**ACADEMIA DE STUDII ECONOMICE, BUCUREȘTI**

**FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**

**SPECIALIZAREA – INFORMATICĂ ECONOMICĂ**

**AN DE STUDIU: III GRUPA: 1098 SERIA: E**

O imagine care conține siglă, Grafică, Font, design grafic

Descriere generată automat

**ANALIZA NUMĂRULUI DE DECESE PE FACTORI DE RISC LA NIVELUL ȚĂRILOR LUMII**

**Coordonator științific**

Conf. univ. dr. Claudiu VINȚE

**Student**

Liviu-Ioan ZECHERU

Gabriel ZĂVOIANU

Andrei-Constantin-Ștefan VASILESCU

**BUCUREȘTI**

**2023**

Cuprins

[Sursa de date 2](#_Toc156056402)

[Referință de timp privind datele 2](#_Toc156056403)

[Descrierea variabilelor 2](#_Toc156056404)

[Descrierea observațiilor 3](#_Toc156056405)

[Abordarea folosind analiza datelor 3](#_Toc156056406)

[Motivația de a opta pentru o anumită metodă de analiză a datelor 6](#_Toc156056407)

[Prezentarea rezultatelor 7](#_Toc156056408)

[ACP 7](#_Toc156056409)

[AEF 8](#_Toc156056410)

[Interpretare 9](#_Toc156056411)

[ACP 9](#_Toc156056412)

[AEF 10](#_Toc156056413)

[Discuții și concluzii 11](#_Toc156056414)

# Sursa de date

Am ales să folosim un set de date pus la dispoziție de un utilizator de pe platforma Kaggle. Acest set de date poate fi accesat la adresa de [aici](https://www.kaggle.com/datasets/varpit94/worldwide-deaths-by-risk-factors/).

# Referință de timp privind datele

Am curățat setul de date astfel încât să-l transformăm din serii cronologice în date nestructurate, alegând anul 2017.

# Descrierea variabilelor

Variabilele setului de date sunt factori ce au dus la deces în rândul populației. Vom enumera fiecare factor cu denumirea sa aferentă în setul de date.

* sursă îndoielnică de apă potabilă (Unsafe water source)
* salubrizare necorespunzătoare (Unsafe sanitation)
* lipsă de acces la spălatul pe mâini (No acces to handwashing facility)
* poluarea aerului la nivelul gospodăriei cauzată de combustibili solizi (Household air pollution from solid fuels)
* alăptare la comun (Non-exclusive breastfeeding)
* discontinuitatea alăptării (Discontinued breastfeeding)
* malnutriție low weight-for-height (Child wasting)
* malnutriție low height-for-age (Child stunting)
* hipogreutate pentru perioada gestațională (Low birth weight for gestation)
* fumatul pasiv (Secondhand smoke)
* consumul de alcool (Alcohol use)
* consumul de droguri (Drug use)
* dietă săracă în fructe (Diet low in fruits)
* dietă săracă în legume (Diet low in vegetables)
* sex neprotejat (Unsafe sex)
* sedentarism (Low physical activity)
* nivelul de glucoză la nivelul jejunului (High fasting plasma glucose)
* nivelul total de colesterol (High total cholesterol)
* BMI mare (High body-mass index)
* presiune sistolică mare (High systolic blood pressure)
* fumat (Smoking)
* deficiență de fier (Iron deficiency)
* deficiență de vitamina A (Vitamin A deficiency)
* demineralizare osoasă (Low bone mineral density)
* poluarea aerului în interior (Air pollution)
* poluarea aerului în exterior (Outdoor air pollution)
* dietă bogată în săruri (Diet high in sodium)
* dietă săracă în cereale integrale (Diet low in whole grains)
* dietă săracă în nuci și semințe (Diet low in nuts and seeds)

# Descrierea observațiilor

Observațiile reprezintă numărul total de decese (estimate, de aceea sunt în virgulă mobilă) pe factorii descriși mai sus la nivel de țară sau entitate pentru anul 2017.

