**Academia de Studii Economice din București**

**Facultatea de Cibernetică, Statistică și Informatică Economică**



**PROIECT**

**Știința datelor în afaceri**

**AACPI**

**Analiza gradului de educație la nivel global**

**Studenți**: Cucu Laura-Lavinia **Profesor:**

Sapasu Anamaria Bizovi Mihai

Stefan Diana Maria

Terteci Leona Iohana

**Prezentare temei – problema analizata: gradul de educatie la nivel global**

Am ales sa analizam sistemul educational la nivel global. In acest scop am ales date din sfera educatiei cu privire la gradul de alfabetizare al persoanelor din tarile respective, la procentele de elevi care incep/termina ciclurile de invatamant, de asemenea am urmarit sa evidentiem si daca fondurile alocate de Guvern din banii publici pentru invatamant constituie un factor reprezentativ ce ar putea detemina diverse comportamente educationale specifice in anumite tari(de exemplu, daca intr-o tara in care educatiei i se aloca un buget destul de mic, nivelul fenomenului de abandon scolar este mai mare decat intr-o tara ce primeste fonduri mari pentru investitia in invatamant). De asemenea, am mai considerat relevanta si rata profesorilor instruiti din sistemul de invatamant, dar si rata de somaj pe care am considerat-o reprezentativa pentru determinarea parcursului celor care abandoneaza scoala pe parcursul ciclurilor acestora.

Astfel, am avut initial 42 de date, iar dupa ce am eliminat outlierii am ajuns la 31 de observatii si 12 indicatori.

**Prezentarea indicatorilor**

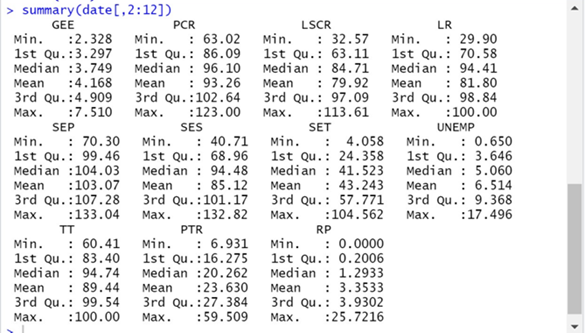
* **GEE-** Government expenditure on education, total (% of GDP)= Investitia guvernului in sistemul de invatamant
* **PCR**- Primary completion rate, total (% of relevant age group)= Procentul persoanelor ce finalizeaza ciclul primar
* **LSCR**- Lower secondary completion rate, total (% of relevant age group)= Procentul persoanelor ce finalizeaza ciclul gimnazial
* **LR**- Literacy rate, adult total (% of people ages 15 and above)= Rata de alfabetizare
* **SEP**- School enrollment, primary (% gross)= Procentul de elevi care incep ciclul primar
* **SES**- School enrollment, secondary (% gross)= Procentul de elevi ce incep ciclul gimnazial
* SET- School enrollment, tertiary (% gross)= Procentul de elevi ce incep ciclul preuniversitar
* UNEMP- Unemployment, total (% of total labor force)= Rata somajului
* TT- Trained teachers in primary education (% of total teachers)= Profesori de specialitate
* PTR- Pupil-teacher ratio, primary= Rata copil-profesor in ciclul primar (masoara decati copii se ocupa in medie un profesor)
* RP- Repeaters, primary, total (% of total enrollment)= Repetenti ciclul primar

**Sursa datelor:**

* Am ales date de pe World Bank. Am cautat ca acestea sa fie cat mai recente, tocmai de aceea am ales acesti idicatori reprezentativi pentru domeniul educatiei. Datele sunt din 2018. In urma eliminarii tarilor pentru care datele au fost incomplete, nu am gasit date in toate tarile pentru unii indicatori, ramanand doar 79 de observatii.

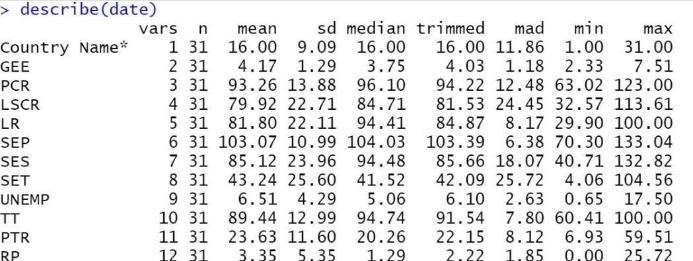
<https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS?view=chart> <https://data.worldbank.org/indicator/SE.PRM.CMPT.ZS?view=chart> <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS?view=chart> <https://data.worldbank.org/indicator/SE.PRM.ENRR?view=chart> <https://data.worldbank.org/indicator/SE.SEC.ENRR?view=chart> <https://data.worldbank.org/indicator/SE.TER.ENRR?view=chart> <https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?view=chart> <https://data.worldbank.org/indicator/SE.PRM.TCAQ.ZS?view=chart> <https://data.worldbank.org/indicator/SE.PRM.ENRL.TC.ZS?view=chart> <https://data.worldbank.org/indicator/SE.PRM.REPT.ZS> <https://data.worldbank.org/indicator/SE.SEC.CMPT.LO.ZS>

**Statistici descriptive**

****

*Fig 1. Statistici descriptive pentru datele initiale*

Observam faptul ca in majoritatea cazurilor, indicatorii alesi iau valori destul de uniform distribuite, caci se pastreaza de cele mai multe ori o distanta aproximativ egala intre minim, quartile si maxim. Exceptie face rata repetentilor pt scoala primara unde maximul este mult mai mare fata de a 3-a quartila si rata somajului unde, de asemenea maximul este semnificativ mai mare fata de a 3-a quartila. Acest lucru indica faptul ca avem putine valori mari pentru ratele somajului in setul de date analizat.

A picture containing text, receipt

Description automatically generated

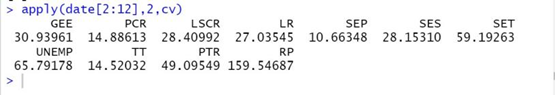
*Fig 2.Statistici descriptive*

Remarcam ca pentru indicatorii GEE, SET, UNEMP, PTR si RP avem valori pozitive pentru skewness, ceea ce face sa avem asimetrie la dreapta, adica in cazul acestor indicatori predomina valorile mari. La polul opus, ceilalti indicatori au valori negative pentru skewness, deci manifesta asimetrie la stanga, ceea ce denota faptul ca predomina valorile mici in cazul lor.

In ceea ce priveste kurtosis avem o singura valoare mai mare decat 3, in cazul RP, ceea ce face ca pentru acest indicator distributia sa fie platikurtika, adica valorile in cazul acestuia nu se grupeaza in jurul mediei. In rest, pentru toti ceilalti indicatori alesi distributiile sunt leptokurtice, valorile grupandu-se in jurul mediei.

