**ACADEMIA DE STUDII ECONOMICE**

Facultatea: Cibernetică, Statistică şi Informatică Economică

Specializarea: ***Informatică economică***



**Analiza cauzelor deceselor la Uniunii Europa**

**Cadru didactic coordonator:** **Studenți grupa 1087:**

Prof.Univ.Dr. VINȚE C CLAUDIU David Cristian-Paul

Dejan Marius-Ionuț

Dinulescu Oana-Adelina

Dorobanțu George-Alexandru

**Cuprins**

[Sursa de date 3](#_Toc167823125)

[Referință de timp privind datele 3](#_Toc167823126)

[Descrierea variabilelor 3](#_Toc167823127)

[Descrierea observațiilor 4](#_Toc167823128)

[Abordarea privind analiza datelor 4](#_Toc167823129)

[Motivația de a opta pentru o anumită metodă de analiză a datelor 4](#_Toc167823130)

[Prezentarea rezultatelor (nu includeti in document grafice care rezulta din cod) 4](#_Toc167823131)

[Interpretare 4](#_Toc167823132)

[Discuții și concluzii 4](#_Toc167823133)

# Sursa de date

Pentru acest proiect am ales să utilizăm datele furnizate de WHO’s Global Health Estimates (GHE), acestea fiind disponibile la următoarea adresă https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death.

# Referință de timp privind datele

Datele utilizate în analiza przentate au fost extrase dintr-un set de date complex, reprezentând valorile observate pe teritoriul Uniunii Europene în anul 2019.

# Descrierea variabilelor

* **Infectious and parasitic diseases (Boli infecțioase și parazitare)**: Rata deceselor cauzate de boli infecțioase și parazitare.
* **Respiratory Infectious (Boli infecțioase respiratorii)**: Rata deceselor cauzate de infecții respiratorii.
* **Maternal conditions (Condiții materne)**: Rata deceselor cauzate de complicații legate de sarcină și naștere.
* **Neonatal conditions (Condiții neonatale)**: Rata deceselor cauzate de condiții care afectează nou-născuții în primele zile sau săptămâni de viață.
* **Nutritional deficiencies (Deficiențe nutriționale)**: Rata deceselor cauzate de deficiențe nutriționale, cum ar fi malnutriția.
* **Malignant neoplasms (Neoplasme maligne)**: Rata deceselor cauzate de cancere.
* **Other neoplasms (Alte neoplasme)**: Rata deceselor cauzate de alte tipuri de tumori, care nu sunt neapărat maligne.
* **Diabetes mellitus (Diabet zaharat)**: Rata deceselor cauzate de diabet.
* **Endocrine, blood, immune disorders (Tulburări endocrine, sanguine și imune)**: Rata deceselor cauzate de tulburări ale sistemului endocrin, sanguin și imunitar.
* **Cardiovascular diseases (Boli cardiovasculare)**: Rata deceselor cauzate de afecțiuni ale inimii și ale vaselor de sânge.
* **Respiratory diseases (Boli respiratorii)**: Rata deceselor cauzate de afecțiuni ale sistemului respirator, cum ar fi bronșita cronică și emfizemul.
* **Digestive diseases (Boli digestive)**: Rata deceselor cauzate de afecțiuni ale sistemului digestiv, cum ar fi ulcerul gastric și hepatita.
* **Genitourinary diseases (Boli genitourinare)**: Rata deceselor cauzate de afecțiuni ale sistemului genitourinar, cum ar fi insuficiența renală.
* **Skin diseases (Boli ale pielii)**: Rata deceselor cauzate de afecțiuni ale pielii.
* **Musculoskeletal diseases (Boli musculo-scheletice)**: Rata deceselor cauzate de afecțiuni ale sistemului musculo-scheletic, cum ar fi artrita.
* **Congenital anomalies (Anomalii congenitale)**: Rata deceselor cauzate de defecte congenitale, prezente la naștere.
* **Sudden infant death syndrome (Sindromul morții subite a sugarului)**: Rata deceselor cauzate de sindromul morții subite a sugarului, o cauză inexplicabilă de deces la sugari aparent sănătoși.
* **Unintentional injuries (Leziuni neintenționate)**: Rata deceselor cauzate de accidente și alte incidente neintenționate, cum ar fi accidentele de circulație.
* **Intentional injuries (Leziuni intenționate)**: Rata deceselor cauzate de acte de violență, cum ar fi sinuciderea și omuciderea.

# Descrierea observațiilor

Fiecare rând din tabel corespunde unei țări și conține valorile numerice asociate pentru diferite categorii de cauze ale deceselor.