# Abordarea folosind analiza datelor

În cadrul ACP (analizei în componente principale), așa cum am menționat și în capitolul de motivație a alegerii metodei, se urmărește reducerea dimensionalității spațiului caracteristic. Motivația acestei reduceri este că există anumite caracteristici care captează un procent mare de variație în setul de date original. Astfel, este important să se găsească „direcțiile” de variație maximă în setul de date. Aceste direcții se numesc componente principale. Iar analiza în componente principale este, în esență, o proiecție a setului de date pe caracteristicile esențiale. În continuare voi explica procedeul de determinare al componentelor principale prin două metode. Prima metodă se numește descompunerea în valori proprii, iar a doua descompunerea valorii singulare.

Presupunând că avem o matrice de date *X* de dimensiune , efectuăm o descompunere a valorilor proprii pe matricea de covarianță a lui *X*. Dacă toate caracteristicile sunt standardizate (medie 0 și dispersie 1), atunci matricea de covarianță este dată de . Dacă caracteristicile nu sunt standardizate, putem scădea media coloanei *i* din fiecare intrare din coloana respectivă și putem calcula matricea de covarianță. Este simplu de observat că matricea de covarianță este o matrice pătratică de ordin *număr\_caracteristici*.

Primele *k* componente principale sunt, prin urmare, vectorii proprii corespunzători celor mai mari *k* valori proprii. Deoarece matricea de covarianță este simetrică și semi-pozitiv definită, descompunerea poate fi scrisă ca:

unde *D* este matricea vectorilor proprii, iar *Λ* este matricea diagonală a valorilor proprii.

A doua tehnică de factorizare a matricei *X* care poate fi utilizată pentru a calcula componentele principale este descompunerea valorii singulare.

Această metodă este definită pentru toate matricele și se realizează în felul următor: dată fiind o matrice *X*, descompunerea valorii singulare a lui *X* va fi egală cu , unde *U*, *Σ* și *V* sunt matricele vectorilor singulari stânga, valorilor singulare și, respectiv, vectorilor singulari dreapta.

Prin urmare, descompunerea singulară a matricei de covarianță a lui *X* este dată de:

Există algoritmi implementați eficient pentru a calcula această descompunere a unei matrice. Pachetul scikit-learn folosește o altfel de implementare pentru a determina componentele principale, pachet pe care îl vom folosi și noi în analiza noastră.

Am folosit AEF (analiza exploratorie a factorilor) pentru a identifica factorii cheie care contribuie la decesul în rândul populației. Prin examinarea acestor factori, se poate dezvălui influența semnificativă a anumitor aspecte asupra stării de sănătate și a ratei de mortalitate.

Aspectele generale ale acestui model și pașii pe care îi avem de parcurs sunt:

1. Găsirea numărului de factori:

Este foarte important să alegem numărul corect de factori, deoarece prea puțini factori produc prea multe încărcături mari, în timp ce prea mulți factori duc la un model fragmentat.

1. Rotire:

Rotirea este un proces care permite analiștilor să facă o soluție FA mai ușor de interpretat. Orice rotire are un obiectiv. Rotirile ortogonale cer ca factorii să nu fie corelați. Scopul rotirilor ortogonale este sporirea generalizării și simplitatea. În contrast, rotirile oblice permit factorilor să fie corelați și scopul lor este să obțină cea mai bună potrivire.

1. Regulile Rotirii:

* Interpretarea este mai ușoară cu încărcături puternice;
* Fiecare rând din matricea factorilor trebuie să conțină cel puțin un zero;
* Fiecare coloană trebuie să conțină cel puțin zerouri;
* Fiecare pereche de coloane a matricei trebuie să aibă variabile ale căror încărcături sunt puternice în una, dar dispar într-o altă coloană;
* Dacă numărul de factori este mai mare de patru, fiecare pereche de coloane trebuie să aibă mai multe variabile cu încărcături zero în perechi specifice de coloane;
* În fiecare pereche de coloane, câteva variabile ar trebui să aibă încărcături non-zero.

1. Tehnici de Rotire:

* Rotirea Ortogonală: Avantajul major constă în simplitate - Încărcările sunt corelații între factori și caracteristicile observate.
* Rotirea Oblică: Mai complexă - Coeficienții modelului factorial sunt coeficienții de regresie care pot fi utilizați pentru a crea caracteristici observabile prin înmulțirea factorilor cu coeficienții modelului factorial. Coeficienții de structură ai factorilor sunt coeficienți de corelație între factori și caracteristicile observate.