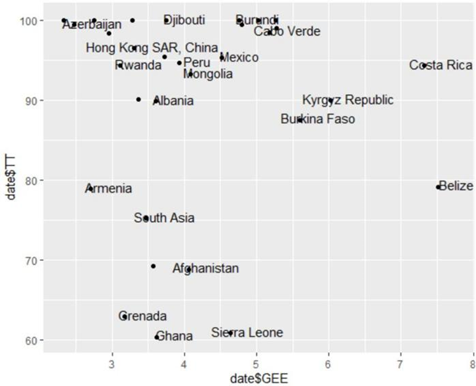
Este de remarcat ca cea mai mare amplitudine se manifesta in cazul indicatorului SET, intrucat acesta masoara procentul elevilor ce incep ciclul preuniversitar de studii. Diferentele pentru acesta sunt semnificative, insa relativ uniform distribuite.

**Coeficientul de variatie**



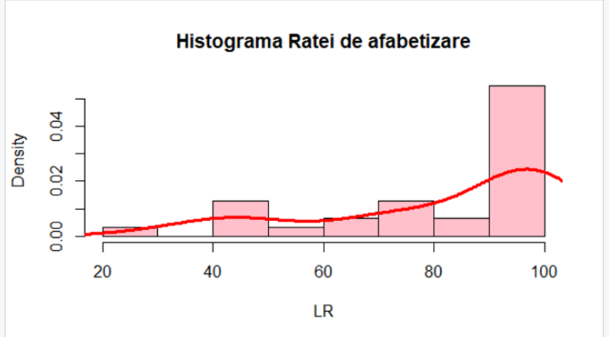
*Fig. 3 output Coeficientul de variatie*

In urma calculului coeficientului de variatie, observam ca acesta ia valori mai mici de 35% in cazul urmatorilor indicatori: GEE, PCR, LSCR, LR, SEP, SES Si TT, adica majoritatea indicatorilor alesi. Pentru acestia, data fiind valoarea coeficientului de variatie, media este reprezentativa si datele sunt omogene. In schimb, pentru ceilalti 4 indicatori, valorile sunt mari de 35% , spre exemplu pentru SET este 59.19% , pentru UNEMP este 65,79% mediile nu sunt representative si datele sunt heterogene . O valoare exagerat de mare remarcam in cazul indicatorului RP, pentru care si diferenta de la a 3-a quartila la maxim era foarte mare.

**

*Fig 5. ggplot GEE-TT*

Pe graficul de mai sus putem observa o distributie neuniforma a tarilor luate in analiza, ceea ce se datoreaza tocmai faptului ca am inclus in modelul de date tari din zone total diferire ale Globului. Putem observa ca Azerbaijan-ul are un grad foarte mare de profesori specializati, in ciuda faptului ca procentul din PIB alocat de catre Guvern catre invatamant este printre cele mai reduse. Tarile fara traditie in invatamant cum sunt Ghana si Afganistan, afectate de-a lungul anilor de probleme mult mai severe se observa ca s-au preocupat mult mai mult de solutionarea acelor probleme caci observam ca atat nivelul de specializare a profesorilor este mic( putem pune acest aspect sub faptul ca cei mai buni specialist, cu un nivel intelectual mai ridicat fata de masa au preferat sa paraseasca aceste tari din cauza conditiilor precare de trai). De asemenea, in cazul acestora nivelul de fonduri alocate invatamantului este destul de mic , aspect sub care ar trebui revenit caci educatia cladeste viitorul si ar putea ajuta la civilizarea populatiei si, ulterior depasirea crizelor prin care trec, caci cu o populatie civilizata si inzestrata intelectual conflictele militare s-ar solutiona cu mult mai mare usurinta . Observam ca pentru Costa Rica avem valori destul de mari in cazul ambilor indicatori.

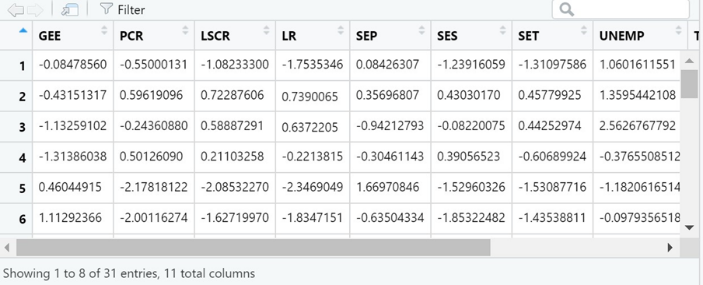


*Fig. 6. Histograma Ratei de alfabetizare*

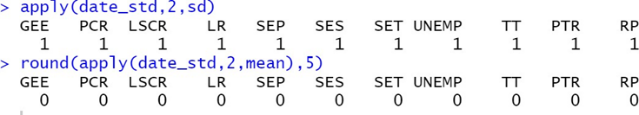
Observam ca distributia ratei de alfabetizare este asimetrica la dreapta. Predomina valorile foarte mari, ceea ce este foarte bine la nivel global. Majoritatea valorilor sunt semnificativ mai mari decat media, ceea ce ne indica evolutia intelectuala la nivel global, un nivel superior de educatie pentru majoritatea tarilor luate sub analiza.

**Prelucrarea datelor**

Standardizam variabilele cu functia scale pentru a putea continua analiza.



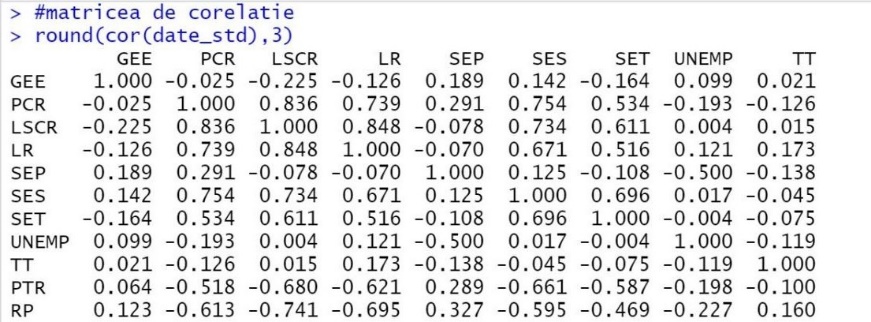
*Fig. 10- Datele standardizate*



*Fig. 7- Media si abaterea standard pentru datele standardizate*

Remarcam faptul ca dupa calculul mediei si a abaterii standard pentru datele standardizate, acestea sunt constante pentru toti indicatorii, abaterea fiind 1, iar media 0. Acest procedeu ne asigura ca am standardizat datele, caci aceasta este o trasatura a datelor standardizate.

**Matricea de corelatie**



*Fig. 8. Matricea de corelatie pentru datele standardizate*

Observam ca exista coeficienti mari de corelatie in cateva cazuri, de pilda intre LR si LSCR de 0,848, intre acesti doi indicatori formandu-se o legatura directa si puternica. Legaturi directe si puternice mai exista si in cazul LSCR si PCR, de 0,836. La polul opus, legaturi indirecte am identificat destul de puternice intre RP si LSCR, acesti doi indicatori fiind destul de corelati, la fel ca si in cazul corelatiei dintre LR si RP, aceasta fiind tot inversa si destul de puternica. Am identificat si valori ale coeficientului de corelatie foarte mici, si valori medii cee ace ne indica faptul ca indicatorii nu sunt foarte bine corelati intre ei.

