Unsupervised Machine Learning

Creating playlists on Moosic:

an analysis involving music, data and machine learning

Critical questions of the project

- **Are the given audio features able to identify “similar songs”, as defined by humanly detectable criteria?**

- **Is K-Means a good method to create playlists? Would you stick with this algorithm moving forward, or explore other methods to create playlists?**

Data preparation

Features dropped:
Nan
Type

Audio perspective
Key - Mode - Tempo - Duration_ms
-Time_signature

Variance and Correlation
Speechiness - Loudness

Features used:
Danceability
Energy
Acousticness
Instrumentalness
Liveness
Valence

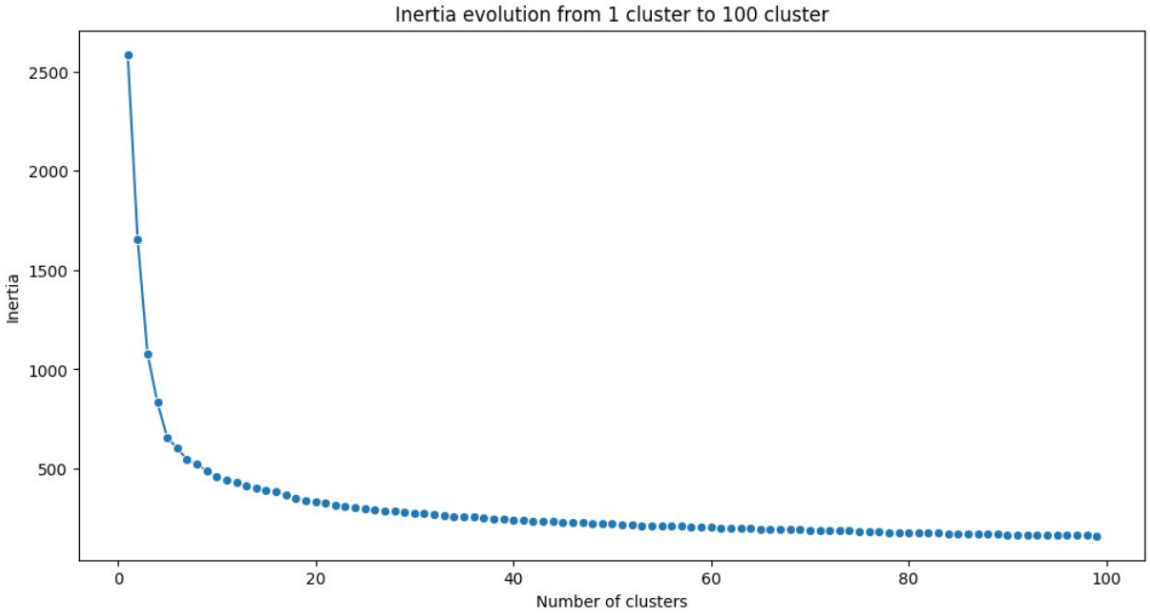
Modelling using the K-Means

- We decided in regards to the features which are left that there is no need to scale because all values have been already in the range of 0 and 1.
- 2 methods to calculate the number of clusters:
  - Elbow method
  - Silhouette coefficient

Modelling - The Elbow method

This the elbow from cluster 1 to 100. It seems that clustering with 6 clusters seems quite good.

But the condition which was set by moosic has been creating a playlist out of a dataset of 5000 songs with a minimum of 50 songs to 250 songs maximum.

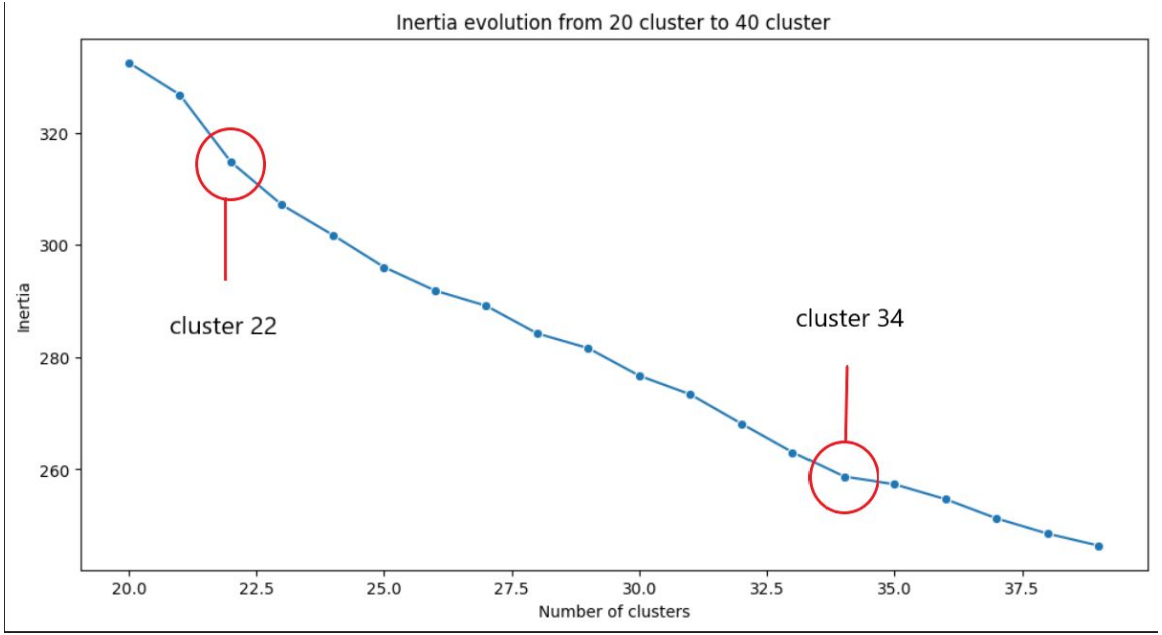


Modelling - The Elbow method

So we took 20 to 100 clusters in consideration.

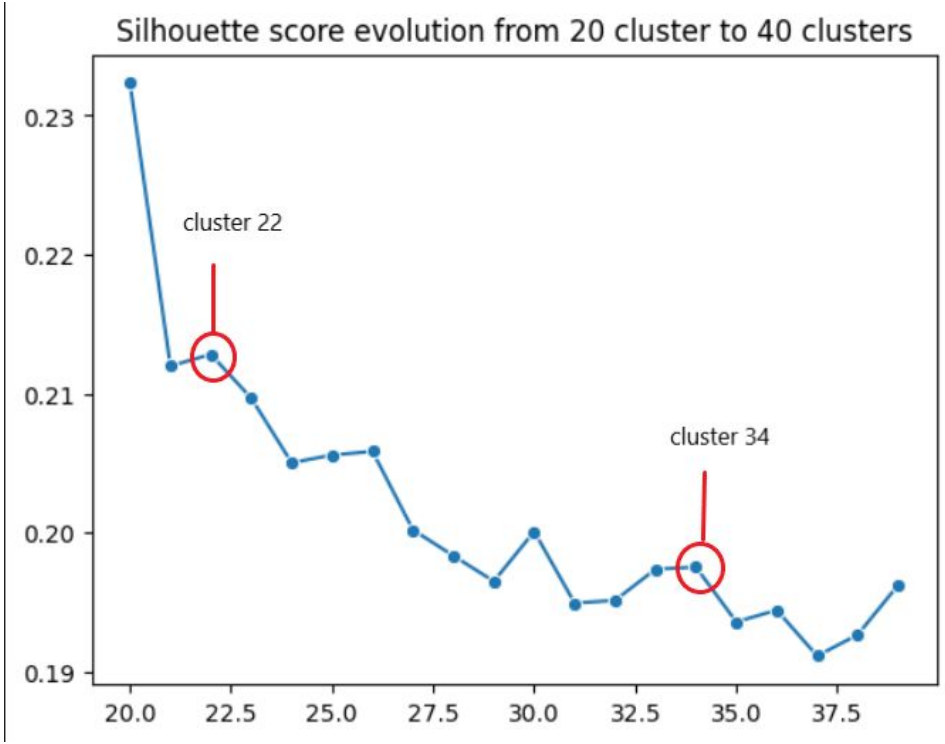
This the elbow from cluster 20 to 40.

It seems that clustering with 22 or 34 clusters seems quite good.



Modelling - The Silhouette Coefficient

Comparing the clusters we got from the elbow method (20 to 40 clusters) with the same amount of clusters but using the silhouette coefficient shows that 22 or 34 clusters seem to work as well.



The final model after implementing into the Spotify playlist (5 samples using 22 clusters)

WORLD MUSIC

Everything fine?

Well...

The final model after implementing into the Spotify playlist (5 samples using 22 clusters)

METAL: Looking at the playlist and the subgenres in detail:

Therion has been playing Death Metal but is nowadays Symphonic Metal (Sweden)

Amorphis is Death Metal (Finnland)

Disincarnate is Death Metal (USA)

Cryptopsy is Technical Death Metal/Metalcore (Canada)

Deceased has been Death metal but is now Thrash Metal/Heavy Metal (USA)

Conclusions

K-means is a good way to categorize unlabeled data as shown in our findings but it could be implemented better:

Optimize Cluster Size for Precision: Using a smaller number of clusters can enhance the precision of cluster definitions.

Incorporate Rich Listener Preferences: Enhance clustering accuracy by including features related to listener preferences, such as genres, subgenres, time periods, analysis of lyrics and country of origin.

Leverage Hybrid Clustering Approaches: Combining K-means with techniques like Hierarchical Clustering or DBSCAN can yield finer results.

Questions?

Thank you for your attention. We are now ready to answer any questions you may have.

The Moosic Machine Learning Team:

Gustavo
Saddam
Mirella