A logo with wings and a letter

Description automatically generated

*ACADEMIA DE STUDII ECONOMICE*

*Facultatea de Cibernetică, Statistică şi Informatică Economică*

*Specializarea: Informatică economică*

*PROIECT DEZVOLTARE SOFTWARE PENTRU ANALIZA DATELOR*

*Analiza cluster si Analiza PCA pentru top 50 de melodii de pe Spotify in anul 2023*

***Cadrul didactic coordonator:***

*Cadrul didactic asociat*

*OBRETIN ALEXANDRU-MARIUS*

***Student:***

***Ariton Alexandru***

*Grupa 1086, Seria C, AN III*

**Bucureşti,**

**2024**

***Listă de Figuri***

Figura 1 Dendograma optima pentru cele 50 de melodii

Figura 2 Graficul Biplot pentru Analiza in Componente Principale

Figura 3 Graficul Scree pentru Analiza in Componente Principale

Descrierea variabilelor

1. track\_name: Numele piesei sau melodiei.
2. artist(s)\_name: Numele artistului sau al artistilor care au realizat piesa.
3. in\_spotify\_playlists: Numarul de playlist-uri Spotify in care apare piesa.
4. in\_spotify\_charts: Prezenta piesei in topurile Spotify.
5. streams: Numarul total de stream-uri pe care piesa le-a acumulat pe Spotify.
6. in\_apple\_playlists: Numarul de playlist-uri Apple Music in care apare piesa.
7. in\_apple\_charts: Prezenta piesei in topurile Apple Music.
8. in\_deezer\_playlists: Numarul de playlist-uri Deezer in care apare piesa.
9. in\_deezer\_charts: Prezenta piesei in topurile Deezer.
10. in\_shazam\_charts: Prezenta piesei in topurile Shazam.
11. danceability\_%: Un scor procentual care reflecta cat de adecvata este o piesa pentru dans, bazat pe o combinatie de elemente muzicale.
12. energy\_%: Un scor procentual care masoara intensitatea si activitatea unei piese.
13. acousticness\_%: Un scor procentual care estimeaza cat de acustica este o piesa.
14. liveness\_%: Un scor procentual care masoara prezenta publicului in inregistrare.

Entitatile supuse observatiei

1. in\_spotify\_playlists
2. in\_spotify\_charts
3. streams
4. in\_apple\_playlists
5. in\_apple\_charts
6. in\_deezer\_playlists
7. in\_deezer\_charts
8. in\_shazam\_charts
9. danceability\_%
10. energy\_%
11. acousticness\_%
12. liveness\_%

Motivatia de a opta pentru o anumita metoda de analiza a datelor

In cazul utilizarii analizei componentelor principale (PCA), motivul este pentru ca faciliteaza vizualizarea si interpretarea datelor fara a pierde multa informatie semnificativa.

Descrierea observatiilor

Observatiile sunt reprezentate de top 50 de melodii de pe Spotify in anul 2023, impreuna cu artistii care le-au compus.

Abordarea privind Analiza Datelor

1. Curatarea si Pregatirea Datelor

* Verificarea si tratarea valorilor lipsa: M-am asigurat ca toate valorile lipsa sunt tratate corespunzator, prin eliminarea inregistrarilor respective.
* Conversia datelor categorice: Variabilele categorice, cum ar fi numele pieselor si ale artistilor, au necesitat codificare one-hot sau etichetare pentru analize numerice ulterioare.
* Normalizarea: Datele numerice au fost normalizate pentru analizele PCA si algoritmul de clusterizare.

1. Reducerea Dimensiunii

* Am aplicat PCA pentru a reduce dimensionalitatea datelor si pentru a identifica componentele care explica cea mai mare parte a variantei.

1. Clusterizare

* Am utilizat clusterizarea ierarhica pentru a gasi grupuri naturale in date si pentru a vizualiza aceste grupuri printr-o dendrograma.
* Am folosit scorul Silhouette pentru a evalua calitatea clusterizarii si pentru a determina numarul optim de clustere.

1. Analiza si Interpretarea Rezultatelor

* Am interpretat clusterele identificand caracteristicile predominante ale pieselor din fiecare cluster.

