, pp. 18–25, Mar. 2024, doi: 10.33480/pilar.v20i1.4937.

[2] X. Li, "Analysis of machine learning-based music recommendation system using Spotify datasets," in *Proc. 2nd Int. Conf. Software Eng. Mach. Learn.*, 2024, pp. 49–55, doi: 10.54254/2755-2721/77/20240650.

[3] P. L. Kumar, "Clustering Spotify songs into moods using Thayer's model with a mood prediction and enhancer recommender system," *Figshare Preprint*, 2024.

[4] S. Mukhopadhyay, A. Kumar, D. Parashar, and M. Singh, "Enhanced music recommendation systems: A comparative study of content-based filtering and K-Means clustering approaches," *Revue d'Intelligence Artificielle*, vol. 38, no. 1, pp. 365–376, Feb. 2024, doi: 10.18280/ria.380138.

[5] M. Pichl, E. Zangerle, and G. Specht, "Understanding user-curated playlists on Spotify: A machine learning approach," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 1, pp. 67–85, 2017, doi: 10.4018/IJSWIS.2017010104.