În cazul nostru, vom adopta metoda VariMax pentru rotirea ortogonală. VariMax are ca rezultat puține încărcături mari și multe aproape de zero.

Abordarea pentru analiza exploratorie a factorilor

Verificări privind adecvarea:

După ce datele au fost preluate, trebuie să verificăm dacă analiza factorială este fezabilă. Testul de Sfericitate Bartlett este o verificare a intercorelației dintre variabilele observabile, ceea ce înseamnă compararea matricei de corelație observată cu matricea identității. Dacă analiza factorială este o metodă potrivită de utilizat, matricea de corelație și matricea identității nu vor fi la fel, iar testul va fi semnificativ. Testul de Sfericitate Bartlett trebuie să producă un p-valoare semnificativ mai mic de 0.05.

În continuare, testul KMO (Kaiser-Meyer-Olkin) ar trebui să verifice dacă este adecvat să utilizăm variabilele observabile pentru analiza factorială. Testul implică calculul proporției de varianță între variabilele observabile. Valorile KMO variază între 0 și 1, iar o proporție sub 0.6 ar sugera că setul de date nu este potrivit pentru analiza factorială.

Numărul de Factori:

Trebuie efectuată o analiză factorială preliminară fără nicio rotire. Acest pas ajută la luarea deciziei cu privire la numărul de factori utilizați într-o soluție. În acest pas, obținem valorile proprii ale soluției noastre inițiale și le plasăm pe un grafic de varianță. Putem observa numărul de factori generate versus valorile proprii. Valorile proprii mai mari sau egale cu 1 ar trebui luate în considerare atunci când se alege numărul de factori. Un factor cu o valoare proprie de 1 explică cel puțin varianța unei singure caracteristici. De asemenea, se poate utiliza metoda subiectivă a cotului (elbow method).

Bazându-ne pe graficul de varianță, vom alege o soluție cu patru factori. Acest lucru poate fi specificat atunci când utilizăm pachetul factor\_analyzer. Putem începe rotirea acestor factori. Deoarece analiza factorială (FA) este o metodă iterativă, este indicat să respectăm un proces general:

1. Începem cu rotirea VariMax.
2. Metoda poate fi setată la minres, ml sau principal. Putem începe cu minres, în timp ce efectuăm rotirea VariMax.
3. Schimbăm metoda la maximum likelihood, dar continuăm să folosim rotirea VariMax.
4. Există două opțiuni logice pentru a decide dacă să folosim coeficientul de corelație multiplu la pătrat ca estimare inițială pentru analiza factorială. Începem întotdeauna cu smc (de exemplu, coeficientul de corelație multiplu la pătrat) și încercăm apoi coeficientul de corelație maximă absolută ca a doua opțiune.
5. Comparăm soluțiile și păstrăm cea care funcționează cel mai bine.
6. Evaluăm încărcările factorilor și luăm în considerare o soluție cu un factor mai mare și una cu un factor mai mic decât k-ul ales.

Dacă împărțim datele, acum putem încerca soluția pe datele de test.

# Motivația de a opta pentru o anumită metodă de analiză a datelor

În ceea ce privește ACP, am decis să alegem această metodă deoarece duce la reducerea dimensiunii setului de date, transformând variabilele inițiale într-un set de componente principale, menținând în același timp cea mai mare parte a variației din date. Acest aspect poate face mai ușoare înțelegerea și interpretarea rezultatelor. Este cunoscut faptul că această metodă identifică corelațiile între diferiți factori, ceea poate dezvălui legături subiacente între factorii de risc și numărul total de decese. Componentele principale produc o reprezentare vizuală facilă a datelor, permițând o bună înțelegere a distribuției relațiilor dintre țări și factorii de risc principali. De asemenea, prin această metodă este eliminată redundanța datelor inițiale, evidențiindu-se doar informațiile semnificative. În ultimul rând, această metodă poate gestiona existența multicoliniarității între factorii de risc.