# **Abordarea privind analiza datelor**

**Analiza Componentelor Principale (ACP):**

Analiza de tip ACP este o tehnică statistică utilizată pentru a transforma un set de date cu multe variabile într-un set mai mic de componente principale. Aceste componente sunt combinații liniare ale variabilelor originale și sunt ordonate în funcție de varianța explicată.

**Standardizarea datelor:**

Pentru a pregăti datele pentru ACP, am standardizat matricea X utilizând funcția de standardizare definită într-un modul extern. După standardizare, am salvat matricea standardizată într-un fișier CSV pentru a putea fi utilizată ulterior.

**Instanțierea modelului:**

Am instanțiat modelul ACP folosind matricea standardizată și am extras valorile proprii, pe care le-am utilizat pentru a realiza un grafic al variantei explicate de componentele principale. Acest grafic ne-a ajutat să vizualizăm importanța fiecărei componente principale în explicarea variabilității datelor.

**Extragerea componentelor principale:**

În continuare, am extras componentele principale și le-am salvat într-un fișier CSV, facilitând astfel analiza ulterioară. Am calculat factorii de corelație și i-am salvat, de asemenea, într-un fișier CSV. Pentru a vizualiza relațiile dintre variabile și componentele principale, am realizat o corelogramă a factorilor de corelație.

**Calcularea scorurilor:**

Am calculat și scorurile pentru fiecare observație în funcție de componentele principale și le-am salvat într-un fișier CSV. Corelograma scorurilor ne-a permis să examinăm distribuția scorurilor și să identificăm eventualele modele sau grupări în date.

**Calcularea calității reprezentării observațiilor:**

Pentru a evalua calitatea reprezentării observațiilor pe componentele principale, am calculat și salvat valorile corespunzătoare în fișiere CSV. Am utilizat corelograme pentru a vizualiza aceste calități de reprezentare.

**Calcularea contribuțiilor la varianța axelor:**

Am calculat contribuția fiecărei observații la varianta axelor (componentelor principale) și am salvat aceste informații pentru a putea analiza impactul fiecărei observații asupra componentelor principale.

**Calculul comunalităților:**

În final, am calculat comunalitățile, care reprezintă partea din variabilitatea originală a variabilelor explicată de componentele principale, și le-am salvat într-un fișier CSV. Am realizat o corelogramă pentru a vizualiza comunalitățile și pentru a înțelege mai bine cât de bine sunt reprezentate variabilele inițiale de către componentele principale.

**Analiza Exploratorie a Factorilor (AEF):**

Această analiză este o metodă utilizată pentru a explora relațiile dintre variabilele dintr-un set de date. AEF ne ajută să înțelegem structurile și modelele subiacente ale datelor. Pentru a realiza analiza de tip AEF asupra setului de date ales, am urmarit urmatorii pasi:

**Importarea și preprocesarea datelor:**

Începem prin a citi datele dintr-un fișier CSV. Valorile lipsă din date sunt completate cu media fiecărei coloane, iar datele sunt standardizate pentru a avea o medie de 0 și o abatere standard de 1.

**Matricea de corelație și selectarea variabilelor:**

Calculăm matricea de corelație pentru datele standardizate și identificăm variabilele care au o corelație mai mare de 0.95. Aceste variabile puternic corelate sunt eliminate pentru a evita redundanța în analiza ulterioară.

**Testul Bartlett și testul KMO:**

Efectuăm testul Bartlett pentru a verifica dacă matricea de corelație este o matrice de identitate, indicând faptul că factorii sunt poriviți pentru analiză. Efectuăm testul KMO (Kaiser-Meyer-Olkin) pentru a măsura adecvarea eșantionului pentru analiza factorială.

**Realizarea corelogramei:**

Realizăm o corelogramă pentru indici KMO folosind Seaborn și Matplotlib, pentru a realiza corelațiile dintre variabile.

**Determinarea factorilor semnificativi (EFA):**

Efectuăm o analiză factorială exploratorie folosind FactorAnalyzer pentru a determina numărul de factori semnificativi pe baza criteriului Kaiser (eigenvalues mai mari de 1). Refacem calculele pentru încărcăturile factoriale cu rotație Varimax și salvăm rezultatele într-un fișier CSV.

**Calcularea scorurile factoriale:**

Calculăm scorurile factoriale pentru fiecare observație și le salvăm într-un fișier CSV. Creăm și salvăm o corelogramă pentru scorurile factoriale.

**Crearea cercului corelațiilor:**

Creăm un cerc al corelațiilor care vizualizează relațiile dintre factorii principali din analiza factorială.