**Matricea de corelatie**

Chart

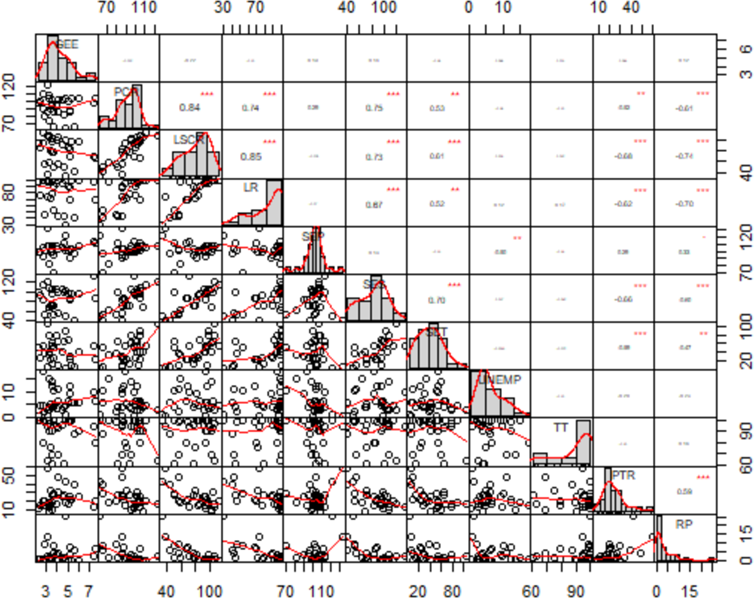
Description automatically generated

*Fig. 9-Corelatie*

Pe figura 22 vom identifica legaturile inverse. Legaturi inverse puternice avem, de pilda, intre RP si LSCR (de -0,74) si intre PTR si LSCR(de -0,68).

Observam din reprezentarea facuta in figura 9 ca avem corelatii puternice intre PCR si LSCR si intre SES- LCR, SES-PCR. Acestea manifesta legaturi directe puternice.

**Performance analytics**



*Fig.11. Performance analytics*

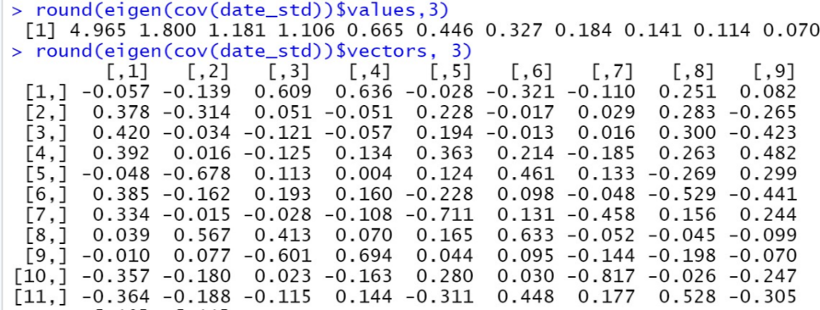
Pe diagonala principala regasim histogramele tuturor indicatorilor pe care sunt trasate densitatile de probabilitate. Putem spune ca cea mai uniforma distributie se gaseste pentru indicatorul SEP, ceea ce e de asteptat intrucat este un indicator ce se refera la ciclul primar de invatamant, unde nu ar trebui sa existe discrepante prea mari intre sistemele de invatamant de la nivel global in scopul de a avea o populatie mondiala uniforma in ceea ce priveste studiile minime de baza cel putin.

Sub diagonala principala avem ploturile, graficele si dreapta de regresie. O dependenta directa putem identifica in cazul SES -LSCR si chiar PC-LSCR. Dependente liniare inverse avem intre PTR-LR, insa aceasta nu este foarte bine conturata. In schimb, avem foarte multe dependente neliniare.

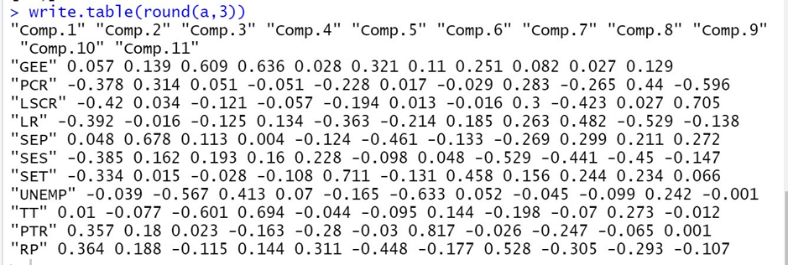
Deasupra diagonalei principale putem identifica coeficientii de corelatie si nivelurile de semnificatie. Pentru cei mai multi indicatori nivelul de semnificatie este maxim, ceea ce inseamna ca sunt reprezentativi in modelul luat spre analiza pentru a determina calitatea invatamantului la nivel global.

***Analiza componentelor principale***

**Valorile proprii ale matricei de covarianta**



*Fig. 13 Vectori si valori propri ale matricei de covarianta*

**

*Fig. 14. Componente principale*

**Forma generala a componentelor principale**

W1=0,057\*GEE-0,378\*PCR-0,042\*LSCR-0,392\*LR+0,048\*SEP-0,385\*SES-0,334\*SET- 0,039\*UNEMP+0,01\*TT+0,357\*PTR+0,364\*RP

W2=0,139\*GEE+0,314\*PCR+0,034\*LSCR-0,016\*LR+0,678\*SEP+0,162\*SES+0,015\*SET- 0,567\*UNEMP-0,077\*TT+0,18\*PTR+0,188\*RP

W3=0,609\*GEE+0,051\*PCR-0,121\*LSCR-0,125\*LR+0,113\*SEP+0,193\*SES- 0,028\*SET+0,0413\*UNEMP-0,601\*TT+0,023\*PTR-0,115\*RP

W4=0,636\*GEE-0,051\*PCR-0,057\*LSCR+0,134\*LR+0,004\*SEP+0,16\*SES- 0,108\*SET+0,07\*UNEMP+0,694\*TT-0,163\*PTR+0,144\*RP

