**Dezvoltare Software pentru Analiza Datelor**

**Analiza Componentelor Principale**

**și Analiza de Cluster**

Cadru didactic: dr. Vințe Claudiu

Studenti: Corodescu Serban, Custura Bogdan, Condrat Claudiu

Grupa 1087, seria C

Cuprins

[1. Prezentare generală 3](#_Toc156154107)

[2. Analiza componentelor principale 4](#_Toc156154108)

[2.1. Introducere 4](#_Toc156154109)

[2.2 Motivatia alegerii metodei de analiza ACP 5](#_Toc156154110)

[2.3 Prezentarea rezultatelor și interpretarea lor 5](#_Toc156154111)

[2.3.1 Analiza variabilității datelor 5](#_Toc156154112)

[3. Analiza de cluster 8](#_Toc156154113)

[3.1. Introducere 8](#_Toc156154114)

[3.2. Motivatia alegerii metodei de analiza HCA 8](#_Toc156154115)

[3.3. Prezentarea rezultatelor și interpretarea lor 9](#_Toc156154116)

[3.3.1 Clasificarea variabilelor 10](#_Toc156154117)

[Bibliografie: 11](#_Toc156154118)

# 1. Prezentare generală

Componentele pe care le-am selectat pentru realizarea acesteor analize sunt urmatoarele:

* **Rata somajului =** procentul de persoane șomere în raport cu forța de muncă totală.Indică nivelul de ocupare a forței de muncă și stabilitatea economică.
* **Forta de munca** = totalul persoanelor apte să lucreze, fie angajate, fie în căutarea unui loc de muncă.
* **PIB/Loc =** produsul Intern Brut (PIB) este împărțit la numărul total al locuitorilor unei țări.
* **Speranta de viata** = numărul mediu de ani pe care îi poate trăi o persoană într-o anumită țară.
* **Undicele de dezvoltare umana(IDU)**= concept utilizat pentru a măsura nivelul de

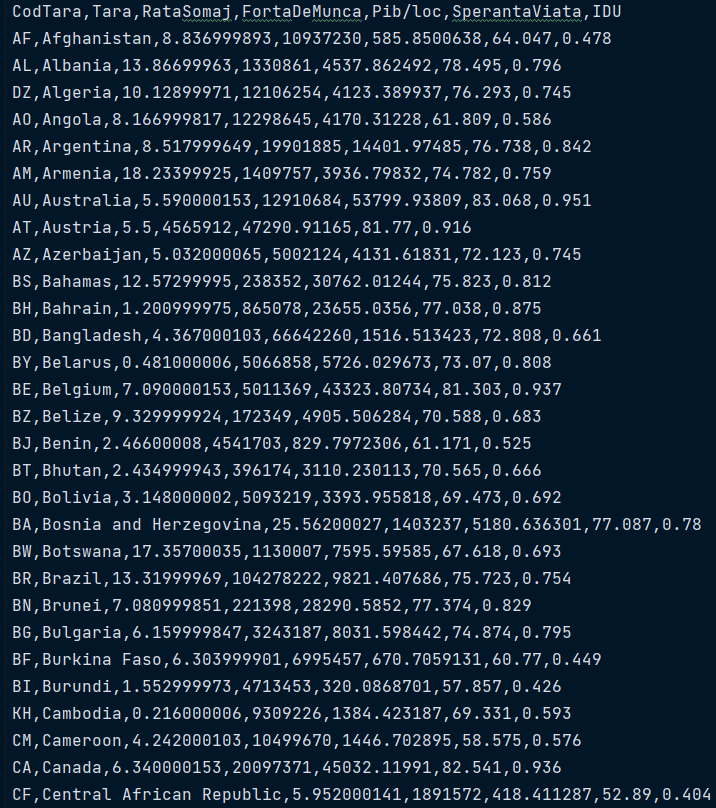
dezvoltare al unei țări.

Datele furnizate sunt din anul 2021, ele constituie informații esențiale privind aspecte socio-economice ale diverselor țări. Rata șomajului oferă o perspectivă asupra stării pieței muncii, evidențiind gradul de ocupare și posibilele provocări economice. Forța de muncă reflectă dimensiunea populației active, având implicații asupra capacității unei țări de a genera creștere economică.