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

Unul dintre principalele motive pentru utilizarea ACP este redimensionarea datelor. Prin eliminarea variabilelor redundante sau corelate, ACP reduce dimensiunea setului de date, făcând analiza mai eficientă și ușurând vizualizarea datelor. O altă utilitate a ACP este identificarea structurilor latente. Proiecția datelor pe noile componente principale dezvăluie modele sau grupuri care nu sunt evidente în datele brute. Astfel, putem descoperi relații ascunse între variabile. Vizualizarea datelor este, de asemenea, îmbunătățită prin ACP. Reprezentarea datelor într-un spațiu multidimensional mai mic facilitează interpretarea și explorarea. Graficele bidimensionale sau tridimensionale pot evidenția relațiile dintre variabile.

Am ales analiza de exploratorie a factorilor deoarece este o tehnică utilizată pentru a explora relațiile dintre variabilele dintr-un set de date. AEF este o metodă puternică pentru a înțelege structurile și modelele subiacente ale datelor. Unul dintre principalele motive pentru utilizarea AEF este identificarea factorilor relevanți. Prin analiza relațiilor dintre variabile, putem descoperi factorii care influențează datele noastre. O altă utilitate a AEF este vizualizarea datelor. Prin reprezentarea grafică a relațiilor dintre variabile, putem observa tendințe, corelații și grupuri. AEF poate ajuta, de asemenea, la detectarea outlier-ilor. Prin identificarea valorilor extreme, putem înțelege mai bine distribuția datelor.

# **Prezentarea și interpretare**a **rezultatelor**

**Analiza Componentelor Principale (ACP):**

În graficul obținut putem observa că numărul minim de componente principale necesare pentru a explica cea mai mare parte din varianța setului de date. Am ales selectarea componentelor principale cu valori proprii peste 1, respectiv primele 5 componente. Acestea explică în proporție de peste 78% varianța observațiilor. Se observă o creștere nesemnificativă după acest prag, ceea ce înseamnă că setul de date poate fi exprimat prin doar 5 din cele 20 de caracteristici fără a pierde o cantitate mare de informație.

Corelograma factorilor de corelație este o diagramă care arată corelațiile între variabilele inițiale și componentele principale rezultate din analiză. De exemplu, Infecțiile Respiratorii au o sarcină pozitivă puternică pe componenta principală C3 (0.6) și o sarcină negativă moderată pe componenta principală C1 (-0.5). Condițiile Materne arată o sarcină pozitivă puternică pe componenta principală C1 (0.7) și o sarcină negativă moderată pe componenta principală C3 (-0.4).

În corelograma scorurilor, fiecare rând reprezintă scorurile unei țări pe fiecare dintre cele 20 de componente principale. De exemplu, România are scoruri negative puternice pe componentele C1, C3, și C8, iar scoruri pozitive pe componentele C5 și C13. Putem compara ușor diferite țări în funcție de scorurile lor pe componentele principale. Țările care au modele de scoruri similare sunt susceptibile să aibă caracteristici similare în variabilele inițiale. De exemplu, Serbia și Rusia au modele similare de scoruri (scoruri negative pe C1 și C3), sugerând o posibilă similaritate în variabilele analizate.

Corelogramă a comunalităților este utilizată în Analiza Componentelor Principale (ACP) pentru a indica proporția varianței fiecărei variabile originale explicată de componentele principale. Variabile precum "Malignant neoplasms" și "Other neoplasms" au comunalități ridicate (aproape de 1) pe majoritatea componentelor, indicând că aceste variabile sunt bine reprezentate de componentele principale. "Infant death syndrome" și "Respiratory diseases" au valori mai mici la început, dar valorile cresc semnificativ pe ultimele componente, sugerând că sunt necesare mai multe componente pentru a explica varianța acestor variabile. Variabilele "Maternal conditions" și "Neonatal conditions" arată comunalități ridicate pe primele componente, ceea ce înseamnă că primele componente principale sunt foarte eficiente în explicarea varianței acestor variabile.

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.

**Analiza Exploratorie a Factorilor (AEF):**

Testul de Sfericitate Bartlett este utilizat pentru a verifica dacă matricea de corelație observată a variabilelor din setul de date diferă semnificativ de matricea identității. Practic, acest test verifică dacă există o anumită redundanță între variabile, pe care o putem rezuma printr-un număr mic de factori.

Bartlett's p-value calculat in cazul nostru est 6.688316706169923e-39, mult sub pragul de seminficatie, asadar putem concluziona că setul nostru de date este potrivit pentru o tehnică de reducere a datelor.