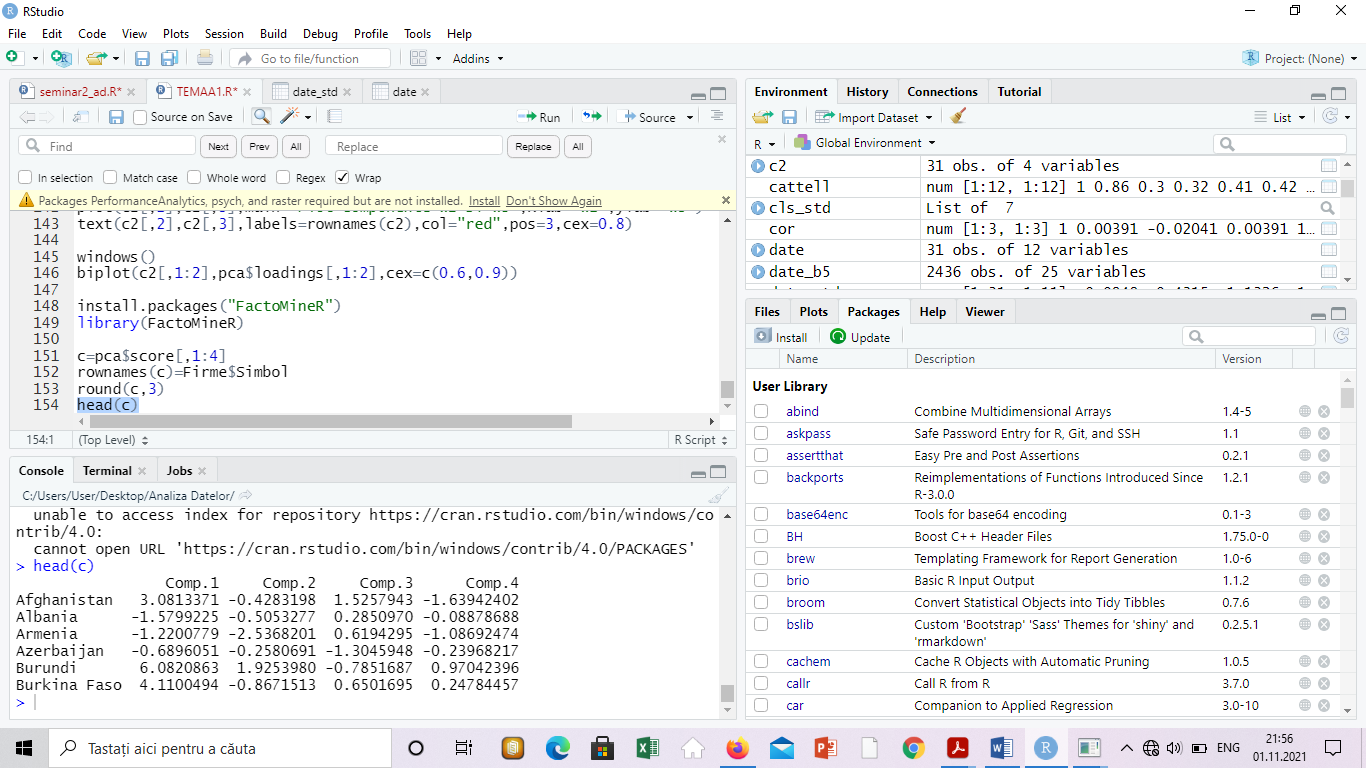
**Scorurile principale**

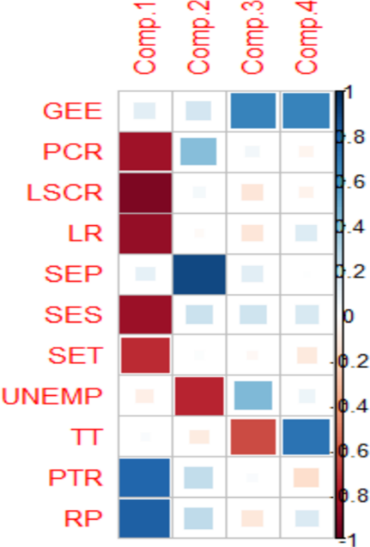
Figura 14. Scorurile principale

In aceasta figura observam scorurile principale.Acestea au fost calculate inmultind vectorii proprii ai matricei de covarianta cu valorile standardizate ale indicatorilor pentru fiecare observatie .

S=0,346\*1.135+0,188\*(-0.468)-0,115\*(-0.575)+0,144\*(-0.302)+0,311\*4.180-0,448\*0.439-0,177\*(-0.565)- 0,528\*(0.618)-0,305\*(0.621)-0.293\*0.821-0.107\*(-194)

**Corelatie intre variabilele originale si componentele principale**

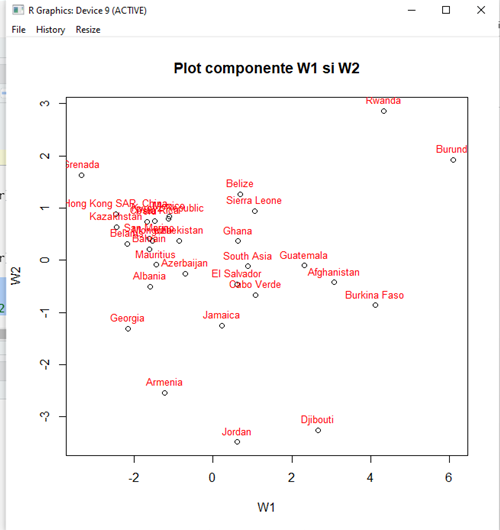
In figura de mai jos am reprezantat grafic legatura dintre indicatori si cele 4 CP retinute in analiza .Componenta 1 se coreleaza puternic indirect cu cei 5 indicatori (PCR, LSCR,LR,SES,SET), iar cu indicatorii PTR si RP exista o corelatie directa puternic. Componenta 2 se coreleaza puternic cu Procentul de elevi care incep ciclu primar (SEP) , cu PCR , iar cu UNEMP se coreaza indirect . Componenta 3 se coreleaza direct puternic cu GEE , UNEMP si indirect cu TT . Iar componenta 4 se coreleaza direct puternic cu GEE si TT.

**

*Fig. 15. Corelatie intre variabilele originale si componentele principale*

Am reprezantat grafic legatura dintre indicatori si cele 4 CP retinute in analiza .Componenta 1 se coreleaza puternic indirect cu cei 5 indicatori (PCR, LSCR,LR,SES,SET), iar cu indicatorii PTR si RP exista o corelatie directa puternic. Componenta 2 se coreleaza puternic cu Procentul de elevi care incep ciclu primar (SEP) , cu PCR , iar cu UNEMP se coreaza indirect . Componenta 3 se coreleaza direct puternic cu GEE , UNEMP si indirect cu TT . Iar componenta 4 se coreleaza direct puternic cu GEE si TT.

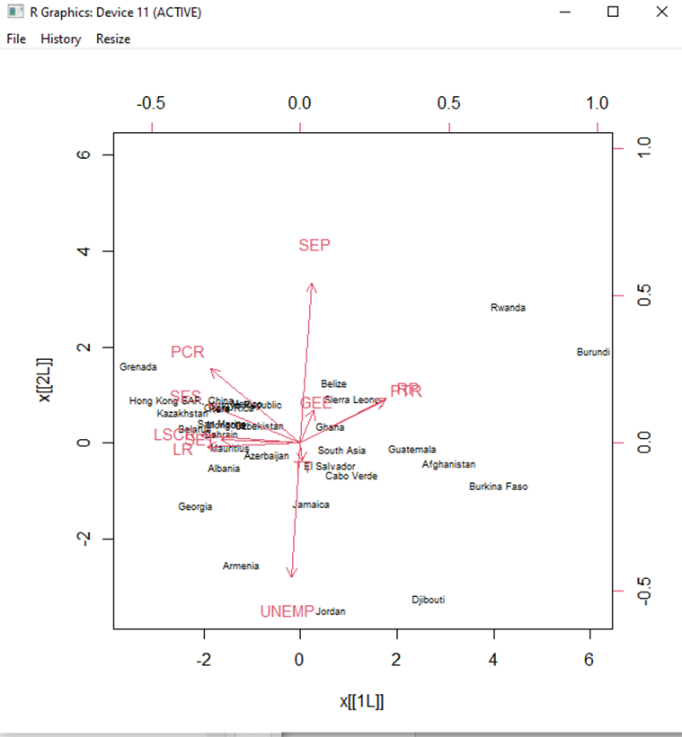
**Plot Componente W1 si W2**



*Fig.16 Plot Componente W1 si W2*

Prima component reflecta rata de alfabetizare.

A doua component reflecta proportia persoanelor care incep studiile primare.