PIB-ul per locuitor este un indicator cheie al standardului de viață, măsurând valoarea totală a bunurilor și serviciilor produse împărțită la nivelul populației. Speranța de viață este un element esențial în evaluarea calității vieții și a stării generale a sănătății populației.

Analizând aceste variabile în contextul fiecărei țări, se pot identifica tendințe, diferențe și relații specifice, oferind o privire asupra dinamicii sociale și economice. Este important să se țină cont de particularitățile fiecărei țări, precum contextul istoric, politic și economic, pentru o înțelegere completă a datelor. Totodată, trebuie avută în vedere dinamica schimbărilor și evoluțiilor ulterioare, deoarece situația socio-economică poate suferi modificări în timp.

IDU este calculat de Programul Națiunilor Unite pentru Dezvoltare (PNUD) și este rezultatul combinării a trei dimensiuni esențiale ale vieții umane: sănătatea, educația și nivelul de trai. Aceste trei componente ale IDU sunt reflectate în indicatorii respectivi: speranța de viață la naștere, durata medie a școlarizării și venitul național brut pe locuitor ajustat la paritatea puterii de cumpărare.

* Figura 1 – O parte din datele analizate*

# 2. Analiza componentelor principale

## 2.1. Introducere

Analiza componentelor principale reprezintă o tehnică multidimensională utilizată pentru a sintetiza informația conținută în variabilele inițiale printr-un set redus de variabile corelate între ele. Scopul este să se creeze noi componente care să fie necorelate, fiecare fiind o combinație liniară a variabilelor originale. Aceste componente, numite componente principale, sunt caracterizate de o variație maximă și sunt esențiale pentru reducerea dimensionalității spațiului inițial și eliminarea redundanței informaționale.

Obiectivul final al analizei componentelor principale este identificarea unor variabile noi care să exprime sintetic informația din variabilele inițiale, menținând o cantitate controlată de informație totală. Pentru a realiza această analiză, este crucială standardizarea datelor. Deoarece dimensiunile și unitățile de măsură pot varia între variabile, standardizarea constă în înlocuirea valorilor fiecărei observații cu un nou set de valori reprezentând raportul dintre valoarea centrată a respectivei observații și deviația standard a variabilei respective. Această operație este esențială pentru a asigura o analiză corespunzătoare a componentelor principale.

## 2.2 Motivatia alegerii metodei de analiza ACP

Am decis sa alegem acest tip de analiza pentru 2 considerente:

**Gestionarea Redundanței:**

* ACP ajută la eliminarea redundanței informaționale, păstrând totuși cea mai mare parte a variabilității. Aceasta face analiza mai eficientă și poate duce la modele mai robuste.