Analiza factorială (AF) este o metodă utilizată pentru a dezvălui relațiile dintre factorii latenți presupuși și variabilele observabile. Un factor latent este un concept care nu poate fi măsurat direct, dar se presupune că are o relație cu mai multe caracteristici măsurabile în date, numite variabile observabile. Variabilele observabile sunt direct măsurabile. Există două forme principale de AF: AF exploratorie și AF confirmatorie. AEF este concepută pentru a descoperi relațiile dintre variabilele observabile și factori fără a presupune că anumite variabile observabile sunt legate de anumiți factori. AEF este folosită în special atunci când avem un set mare, cum avem in cazul nostru, 29 de variabile observabile. Am ales analiza exploratorie a factorilor, deoarece aceasta permite identificarea principalilor factori care contribuie la deces în rândul populației. Prin evaluarea acestor factori, se poate determina care aspecte au o influență semnificativă asupra sănătății și mortalității.

# Prezentarea rezultatelor

În primul rând, am citit setul de date inițial situat la calea **dataIN/number-of-deaths-by-risk-factor.csv** într-un tabel în memorie sub forma unui Pandas Dataframe în care indexul este prima coloană a tabelului. Această coloană reprezintă țara sau zona geografică în care s-au înregistrat nivelul de morți pe fiecare factor de risc. Am curățat setul de date parcurgând fiecare înregistrare și editând valorile lipsă prin aproximarea acestora cu media înregistrării. În final, am standardizat datele folosindu-ne de **StandardScaler** din pachetul **sklearn.preprocessing**.

## ACP

Rezultatele analizei în componente principale vor fi găsite la căile **dataOUT/ACP/Files** (pentru fișiere)și **dataOUT/ACP/Plots** (pentru grafice). Folosindu-ne de modulul **PCA** pus la dispoziție de biblioteca **sklearn.decomposition**, am instanțiat un astfel de obiect pentru un număr variabil de componente principale, folosindu-ne de valorile standardizate determinate anterior. Am parcurs acest demers deoarece ne-a interesat care este numărul minim de componente principale care determină o varianță aproape maximă (0,999...) a setului de date. Rezultatul poate fi găsit la calea **dataOUT/ACP/Plots/varianța\_explicată\_de\_nr\_componentelor.png**1. În continuare, am vrut să vedem care este varianța explicată de fiecare componentă în parte. Pentru a realiza acest lucru, am construit un grafic folosind ca date de intrare valorile proprii ale instanței de **PCA**, anume **explained\_variance\_**. Rezultatul poate fi găsit la calea **dataOUT/ACP/Plots/varianța\_explicată\_de\_componentele\_principale.png**2. Vom salva factorii de sarcină specifici analizei noastre la calea **dataOUT/ACP/Files/factori\_de\_sarcină.csv**3. De asemenea, vom face și o corelogramă găsită la calea **dataOUT/ACP/Plots/corelograma\_factorilor\_de\_sarcină.png**3. Valorile observațiilor inițiale transformate în spațiul componentelor principale (scorurile) le-am salvat la calea **dataOUT/ACP/Files/scoruri\_în\_noul\_spațiu.csv**4. În fine, vom calcula calitatea reprezentării observațiilor5, contribuția observațiilor la varianța axelor componentelor principale6 și comunalitățile7 variabilelor noi și le vom salva corespunzător.

## AEF

Pentru analiza exploratorie a factorilor (AEF), rezultatele analizei pot fi găsite la căile **dataOUT/AEF/Files** (pentru fișiere) și **dataOUT/AEF/Plots** (pentru grafice). Am efectuat testul Bartlett pentru a evalua sfericitatea datelor în contextul analizei exploratorie a factorilor (AEF). Am calculat și salvat indicele KMO (Kaiser-Meyer-Olkin) pentru a măsura adecvarea datelor pentru analiza factorilor. Rezultatele acestui indice au fost stocate în fișierul **dataOUT/AEF/Files/indici\_kmo.csv8**, iar o corelogramă corespunzătoare acestor indici a fost creată și salvată la calea **dataOUT/AEF/Plots/corelograma\_indicilor\_kmo.png9**. Am calculat și salvat factor loadings, indicând asocierea dintre variabile și factorii identificați, în fișierul **dataOUT/AEF/Files/factor\_loadings.csv10**. De asemenea, am creat o corelogramă a factorilor de corelație, oferind o perspectivă vizuală asupra interdependențelor dintre factori, iar aceasta a fost salvată la calea **dataOUT/AEF/Plots/corelograma\_factorilor\_de\_corelație.png11.**  În continuare, am examinat valori proprii pentru a evalua varianța explicată de factorii comuni, iar rezultatele au fost ilustrate într-un grafic disponibil la **dataOUT/AEF/Plots/varianța\_explicată\_de\_factorii\_comuni\_fa.png12**. Scorurile factoriale obținute din AEF au fost salvate la **dataOUT/AEF/Files/scoruri\_factoriale.csv13**, iar corelograma scorurilor este disponibilă la calea **dataOUT/AEF/Plots/corelograma\_scorurilor.png14**. Pentru a examina relațiile între factori, am realizat un cerc al corelațiilor, ilustrat în fișierul **dataOUT/AEF/Plots/cercul\_corelațiilor.png15**, unde poziționarea variabilelor pe cerc indică corelațiile lor cu factorii obținuți.