****

*Fig.17 Graficul Biplot*

In figura 17, tarile sunt reprezentati ca puncte in planul format de primele 2 componente principale. In plus , fata de reprezentarea observatiilor , graficul indica variabilele originale ca vectori pornind din origine. Cu privire la vectori , putem analiza 3 elemente importante:

1.Orientarea(directia) vectorului in raport cu spatiul CP .Cu cat este mai paralel cu o axa(cu o componenta principala) , cu atat contribuie mai mult la acea CP. Spre exemplu SEP si UNEMP vor contribui mai mult la componenta 2 .Iar LR si LSCR vor contribui mai mult la componenta 1.

2.Lungimea vectorului . Cu cat vectorul este mai lung, cu atat variabilitatea indicatorului reprezentat de cele 2 CP este mai mare. Spre exemplu vectorul SEP este cel mai lung , deci variabilitatea indicatorului este mai mare.

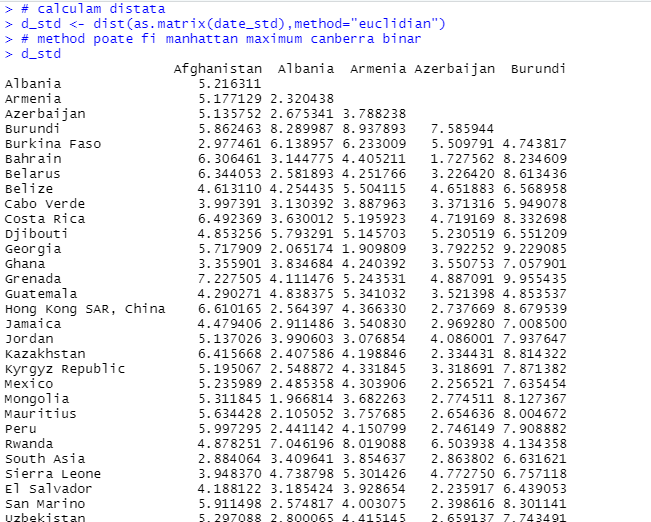
3. Unghiul dintre vectori . Cu cat e mai strans, cu atat corelatia dintre variabile este mai puternica . Spre exemplu corelatia dintre SES si PCR este una puternica Unghiurile drepte denota absenta corelatiei ,vectorii opusi indica legaturi opuse, cum ar fi : SEP si UNEMP .

**Concluzie ACP**

Asadar, la finalul analizei componentelor principale prin metoda ACP am ajuns sa avem 4 componente principale, de la cele 11 cu care am pornit initial. Aceste 4 componente principale obtinute conserva integral informatiile din cele 11 variabile originale, in sens de varianta totala si generalizata. Componentele principale sunt necorelate intre ele doua cate doua si au o varianta maximala descrescatoare. Prin reducerea dimensionalitatii datelor prin metoda ACP am eliminat redundanta informationala.

**Analiza Cluster**

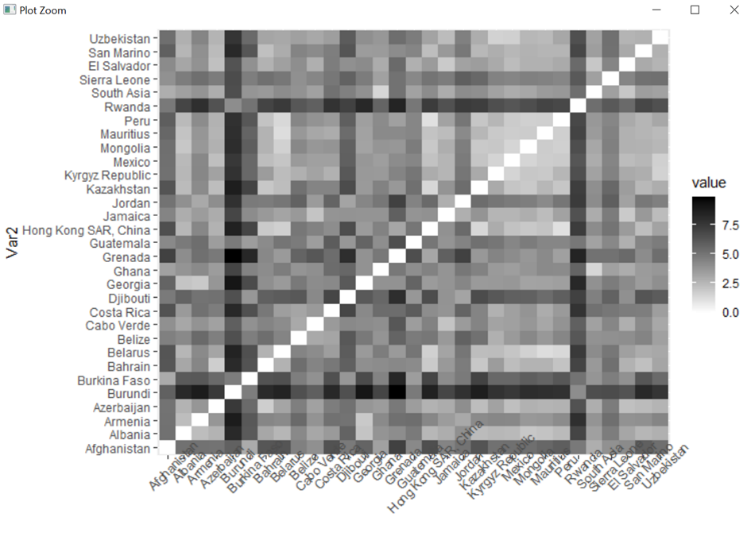
**Matricea Distantelor**



*Figura1- Output matricea distantelor*

In figura de mai sus avem matricea distantelor euclidiene intre toate tarile. De exemplu distanta euclidiana intre Armenia si Albania este 2,32 sau distanta intre Jamaica si Afganistan este 4,4794 . Un alt exemplu este distanta este intre Grenada si Burundi este de 9,9554, fiind o distanta mai mare fata de celelalte .

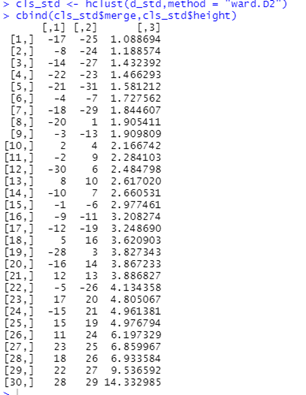
**Distanta euclidiana intre primele 2 obiecte**

**

*Figura. 3- Distante euclidiene-Plot*

In figura 3 putem observa intre ce tari exista o similitudine mai mare sau mai mica. Cu cat distanta euclidiana este mai mica, cu atat exista intre cele doua tari similitudini mai mari, deci tinand cont de legenta plot-ului, culorile mai deschise, mai apropiate de alb reprezinta valori mici si, prin urmare, similitudini mari intre tari. Astfel, putem spune ca avem similitudini mari, de pilda, intre Peru si Guatemala, fapt ce nu ne surprinde, fiind tari destul de similar la modul general, nu doar din punctul de vedere al sistemului de invatamant. Daca luam in considerare si pozitionarea lor geografica in America de Sud, respectiv Centrala, zone cu specific destul de asemanator, deci este firesc ca fondurile alocate pentru sistemul de invatamant sa fie destul de asemanatoare, la fel ca si nivelul de pregatire al elevilor, caci nu sunt mari diferente culturale intre cele doua. Pe cale de consecinta, diferentele obtinute intre valorile indicatorilor luati in considerare nu vor fi foarte mari, ulterior rezultand o diferenta mica intre cele doua per ansamblu. In mod firesc, daca am compara aceleasi doua tari cu tari de pe alt continent, unde exista diferente culturale, strategice sau financiare semnificative, in mod evident distantele euclidiene ar fi mult mai mari si pe cale de consecinta similitudinile mult mai mici. Observam pe grfic nuante inchise pentru foarte multe tari atunci cand se intersecteaza cu Rwanda sau Burundi, indiferent fata de ce alta tara am calcula distanta euclidiana de la acestea. Acest lucru se intampla pentru ca sunt tari Africane foarte slab dezvoltate. Se pare ca in ceea ce priveste sistemul educational acestea sunt codasele clasamentului, caci diferentele intre ele si restul sunt foarte mari. De asemenea, si in cazul acestora sunt localizate una langa cealalta, deci, iarasi, stereotipiile culturale, financiare, contextele sociale si politice isi pun amprenta, facand ca in aceste tari sa nu se aloce probabil suficienti bani sistemului de invatamant, sa nu fie un obicei mersul la scoala, terminarea ciclului gimnazial sau chiar liceal, alocarea unui numar semnificativ de profesori pentru nevoile sistemului, si astfel aceste sisteme educationale se plaseaza cel mai departe de cele ale celorlalte tari luate in considerare in analiza noastra.

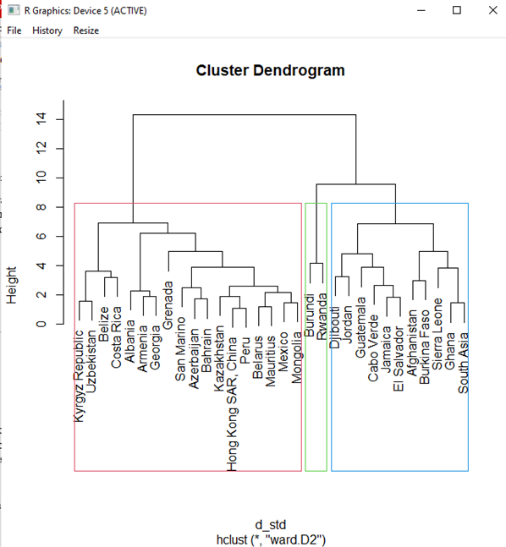
**Metode Ierarhice: Metoda Ward**



*Fig. 4- Clusterizare metoda Ward*

Observam ca pe primele doua coloane avem etichetele la fiecare etapa de clusterizare, iar pe ultima distantele de agregare. In ceea ce priveste distanta de agregare, valorile acesteia observam ca sunt crescatoare de la o etapa la alta. Acest lucru este firesc intrucat metoda Ward este o metoda de clusterizare ierarhica. La prima etapa de clusterizare componentele 17 si 25 au format un cluster la distanta de agregare de 1,0886. La etapa 2 de clusterizare componentele 8 si 24 formeaza un cluster la o distanta de agregare de 1,188574. In etapa 8 de clusterizare, componenta 20 se adauga clusterului 1 deja format, la o distanta de comasare de 1,905411.

**DENDOGRAMA**



*Fig. 10- Dendograma cu evidentierea clusterelor*

In figura de mai sus sunt evidentiate pe dendograma si cele trei clustere. Observam ca cele mai slab dezvoltate tari, asa cum am evidentiat si anterior in analiza, sunt Burundi si Rwanda, tari africane mult mai slab pregatite din punctul de vedere al educatiei din cauza diferentelor culturale, accentului slab plasat pe educatie, dar si a deficitului financiar. In clusterul cu volum mai mare evidentiat cu rosu, adica cel de-al doilea, sunt tarile mai bine dezvoltate din setul nostru de date. De pilda China face parte din acest cluster, ceea ce e firesc daca ne gandim la diferentele culturale si economice dintre China si Jamaica sau Afghanistan.

Dendograma ne releva cate clustere vom lua in considerare in analiza noastra. Se va trasa o taietura care in functie de cate locuri va intersecta, acela va fi numarul de clustere retinut. Vom face taietura unde distanta e cea mai mare intre 2 pasi consecutivi, aceasta va fi pozitionata la nivelul 8 si va intersecta dendograma in 3 locuri, deci vom obtine 3 clustere.

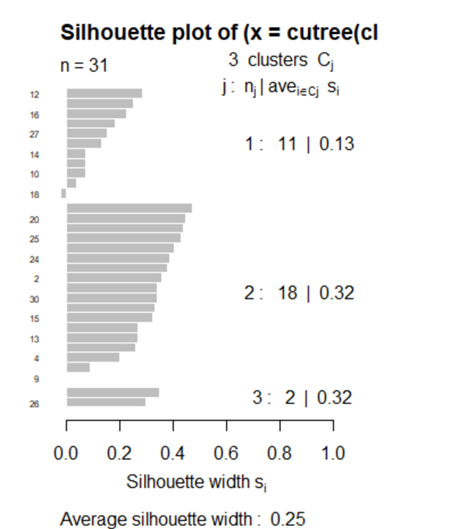
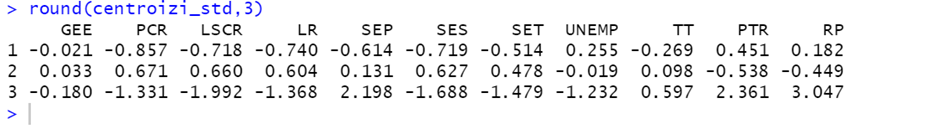
**

Fig. 7- Grafic Silhouette

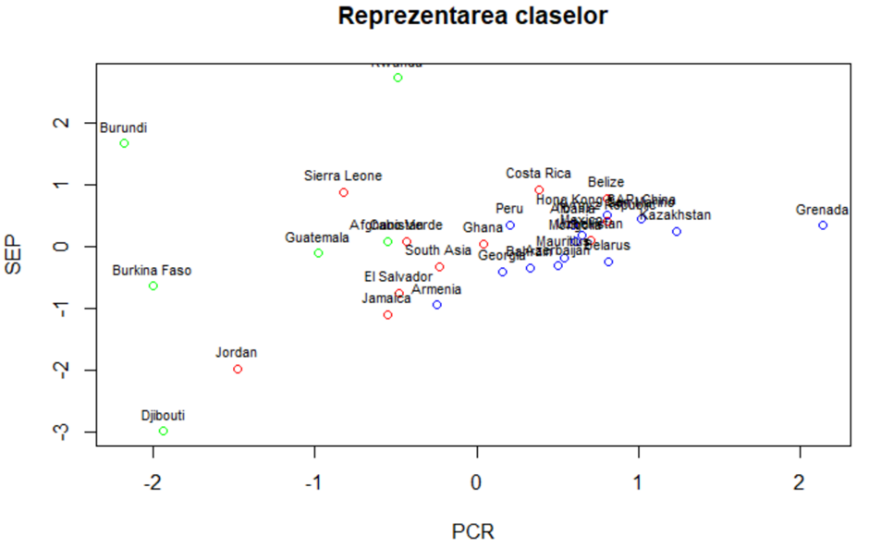
Graficul Silhouette arata cat de bine au fost clasificate datele in clustere, coerenta acestora si nivelul de similitudine al datelor in cadrul unui grup sau intre grupuri. Cele 31 de observatii in 3 clustere. Primul va contine 11 observatii care vor avea media S(i) 0,13, al doilea cluster va contine 18 observatii cu media S(i) 0,32, iar ultimul cluster, cel de-al treilea, doar doua observatii cu media S(i) 0,32.Acuraterea solutiei de clasterizare poate fi apreciata prin graficul siluetei medii , in cazul de fata silueta medie globala atinge o valoare de 0.25 , pentru toate observatiile.

**

*Fig.9- Centroizii clusterelor.*

Centroidul reprezinta valoarea medie a clusterului. Comparand vactorul de medii observam diferentele intre clase, putand sa le caracterizam in aceasta modalitate. Observam valori negative pentru centroizii claselor 1 si 3 pentru investitiile guvernamentale in invatamant, deci aceste tari incadrate in aceste clase ofera investitii din PIB catre sistemul de invatamant mai mici fata de cele din clasa a doua. Continuand cu caracterizarea primului cluster, obtinem tot valori mici pentru persoanele care incep si finalizeaza cilclurile de invatamant, comparativ cu celelalte doua clase. Tot valori mici pentru acesti indicatori sunt si in cazul clasei a treia, spre deosebire de valorile medii pentru clasa a doua. Rata alfabetizare are tot o valoare negativa in cazul claselor 1 si 3 si o valoare semnificativ mai mare fata de acestea in cazul clasei a doua.

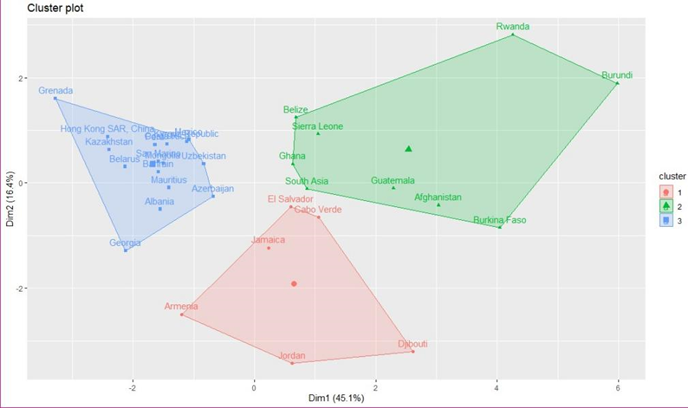
Analizand comparativ cele 3 clustere putem spune ca cea de-a doua clasa este cea care contine cele mai bine dezvoltate tari caci are cele mai mari investitii, dar si cele mai mari rate pentru elevii care termina ciclurile de invatamant. Asadar, putem coclude ca cele mai performante sisteme de invatamant sunt in tarile din clusterul 2. Clusterul 1 cuprinde tarile medii si slab dezvoltate, cu sisteme de invatamant mai putin performante, in timp ce clusterul 3 contine doar tari foarte slab dezvoltate cu sisteme de invatamant foarte slabe ca performanta. Luand in considerare faptul ca cel de- al doilea cluster contine 18 observatii, deci semnificativ mai multe fata de celelalte, care au 11 respectiv doua observatii, putem spune ca sunt mai multe tari bine dezvoltate in setul nostru de date decat slab dezvoltate sau foarte slab dezvoltate.



*Fig. 14- Reprezentarea claselor*

Din reprezentarea claselor, Djiboudi este tara cu cea mai mica rata a persoanelor care termina ciclul primar, dar si cu cel mai mic grad de persoane care se inscriu in ciclul primar, asadar aceasta poate fi caracterizata drept o tara cu un sistem educational foarte slab dezvoltat, cu o educatie precara. Grenada este tara cu cea mai mare rata a persoanelor care incep ciclu primar, dar si cu o rata ridicata a persoanelor care termina in ciclu primar , asadar este o tara cu un sistem educational ridicat , elevii reudind sa termine primul ciclu.

Cum ar fi Borundi are procentul persoanelor cliclu primar ridicat (aproape 2 ) , iar procentul persoanelor ce finalizeaza extrem de scazut , aproape -2 , ceea ce indica ca elevii incep studiile si intre timp renunta , nefinalizandu-le.



*Fig. 15- Reprezentarea claselor*

De aceasta data clusterele sunt reprezentate in planul principal. Pe graficul de mai sus observam cum tarile sunt grupate in functie de specificul fiecarei clase. Observam ca primul cluster, reprezentat pe grafic cu rosu, are un numar destul de mic din observatii ce sunt destul de diferite, variaza destul de mult cele doua coordonate. Cel de-al doilea cluster este reprezentat pe grafic cu varde, dupa cum indica legenda, tarile fiind reprezentate prin triunghiuri. Acesta are o acoperire mai mare decat primul cluster si contine si ceva mai multe observatii decat in cazul anterior, insa seminificativ mai putine decat in cazul clusterului cu numarul 3, ce este reprezentat in planul principal cu albastru. Tarile, reprezentate cu patrate, dupa cum ne indica legenda, sunt semnificativ mai multe, observatiile sunt mai putin imprastiate decat in celelalte cazuri. Observam, in plus ca dimensiunea 1 retine 45,1% din informatie si cea de-a doua 16,4%.

Mentionam ca acest grafic nu este in concordanta cu cel anterior din cauza faptului ca am aplicat ulterior aceasta metoda pentru a obtine graficul fviz si l-am introdus ulterior in analiza, si nu de la inceput, si algoritmul K-means a generat desigur alt set de valori aleatorii.

***Concluzii analiza cluster:***

Luand toate cele de mai sus in considerare, putem ajunge la concluzia ca analiza cluster ne releva impartirea setului de date pe care am lucrat in 3 clase, in functie de indicatorii alesi. Daca tinem cont de scopul initial in care am efectuat aceasta analiza, anume a determina calitatea sistemului de invatamant din cele 31 de tari pentru care avem date disponibile in ceea ce priveste cei 11 indicatori alesi, relevanti pentru acest domeniu, putem conchide faptul ca cele 3 clustere se impart in functie de gradul de pregatire academica si de calitate a sistemului de invatamant in o grupa care contine o treime din observatii ce au o calitate buna a sistemului de invatamant, grupa de mijloc care contine aproape toate celelalte observatii ce au un sistem de invatamant slab calitativ si inca o grupa cu doar doua observatii cu un nivel foarte slab. Din pacate, din setul de date pe care l-am avut la dispozitie pentru efectuarea analizei, majoritatea tarilor au un sistem de educatie slab dezvoltat.

**Solutie:**

Asa cum reiese din analizele de mai sus, majoritatea tarilor au un sistem educational slab dezvoltat, o solutie propusa de noi este investirea de catre guvern sau alte asociatii in educatie in tarile slab dezvoltate, eventual realizarea unor reforme educationale privind programa de invatamant. De asemenea, promovarea si incurajarea la educatie in tarile slab dezvoltate este esentiala pentru a creste gradul de educatie la nivel global. Investitia in tehnologie poate fi o metoda de a dezvolta invatamantul in toata lumea.

**Soluții abordate în alte cercetări realizate:**

Nike Carstarphen analizează aceeași temă în lucrarea sa, venind cu soluții ce constă în oferirea unor locuri de muncă absolvenților, motivându-i astfel pe aceștia să-și finalizeze și continue studiile. De asemenea, Carstarphen propune angajarea cadrelor didactice cu o experiență aprofundată în domeniu, ce pot oferi și dezvoltare practică.

( <https://inee.org/sites/default/files/resources/USIP_Report.pdf> )

Chulani Herath propune revizuirea politicii educaționale ce constă în calitatea și cantitatea programelor pentru a putea ajunge la o egalitate globală.