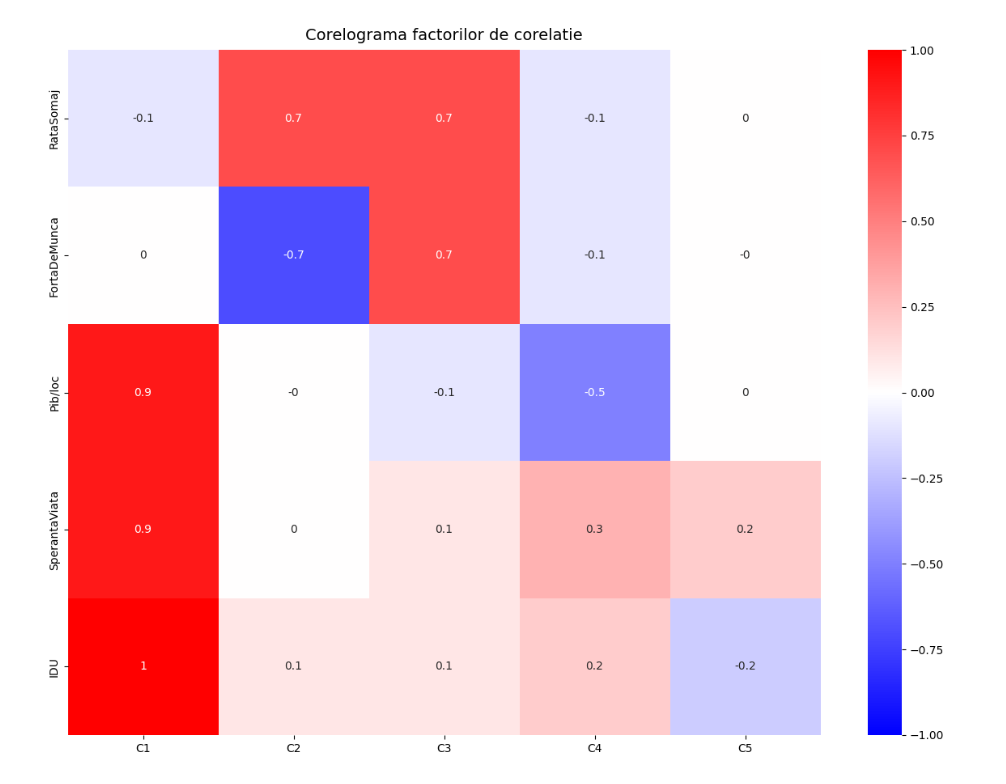
**Reducerea Dimensionalității:**

* In momentul in care se lucreaza cu seturi mari de date sau cu un număr mare de variabile, ACP ajută la reducerea dimensiunii spațiului de date. Astfel, se pot păstra doar acele componente care capturează cea mai mare variație a informației inițiale, eliminând variabilele redundante sau mai puțin semnificative.

## 2.3 Prezentarea rezultatelor și interpretarea lor

### 2.3.1 Analiza variabilității datelor

Pentru a realiza analiza în componente principale (ACP) în mod programatic, am folosit codul dezvoltat în cadrul seminariilor, care include o clasă de model pentru ACP, fișierul cu funcții pentru grafice și un fișier de comandă principal, main*ACP.py*. În proiect, am inclus și fișierul TariPtAnaliza.csv, acesta conținând datele de intrare.



*Figura 2 - Corelograma*

Analiza în componente principale începe prin determinarea vectorilor și valorilor proprii ale matricei de corelație. Valorile proprii reflectă varianțele asociate componentelor, în timp ce vectorii proprii indică coeficienții legăturii liniare dintre variabilele observate și componentele principale.

În figura 2, tabelul prezintă calculul varianței pentru cele 5 componente, oferind o perspectivă asupra contribuției fiecărei componente la variația totală a datelor.

*A graph with a line and a line

Description automatically generated*

*Figura 3 - Graficul de valori*

Conform criteriilor Cattel și Kaiser, sunt semnificative doar primele 2 componente, deoarece au o varianță mai mare decât 1, lucru evidențiat și de graficul din figura 3.

# 3. **Analiza de cluster**

## 3.1. Introducere

Pentru a determina grupările naturale ale datelor sau pentru an oferi o împărțire convenabilă a datelor în grupuri, sunt utilizate algoritmi de clusterizare sau clasificare nesupervizată. Clasificarea nesupervizată nu folosește informații prealabile despre grupurile care vor fi clasificate, spre deosebire de clasificarea supervizată, așa cum este cazul analizei discriminante.

Datele care sunt utilizate pentru analiză sunt valori ale relațiilor dintre oameni și variabilele studiate, luate în două moduri diferite: distanțe sau disimilarități. Matricele de distanță sunt utilizate pentru înregistrarea acestor valori. Din acest motiv, valorile care urmează să fie analizate sunt deja rezultatul unor calcule care au determinat măsurători de disimilaritate între diferite obiecte, fie ele variabile sau instanțe. Grupurile clasificate conțin oameni care sunt foarte asemănători și au doar o mică diferență.

