**Clasificarea tarilor din Europa**

**dupa exporturile de bunuri**

**si servicii ca procent din PIB**

**(Tema 3 - Analiza de clusteri)**

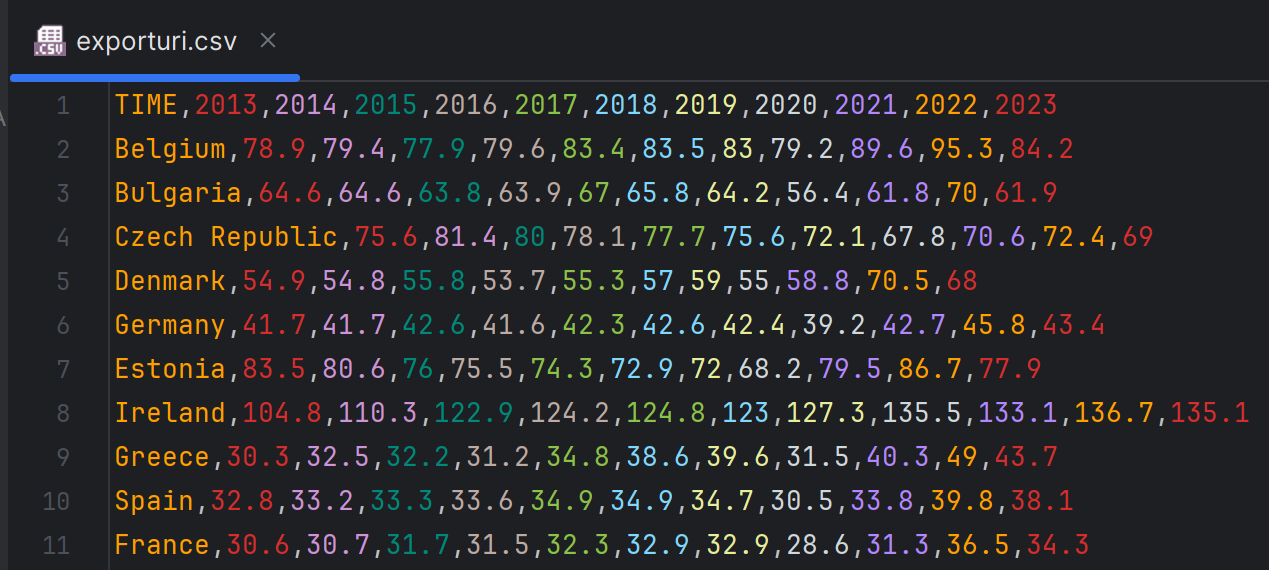
**Sursa datelor**Eurostat (Exports of goods and services in % of GDP) **Fișier de intrare:**

Fișierul **exporturi.csv** din directorul data\_in conține date referitoare la exporturile de bunuri și servicii ale țărilor europene, exprimate ca procent din PIB. Datele sunt organizate pe linii si coloane, fiecare linie reprezintă datele despre exporturile dintr-o țară dintr-un anumit an, în timp ce fiecare coloană reprezintă un an diferit, din 2013 până în 2023.

Structura tabelului este:

* **Coloana "TIME"**: Aceasta conține numele țărilor (de exemplu, Belgia, Bulgaria, etc.).
* **Coloanele de la 2013 până la 2023**: Fiecare coloană reprezintă exporturile pentru un anumit an din intervalul 2013-2023 pentru țara respectivă.

Exemplu cum arata datele:



***Analiza de clusteri:***

**Dendrograma ierarhică**: Am aplicat metoda Ward de clusterizare ierarhică pentru a construi o ierarhie a instanțelor din setul de date. Rezultatul a fost vizualizat sub forma unei dendrograme, care arată cum sunt grupate treptat instanțele (tarile) pe măsură ce distanțele dintre ele scad.

A diagram with blue lines

Description automatically generated

Dendrograma ne ajută să înțelegem cum sunt grupate țările pe măsură ce distanțele dintre ele scad. În cazul nostru, putem observa că țările din Europa Centrală și de Est sunt mai strâns legate între ele, în timp ce țările din Europa de Vest formează grupuri distincte

**Metoda Elbow:** Am folosit metoda Elbow pentru a determina numărul optim de clustere. Aceasta presupune analiza distanțelor dintre clustere pe măsură ce numărul de clustere crește și identificarea punctului în care rata de reducere a distanței scade semnificativ (acesta indică numărul optim de clustere).

În graficul Elbow, punctul de cotitură vizibil la 2 clustere sugerează că adăugarea unui al treilea cluster nu aduce o reducere semnificativă a distanței dintre clustere, indicând astfel numărul optim de clustere.A diagram of a diagram

Description automatically generated with medium confidence

A diagram of a group of people

Description automatically generated with medium confidence

**Graficul Silhouette:** arată cât de bine sunt grupate instanțele în cadrul clusterelor.

Scorul Silhouette mediu pentru partitia optimă este de 0.702, ceea ce indică o bună separare între clustere. Instanțele sunt relativ bine grupate, iar distanța dintre clustere este suficient de mare pentru a evidenția diferențele

A graph of a graph

Description automatically generated with medium confidence

Pentru partitia cu 3 clustere, scorul Silhouette a scăzut la 0.56, ceea ce sugerează o separare mai slabă între clustere comparativ cu partitia optimă.

A chart of different colored shapes

Description automatically generated with medium confidence

**Scatter plot-ul** reprezintă instanțele (țările) plasate pe primele două componente principale (obținute prin PCA), fiecare țară fiind colorată în funcție de clusterul din care face parte. Acesta arată distribuția țărilor în funcție de variabilele selectate.

Z1 si Z2 sunt cele 2 componente PCA.

A graph with blue and orange dots

Description automatically generated

A graph with many colored dots

Description automatically generated

**Histogramele** reprezintă distribuția fiecărei variabile (în acest caz, exporturile de bunuri și servicii ca procent din PIB) în cadrul fiecărui cluster. Clusterul 1 (țările cu exporturi mari) prezintă o frecvență mare a valorilor ridicate, în timp ce în clusterul 2, distribuția este mult mai variată.A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

A group of blue bars

Description automatically generated

A graph of different sizes and colors

Description automatically generated with medium confidence

**Observații:** In directorul data\_out se salvează următoarele fișiere:

* ierarhie.csv: conține rezultatele ierarhiei de clusterizare, incluzând distanțele și frecvențele de combinare a clusterelor la fiecare pas.
* partitii.csv: salvează etichetele clusterelor și scorurile Silhouette pentru partitia optimă și cea cu 3 clustere.