# Interpretare

## ACP

1 – În graficul obținut putem observa că numărul minim de componente principale pentru care obținem o varianță aproape perfectă este de 10. Astfel, observăm o creștere nesemnificativă după acest prag, ceea ce înseamnă că setul de date poate fi exprimat aproape perfect prin doar 10 din cele 29 de caracteristici.

2 – Am ales să limităm domeniul pe axa Oy pentru a putea sesiza mai bine diferențele subtile dintre componentele de la 3 în sus. Prima componentă explică cea mai mare parte din varianță, a doua fiind la distanță mare, însă menținând tot o valoare semnificativă, semn că este esențială modelului. De la a treia componentă se remarcă scăderea în varianța explicată a componentelor. Conform criteriului lui Kaiser, deoarece variabilele observate sunt standardizate și au varianță 1, apare firesc în ceastă situație ca noile variabile, componentele principale, să fie considerate importante, semnificative, în măsura în care cumulează mai multă varianță decât o variabilă observată. Așadar, se recomandă reținerea acelor componente principale care au varianță **mai mare decât 1**. În cazul nostru, C1 și C2, deoarece ele se află peste orizontala , colorată cu roșu.

După criteriul lui Cattell în varianta grafică, se determină primul unghi dintre două pante consecutive mai mare decât 180°. Se rețin doar componentele până la acel punct inclusiv. Conform graficului nostru, este vorba despre componentele C1, C2, C3 și C4.

3 – Valorile din matrice arată nivelul de contribuție al variabilelor inițiale la noile variabile determinate (componentele principale). O coloană din corelogramă ne arată în ce măsură și în ce direcție contribuie fiecare risc la noua variabilă. Putem observa, spre exemplu, că la nivelul componentei 3, factorii de risc care contribuie în mod direct și puternic sunt **poluarea aerului în exterior** și **fumatul**, ceea ce ne duce cu gândul că această componentă este strâns legată de factorii de risc ce duc la morțile cauzate de bolile respiratorii.

4 – Fiecare valoare din această matrice reprezintă proiecția observației inițiale pe axele formate de noile variabile. Matematic, se obține înmulțind matricea de date standardizată cu matricea vectorilor proprii.

5 – Calitatea reprezentării observațiilor ne arată cât de bine este reprezentată fiecare observație în noul spațiu determinat de componentele principale. Cu cât valorile sunt mai apropiate de 1, cu atât reprezentarea observației este mai bună.

6 – Contribuția observațiilor la varianța axelor ne arată cât de mult influențează fiecare observație fiece componentă principală.

7 – Comunalitățile oferă informații despre cât de mult din variația fiecărei variabile originale este reprezentată de componentele principale.

## AEF

8 – KMO (Kaiser-Meyer-Olkin) este un indice utilizat pentru a evalua adecvarea datelor pentru analiza factorială. Interpretarea valorilor indicelui KMO (Kaiser-Meyer-Olkin) în contextul factorabilității datelor:

* 0,90 - 1,00: Factorabilitate foarte bună
* 0,80 - 0,89: Factorabilitate bună
* 0,70 - 0,79: Medie
* 0,60 - 0,69: Mediocră
* 0,50 - 0,59: Slabă
* 0,00 - 0,49: Lipsesc factori

Cu majoritatea variabilelor având un indice de peste 0,9, iar cel mai mic indice înregistrat fiind de 0,83 pentru "unsafe water source", concluzionăm că datele prezintă o factorabilitate foarte bună.