## 3.2. Motivatia alegerii metodei de analiza HCA

Am decis sa alegem acest tip de analiza deoarece:

**Contribuie la imbunătățirea proceselor de marketing și vânzare:**

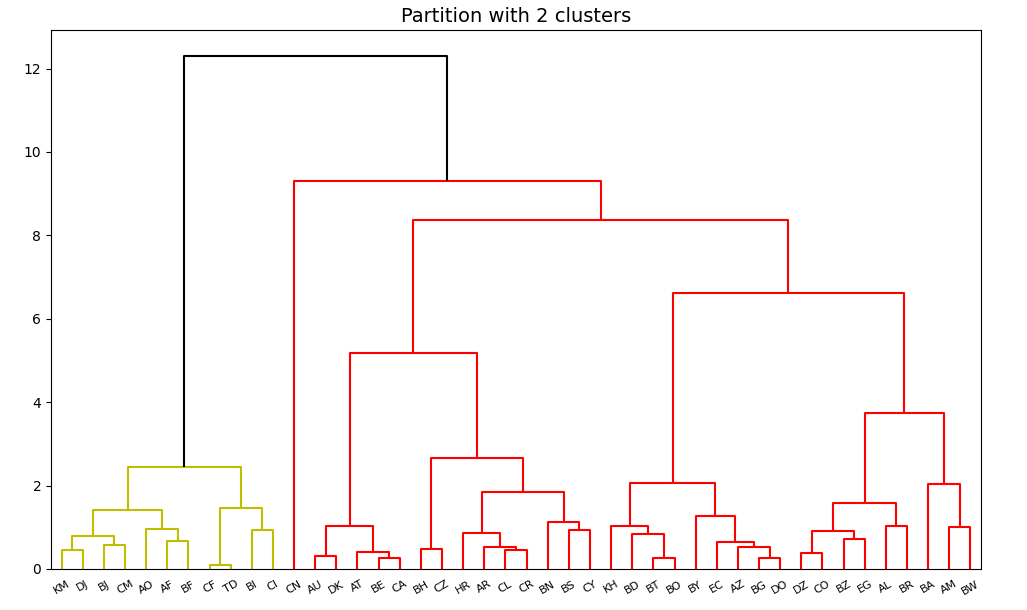
* În domeniul afacerilor, analiza de cluster poate fi folosită pentru a identifica segmentele de piață și pentru a personaliza strategiile de marketing și vânzare în funcție de nevoile și preferințele specifice ale fiecărui segment.

**Ajuta la validarea hipotezelor:**

* Analiza de cluster poate fi utilizată pentru a valida sau infirma ipotezele existente privind structura sau segmentarea datelor. Acest aspect contribuie la consolidarea fundației teoretice a studiului.

## 3.3. Prezentarea rezultatelor și interpretarea lor

Conform dendogramei din figura de mai jos, este posibil să se separe al doilea cluster în două. Acest lucru ar produce o nouă dendogramă cu trei clustere, care ar oferi o împărțire mai precisă a datelor în cele trei grupe:



*Figura 4 – Partitia oprimala*

### 3.3.1 Clasificarea variabilelor

Analiza de clusteri poate fi folosită și pentru gruparea variabilelor, dacă sunt alese metricile potrivite. Este de obicei folosită pentru clasificarea instanțelor. Pentru aceasta clasificare de variabile, o metodă de analiză este metoda medie, care utilizează metrica de corelație și produce două clustere. Dendograma clasificării variabilelor este ilustrată în figura de mai jos:

Așa cum se poate vedea, rata somajului și forta de munca formează un cluster, iar PIB/loc este grupat cu restul indicatorilor, si anume speranta de viata și indicele de dezolvtare umana care formează un alt cluster. De asemenea, este evident că aceste clustere corespund elementelor principale care au fost identificate în prima analiză.

*A diagram with blue and orange squares

Description automatically generated*

*Figura 5 – Dendograma variabilelor*

# Bibliografie:

Cursuri – prof. Furtuna Titus Felix

Laboratoare – prof Vinte Claudiu

<https://hdr.undp.org/data-center/documentation-and-downloads>

<https://hdr.undp.org/data-center/documentation-and-downloads>

<https://data.worldbank.org/indicator/NY.GDP.PCAP.CD>

<https://data.worldbank.org/indicator/SL.TLF.TOTL.IN>

https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS