**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](#_Toc155292401)

[Referință de timp privind datele 2](#_Toc155292402)

[Descrierea variabilelor 2](#_Toc155292403)

[Descrierea observațiilor 3](#_Toc155292404)

[Abordarea folosind analiza datelor 3](#_Toc155292405)

[Motivația de a opta pentru o anumită metodă de analiză a datelor 4](#_Toc155292406)

[Prezentarea rezultatelor 5](#_Toc155292407)

[ACP 5](#_Toc155292408)

[Interpretare 5](#_Toc155292409)

[Discuții și concluzii 6](#_Toc155292410)

# 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 eficienți 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.

În cadrul [PLACEHOLDER pentru voi, Gabi și Ardeu] …

# 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.

În ceea ce privește [PLACEHOLDER pentru voi, Gabi și Ardeu] …

# 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.

# Interpretare

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?

# Discuții și concluzii

TBA