9 – Corelograma indicilor KMO, prezentată în figura 9, evidențiază cu culori variabilele analizate în funcție de coeficienții lor de corelație. Toți indicii au culoarea roșie, reflectând corelații directe și puternice între variabile.

10 – Indică cât de mult fiecare variabilă este influențată de cei doi factori comuni identificați în cadrul analizei factorilor exploratori (AEF). Cu cât încărcarea factorială pentru o anumită variabilă și un anumit factor este mai mare (aproape de 1 sau -1), cu atât acea variabilă este mai puternic asociată și influențată de acei factori specifici.

11 – Această vizualizare grafică oferă o perspectivă asupra gradului de asociere între factori, facilitând înțelegerea modului în care aceștia interacționează și contribuie la variabilitatea datelor. Se evidențiază două grupări distincte în analiza exploratorie a factorilor, unde 15 variabile sunt asociate predominant cu factorul 1, în timp ce 12 variabile sunt corelate mai puternic cu factorul 2.

12 – Valorile proprii (eigenvalues) situate deasupra liniei în graficul de varianță indică importanța factorilor corespunzători în explicarea variației datelor. Selectăm numărul de factori asociat punctului unde linia începe să scadă mai lent, evidențiind o contribuție semnificativă a acestor factori la variația totală. Acești factori sunt reținuți în analiza exploratorie a factorilor pentru a captura esența variației semnificative a datelor. În corelograma noastră, se pot distinge clar 2 factori semnificativi.

13 – Furnizează informații despre contribuția variabilelor la factorii identificați în analiza exploratorie a factorilor (AEF). Prin analiza acestor scoruri, se poate evalua modul în care variabilele inițiale contribuie la definirea și influențarea factorilor identificați în procesul de analiză.

14 – Această corelogramă prezintă grafic scorurile factoriale pentru fiecare observație în raport cu factorii relevanți, oferind o imagine vizuală asupra distribuției și relațiilor acestora în noul spațiu al factorilor. Valorile pozitive și negative indică direcția relației dintre observații și factori. Valorile pozitive sugerează o asociere pozitivă între observații și factorii respectivi, ceea ce înseamnă că, pe măsură ce o variabilă observabilă crește, scorul factorului asociat crește, și viceversa. Pe de altă parte, valorile negative indică o asociere negativă, semnificând că o creștere a variabilei observate este asociată cu o scădere a scorului factorului, și invers.

15 – Reprezintă grafic cercul corelațiilor în cadrul analizei exploratorie a factorilor (AEF). Acest grafic oferă o vizualizare a relațiilor dintre variabile și factori în spațiul bidimensional, evidențiind corelațiile variabilelor cu factorii identificați. axa Ox reprezintă factorul 2 și axa Oy reprezintă factorul 1 în cadrul cercului corelațiilor în analiza exploratorie a factorilor (AEF), poziționarea variabilelor în primul cadran indică o asociere pozitivă între variabilele observate și acești doi factori. Mai specific, creșterea valorilor variabilelor este corelată pozitiv cu scorurile factorului 2 de-a lungul axei Ox și cu scorurile factorului 1 de-a lungul axei Oy.

# Discuții și concluzii

Aceste două analize pe care le-am întreprins în lucrarea noastră se pretează cel mai bine pentru setul de date ales, deoarece reducerea dimensionalității reprezintă un aspect cheie care trebuia întreprins pentru a extrage cât mai mult informație relevantă din setul brut de date.

După realizarea analizei în componente principale, am ajuns la concluzia că setul de date inițial putea fi minimizat la doar 10 componente principale pentru a explica aproape perfect varianța datelor. Este greu de precizat ce ar putea include fiecare dintre aceste 10 componente principale din cei 29 de factori care duc la deces, dar din corelograma factorilor de sarcină putem intui, spre exemplu, că cea de-a treia componentă este strâns legată de poluarea mediului.

În ceea ce privește analiza exploratorie a factorilor, ne-a fost confirmată încă o dată buna alegere a metodelor de analiză atât prin valorile foarte apropiate de 1 ale indicilor KMO calculați, cât și prin corelograma scorurilor care ne arată cât de mult contribuie factorii determinați de noi la explicarea varianței din setul de date original.