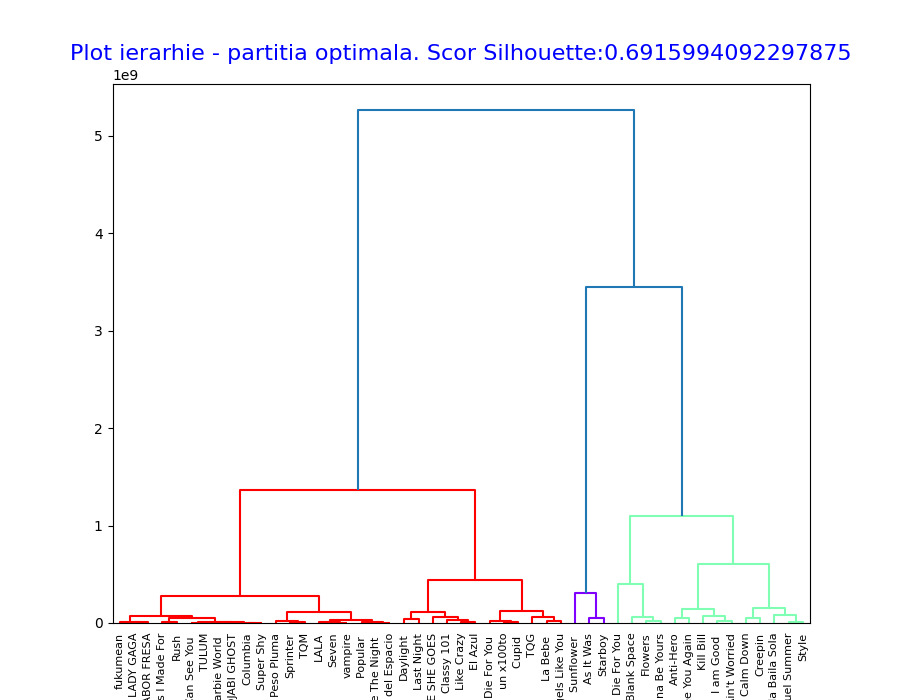


Figura 1 Dendograma optima pentru cele 50 de melodii

Dendrograma de mai sus reprezinta rezultatul unei analize de clustering ierarhic efectuata pe setul de date muzicale. Acest tip de grafic arata modul in care diferitele piese sunt grupate impreuna in clusteri bazate pe similaritatile lor.

Clusterul Reprezentat de Culoarea Rosie

Sa presupunem ca clusterul rosu include piesele "Seven", "LALA", "vampire", si "Cruel Summer". Aceste piese au numere mari de stream-uri si apar intr-un numar semnificativ de playlist-uri pe Spotify si Apple Music, indicand o popularitate inalta. Putem presupune ca factorii cheie pentru gruparea acestor piese in acelasi cluster includ:

* Prezenta in Playlist-uri si Clasamente: Un numar mare de aparitii in playlist-uri si pozitii inalte in clasamente sugereaza ca piesele sunt populare si bine primite de publicul larg.
* Numarul de Stream-uri: Piesele cu un numar mare de stream-uri indica o audienta larga si o popularitate sustinuta.
* Distributia pe Platforme de Streaming: Prezenta consistenta pe multiple platforme de streaming poate fi un indicator al unei popularitati globale si al unui apel universal.

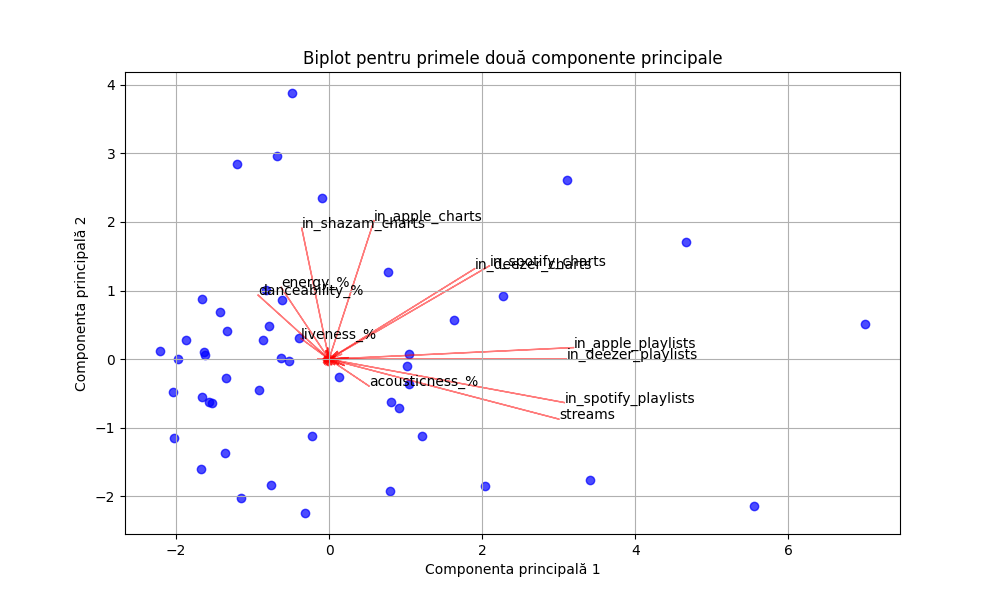
Clusterul Reprezentat de Culoarea Verde

Presupunand ca clusterul verde include piesele "Flowers", "Daylight", "As It Was", si "Kill Bill", Acest cluster ar putea fi caracterizat de similaritati in trasaturile muzicale, cum ar fi:

* Dansabilitate si Energie: Niveluri similare ale dansabilitatii si energiei pot indica faptul ca aceste piese sunt preferate pentru anumite contexte, cum ar fi cluburi sau petreceri.
* Atributul Acousticness: O masura a acusticii care sugereaza un stil mai natural sau mai putin electronic, posibil asociat cu un gen specific cum ar fi baladele sau muzica acustica.
* Artistul sau Genul Muzical: Colaborarile intre anumiti artisti sau apartenenta la un anumit gen muzical pot fi, de asemenea, factori care contribuie la formarea acestui cluster. Scorul Silhouette de 0,69 indica o calitate buna a clustering-ului, sugerand ca piesele din fiecare cluster sunt relativ similare intre ele si destul de diferite de piesele din alti clusteri.

Relatiile dintre variabilele initiale si elementele rezultate in urma aplicarii algoritmului PCA:

In cazul setului de date despre care discutam, variabilele initiale includ numarul de stream-uri, prezenta in playlist-uri si topuri de pe diverse platforme de streaming si caracteristici ale pieselor, cum ar fi dansabilitatea si energia. PCA le-ar transforma intr-un set de componente principale care ar putea reflecta diferite aspecte ale succesului si caracteristicilor muzicii, cum ar fi popularitatea, atractivitatea pentru ascultatori si stilurile muzicale.

Figura 2 Graficul Biplot pentru Analiza in Componente Principale

Interpretarea Biplot-ului:

* Punctele albastre reprezinta piesele individuale.
* Vectorii rosii indica modul in care caracteristicile muzicale (dansabilitate, energie, acustica, prezenta live) si prezenta in playlist-uri si topuri sunt reprezentate in spatiul bidimensional al componentelor principale.

Analiza Relatiilor:

* Prezenta in Playlist-uri si Topuri: Observam ca variabilele precum in\_spotify\_playlists, in\_apple\_playlists, streams, sunt relativ apropiate unele de altele, sugerand ca piesele care sunt populare pe una dintre platforme tind sa fie populare si pe celelalte.
* Caracteristicile Muzicale: Variabilele danceability\_%, energy\_%, si liveness\_% sunt orientate in aceeasi directie si sunt relativ apropiate, indicand o asociere intre aceste caracteristici ale pieselor. Acest lucru ar putea sugera ca piesele cu niveluri mai ridicate de dansabilitate, energie si prezenta live sunt mai susceptibile sa aiba un succes similar pe platformele muzicale.
* Acustica: acousticness\_% este orientat in directia opusa fata de energy\_% si danceability\_%, ceea ce poate indica faptul ca piesele cu o compozitie mai acustica au tendinta de a avea mai putina energie si dansabilitate.
* Corelatii: Daca sagețile sunt mai lungi, acest lucru indica o influenta mai mare asupra variatiei in date. De exemplu, daca streams are o sageata lunga si este orientata spre punctele albastre concentrat intr-o anumita zona, aceasta sugereaza o legatura puternica intre succesul comercial al pieselor si caracteristicile muzicale asociate cu acea directie.

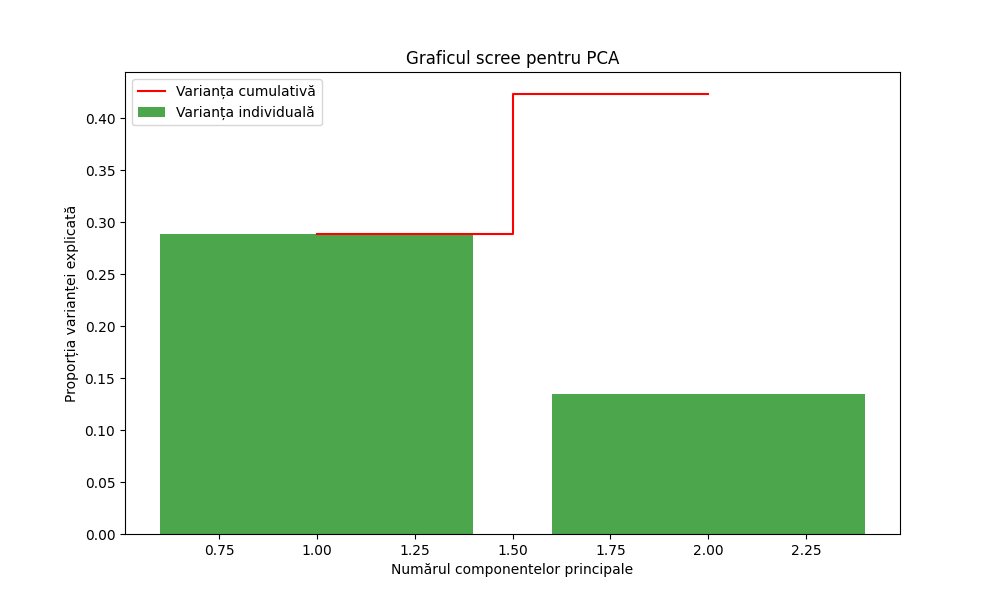


Figura 3 Graficul Scree pentru Analiza in Componente Principale

Interpretarea Scree-ului:

* Barele Verzi (Varianța Individuală): Fiecare bara verde reprezinta proportia variantei explicata de fiecare componenta principala individuala. In graficul tau, prima bara este mult mai inalta decat a doua, ceea ce indica faptul ca prima componenta principala explica o proportie mai mare din varianta totala a datelor.
* Linia Rosie (Varianța Cumulativa): Aceasta reprezinta suma progresiva a variantei explicate pe masura ce fiecare componenta principala este adaugata. Linia incepe la valoarea variantei explicata de prima componenta si creste pe masura ce este inclusa a doua componenta.

Observatii:

* Prima componenta principala este semnificativ mai importanta in termeni de varianta explicata comparativ cu a doua componenta, asa cum se vede din inaltimea primei bare verzi.
* Varianta cumulativa dupa includerea primelor doua componente principale pare sa fie de aproximativ 40% sau putin mai mult, ceea ce inseamna ca restul componentelor principale (nereprezentate in acest grafic) trebuie sa explique restul de 60% sau mai putin din varianta totala.

CONCLUZII

Analiza a evidențiat relațiile complexe dintre caracteristicile pieselor muzicale și succesul lor pe platformele de streaming. Prin reducerea dimensionalității și clusterizare, am câștigat insight-uri valoroase care pot ghida deciziile în industria muzicală. Totuși, este important să reținem că aceste metode de analiză pot dezvălui doar tendințe și asocieri, nu cauzalitate directă. Pentru a trage concluzii definitive, ar fi necesare analize suplimentare și, posibil, colectarea de date adiționale.

***BIBLIOGRAFIE***

* Cursuri „Analiza Datelor”, prof. univ. dr. Furtună Titus Felix
* Seminare „Analiza Datelor”, prof. univ. dr. Furtună Felix
* Sursa datelor: <https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023>