Corelograma indicilor KMO ajută la evaluarea adecvării datelor pentru analiza factorială, indicând care variabile sunt potrivite și care ar putea necesita transformări suplimentare. Valorile KMO sunt prezentate lângă fiecare variabilă și variază între 0 și 1. Un indice KMO mai aproape de 1 indică faptul că variabilele sunt potrivite pentru analiza factorială.

0.8 - 1.0: Factorabilitate excelentă.

0.7 - 0.8: Factorabilitate bună.

0.6 - 0.7: Factorabilitate medie.

0.5 - 0.6: Factorabilitate slabă.

<0.5: Inadecvată pentru analiza factorială.

Variabilele precum "Digestive diseases" (0.84) și "Cardiovascular diseases" (0.77) au indici KMO relativ ridicați, indicând că sunt foarte potrivite pentru analiza factorială. "Malignant neoplasms" are un indice KMO de 0.48, ceea ce sugerează o adecvare scăzută pentru analiza factorială. Majoritatea variabilelor au indici KMO peste 0.6, ceea ce indică o adecvare moderată spre bună pentru analiza factorială.

Pentru a decide numărul de factori am utilizat criteriul Kaiser, care se bazează pe valorile proprii (eigenvalues) ale matricei de corelație. Fiecare factor are un eigenvalue asociat. Dacă eigenvalue-ul este mai mare decât 1, atunci factorul este considerat semnificativ. Se observă că primele 5 valori proprii depășesc pragul de 1.

Corelogramă a factorilor de corelație din FA (Factor Analysis)" și este utilizată pentru a vizualiza relația dintre variabilele originale și factorii extrași în analiza factorială. Valorile din celule indică coeficientul de corelație între variabilele inițiale și factorii corespunzători: o valoare apropiată de 1 sau -1 indică o corelație puternică, pozitivă sau negativă, între variabila inițială și factorul respectiv. Other neoplasms" are o corelație puternică cu Factorul 1 (0.62) și o corelație moderată cu Factorul 4 (0.83), ceea ce sugerează că acești factori sunt importanți pentru această variabilă. "Cardiovascular diseases" are o corelație negativă puternică cu Factorul 1 (-0.78), ceea ce sugerează o relație inversă puternică cu acest factor.

Graficul varianței explicată de factorii comuni ne ajută să vizualizăm varianta explicată de fiecare componentă principală din analiza factorială. Prima componentă principală (C1) explică cea mai mare parte a varianței, cu o valoare proprie de aproximativ 6. A doua componentă principală (C2) explică o parte semnificativă a varianței, dar mai puțin decât prima componentă. După a cincea componentă principală, valorile proprii scad sub 1, indicând că aceste componente sunt mai puțin semnificative. Numărul de componente principale semnificative este determinat de punctul unde curba se aplatizează. În cazul nostru, se observă o scădere semnificativă la a 3-a componentă, acestea rămânând totuși relativ relevante până la cea de a 5-a, unde se observă un nou colaps în importanță.

Cercul corelațiilor în cadrul analizei exploratorie a factorilor (AEF) oferă o vizualizare a relațiilor dintre variabile și factori în spațiul bidimensional, evidențiind corelațiile variabilelor cu factorii identificați. Pe axa Ox este reprezentat factorul 2, iar pe axa Oy factorul 1. Fiecare punct din grafic reprezintă o variabilă inițială și este poziționat în funcție de valorile de corelație cu Factorul 1 și Factorul 2. Variabilele care sunt situate aproape de marginea cercului au o corelație puternică cu factorii respectivi.

# Concluzii

Analiza efectuată în cadrul lucrării noastre este potrivită setului de date selectat, deoarece reducerea dimensionalității este esențială în extragerea informațiilor semnificative din datele brute. În special, analiza în componente principale a fost extrem de utilă în acest sens.

După finalizarea analizei în componente principale, am constatat că putem reduce setul nostru de date inițial la doar 5 componente principale pentru a explica în cea mai mare parte variația datelor. Deși este dificil să identificăm precis conținutul acestor 5 componente principale din cei 20 de factori care contribuie la rata mortalității.

Utilizarea analizei exploratorie a factorilor a fost justificată prin valorile ridicate ale indicelui KMO, care sunt apropiate de 1, cât și de corelograma scorurilor, care arată contribuția factorilor identificați la explicarea variației din setul nostru de date original.

Așadar, prin reducerea numărului de componente ce descriu informația studiată, putem obține o înțeșegere superioară asupra acesteia. Astfel, descoperim corelațiile și legăturile dintre variabilele inițiale dinb setul de date și componentele noi create de noi, pentru a descoperi cauzalitatea deceselor, precum și modalitatea în care cauzele de deces se influențează reciproc.