**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI, FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**



**Proiect Dezvoltare Software pentru Analiza Datelor**

Profesor coordonator: Vințe Claudiu

Studenți:

Simion Ștefan-Alexandru

Stanca Munteanu Patrick

Stancu Diana-Elena

Szabo Alexandru-Ștefan

Grupă:

1096 E

# Sursa de date

În acest proiect datele de intrare sunt extrase de pe Eurostat și se referă la nivelul de cunoaștere ale skill-urilor digitale pe care populația fiecărei țări europene le are, în funcție de regimul de muncă și de industria în care activează.

# Referință de timp privind datele

Setul de date conține date din anul 2022. Am ales un set de date dintr-un an cât mai apropiat de anul în care suntem.

# Descrierea variabilelor

Am ales acești indicatori pentru că am considerat că sunt semnificativi în descrierea evoluției țărilor europene în ceea ce privește skill-urile digitale.

Coloana All Individuals 🡪 se referă la procentul populației cu un nivel peste medie al skillu-rilor digitale, la nivel de țară.

Coloanele de la Students până la Full Timer 🡪 se referă la regimul de muncă pe care îl desfășoară indivizii din acea țară și este exprimat tot în procente din populația totală.

Coloanele rămase 🡪 fac referire la industria în care activează persoanele din țările europene.

# Descrierea observațiilor

În ceea ce privește observațiile, am ales 37 de țări europene, despre care am extras valori pentru variabilele de mai sus. Acestea sunt întâlnite în cadrul csv-ului sub forma denumirii complete a acestora.

# Abordarea privind analiza datelor

Din punct de vedere al analizei datelor, variabilele alese și valorile extrase pentru observații prezente și în csv-ul cu date de intrare vor fi preluate în structuri de date specifice limbajului Python, mai precis DataFrame din librăria pandas. Utilizând funcții din alte librării folosite frecvent în analiza datelor, vom aplica algoritmi reprezentativi și vom genera grafice sugestive.

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

Am ales Analiza Componentelor Principale, deoarece prin această metodă setul de date își reduce dimensiunea și permite o mai bună vizualizare a datelor, toate acestea realizându-se fără a pierde informații și surprinzând cât mai bine datele semnificative.

Am ales Analiza de Cluster, deoarece ne dorim să grupăm observațiile în variante multiple și să descoperim partiția optimă.

# Prezentarea rezultatelor și interpretarea acestora

Analiza Componentelor Principale:

Prima imagine reprezintă un biplot de la ACP, afișând țările ca puncte în spațiul ale primelor două componente principale (C1 și C2). ACP este o tehnică de reducere a dimensiunii folosită pentru a reduce numărul de variabile într-un set de date, păstrând totuși cât mai multă variație posibilă. Țările sunt poziționate în funcție de scorurile lor pe aceste două componente, care reprezintă combinații ale variabilelor originale.

A doua imagine este o corelogramă a scorurilor ACP. Fiecare celulă din matrice reprezintă corelația dintre componentele principale pentru fiecare țară. Albastru indică o corelație negativă, iar roșu indică o corelație pozitivă. Acest lucru ajută la înțelegerea modului în care fiecare pereche de componente se relatează.

A treia imagine este similară cu prima, dar reprezintă diferite sectoare de muncă în loc de țări. Poziționarea fiecărui sector pe grafic indică modul în care acele sectoare se corelează cu cele două componente principale.

Cea de-a patra imagine este un scree plot care arată varianța explicată de fiecare componentă principală. Este folosit pentru a determina numărul de componente de păstrat într-un ACP. "Cotul" plotului este de obicei folosit ca punct de tăiere pentru selecția componentelor, deoarece din acest punct înainte, componente adiționale contribuie mai puțin la explicarea varianței.

Analiza de Cluster:

În prima imagine există **șase clustere** diferite (C0 până la C5), fiecare reprezentând un grup de țări cu niveluri similare de competențe digitale, raportat la variabilele selectate de noi din setul de date.

A doua imagine este un dendrogram, care este o diagramă în formă de arbore utilizată frecvent pentru a ilustra aranjamentul clusterelor produse de gruparea ierarhică. Acest tip de diagramă este util pentru a vizualiza și interpreta relațiile dintre sectoarele de muncă extrase, permițându-ne să înțelegem cum sunt grupate și să observăm distanța relativă între clustere.

Următoarea imagine reprezintă un dendrogram care ilustrează clusterizarea variabilelor folosind metoda mediei metrice de corelație. Această metodă sugerează faptul că distanța dintre clustere este calculată pe baza mediei corelațiilor dintre variabilele din fiecare cluster. Prin intermediul ei, putem afla dacă sectoarele de muncă extrase au tendințe similare în ceea ce privește competențele digitale ale forței de muncă.

Ultima imagine arată un grafic de dispersie pe care l-am numit “Partiția optimă”. Cu ajutorul acestuia putem identifica grupuri de țări cu caracteristici similare în ceea ce privește competențele digitale ale populației, bazându-ne pe anumite variabile sau indicatori analizați în studiu.