( <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5617683/> )

**Anexa 1-Datele initiale**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Country Name | GEE | PCR | LSCR | LR | SEP | SES | SET | UNEMP | TT | PTR | RP |
| Afghanistan | 4,058869839 | 85,62532806 | 55,34642029 | 43,01971817 | 103,9961624 | 55,42520905 | 9,686420441 | 11,05700016 | 68,903 | 48,78979 | 4,07978 |
| Albania | 3,611720085 | 101,5378571 | 96,33482361 | 98,14115143 | 106,9934235 | 95,43232727 | 54,96133041 | 12,34000015 | 89,92080688 | 17,57287 | 0,84806 |
| Armenia | 2,707590103 | 89,87895966 | 93,29219818 | 95,8902 | 92,71524811 | 83,15068054 | 54,57048035 | 17,49600029 | 78,9372 | 15,41859 | 0,27312 |
| Azerbaijan | 2,473819971 | 100,2199478 | 84,71309662 | 76,9026 | 99,7220993 | 94,48007965 | 27,70849991 | 4,900000095 | 99,52143097 | 15,42982 | 0,12816 |
| Burundi | 4,762020111 | 63,02138901 | 32,57291031 | 29,8976 | 121,4215622 | 48,46501923 | 4,057660103 | 1,447999954 | 100 | 42,52383 | 25,72155 |
| Burkina Faso | 5,603469849 | 65,47892761 | 42,97488022 | 41,22444916 | 96,09037018 | 40,70972824 | 6,501870155 | 6,093999863 | 87,57572937 | 39,72173 | 5,70354 |
| Bahrain | 2,327810049 | 97,87412262 | 94,31429291 | 97,46418762 | 99,36305237 | 98,59906769 | 50,48189926 | 0,649999976 | 100 | 11,92315 | 0,32823 |
| Belarus | 4,794990063 | 104,6179123 | 97,84088898 | 99,75656128 | 100,5002136 | 102,4395294 | 87,42909241 | 4,760000229 | 99,55487823 | 19,23259 | 0,04302 |
| Belize | 7,509890079 | 104,4744797 | 67,19004822 | 68,907 | 111,6968765 | 85,37245941 | 24,54084015 | 6,512000084 | 79,22837067 | 19,78332 | 6,67993 |
| Cabo Verde | 5,181749821 | 87,29020691 | 68,22319794 | 67,90282 | 104,0279694 | 88,15511322 | 23,61651039 | 12,17000008 | 98,55406189 | 21,07361 | 7,74608 |
| Costa Rica | 7,315720081 | 98,59436798 | 70,29882813 | 97,86379242 | 113,2942886 | 132,816925 | 55,20793915 | 9,631999969 | 94,36399078 | 12,2047 | 2,31226 |
| Djibouti | 3,748559952 | 66,38028717 | 49,22512054 | 52,3683 | 70,29769135 | 51,45777893 | 50,89 | 10,25699997 | 100 | 29,37453 | 8,66638 |
| Georgia | 3,574429989 | 95,50823212 | 101,8094482 | 99,829 | 98,63083649 | 105,9787216 | 60,33444977 | 13,78499985 | 69,28191 | 8,98194 | 0,38969 |
| Ghana | 3,617980003 | 93,80911255 | 78,04167938 | 79,03964233 | 103,5684662 | 71,31968689 | 15,6917696 | 4,157000065 | 60,41054916 | 27,24556 | 1,80744 |
| Grenada | 3,173799992 | 123,0012283 | 106,6802979 | 99,252782 | 106,8507614 | 120,1183395 | 104,561882 | 2,7920627 | 62,94918823 | 16,35068 | 2,40243 |
| Guatemala | 2,951819897 | 79,69037628 | 56,40378189 | 45,792 | 101,9026184 | 52,72243118 | 38,79 | 2,40199995 | 98,37829 | 20,26228 | 9,01374 |
| Hong Kong SAR, China | 3,310100079 | 104,4942322 | 106,7035065 | 100 | 108,6176682 | 107,4897919 | 76,92223358 | 2,903000116 | 96,55108643 | 13,34788 | 0,40221 |
| Jamaica | 5,260620117 | 85,61917877 | 82,38955688 | 87,782 | 90,99539185 | 82,2886734 | 31,8922 | 9,104000092 | 100 | 24,79434 | 2,84127 |
| Jordan | 3,284049988 | 72,70050812 | 59,02954865 | 98,22711182 | 81,45890045 | 63,11653137 | 34,41532135 | 14,95899963 | 100 | 18,53699 | 1,15416 |
| Kazakhstan | 2,750819921 | 110,3839035 | 113,6107635 | 99,78163147 | 105,836731 | 114,2437134 | 53,98825073 | 4,824999809 | 100 | 19,63913 | 0,04435 |
| Kyrgyz Republic | 6,026189804 | 104,5162888 | 95,10778809 | 99,58599854 | 107,5654068 | 95,05480957 | 41,26702118 | 5,960000038 | 89,9993 | 24,98946 | 0,01892 |
| Mexico | 4,522819996 | 102,2558899 | 91,48665619 | 95,37991333 | 105,0280685 | 105,103363 | 41,52280045 | 3,282999992 | 95,34085083 | 25,73903 | 0,389 |
| Mongolia | 4,088590145 | 101,7796478 | 105,1724091 | 98,42311859 | 104,0403595 | 97,89 | 65,59544373 | 6,254000187 | 93,27017212 | 30,38318 | 0,06013 |
| Mauritius | 5,02312994 | 100,7911835 | 86,75364685 | 91,32539368 | 101,1075974 | 95,09619904 | 68,2892 | 6,657000065 | 100 | 16,19841 | 0,01004 |
| Peru | 3,931309938 | 96,10073853 | 97,97501373 | 94,40827179 | 106,9451218 | 106,4454117 | 78,902 | 3,390000105 | 94,73621368 | 17,38724 | 2,74198 |
| Rwanda | 3,110830069 | 86,54634857 | 36,7946701 | 73,21559143 | 133,0415497 | 40,89606094 | 6,725719929 | 1,01699996 | 94,40020752 | 59,50859 | 13,58886 |
| South Asia | 3,47919488 | 90,06334686 | 79,69064331 | 72,24388123 | 99,5641 | 70,06026 | 24,17538071 | 5,059690961 | 75,22470093 | 33,24669 | 1,29332 |
| Sierra Leone | 4,635769844 | 81,78533936 | 51,01734161 | 43,20632935 | 112,75457 | 99,902 | 65,892 | 4,416999817 | 60,91664886 | 27,52246 | 1,48204 |
| El Salvador | 3,725229979 | 86,60769653 | 77,42192841 | 89,00859833 | 94,82711029 | 71,66178894 | 29,37192917 | 4,006000042 | 95,45491028 | 26,89473 | 3,78063 |
| San Marino | 3,370330095 | 107,4193497 | 103,2467499 | 99,91642761 | 108,0840225 | 67,85713959 | 42,47126007 | 3,90198 | 90,20407867 | 6,93061 | 0 |
| Uzbekistan | 5,281340122 | 103,0236282 | 95,90341949 | 99,9928894 | 104,2329788 | 94,98806763 | 10,07635021 | 5,736000061 | 99,03334808 | 21,50664 | 0,00125 |

*Tabel 2- Date initiale*

*Anexa 2-Codul R*

*getwd() setwd("C:/Users/Diana/Desktop/sda") getwd()*

*date<- View(date) #statistici descriptive*

*summary(date[,2:12]) library(psych) describe(date)*

*#coef de variatie install.packages("raster") library(sp) library(raster)*

*cv(date$GEE) apply(date[2:12],2,cv)*

*boxplot(date$GEE) plot(date$TT,date$LR) plot(date$PCR, date$SES) abline(lm(date$PCR~date$SES))*

*library(ggplot2) ggplot(date,aes(x=date$GEE, y=date$TT))+ geom\_point()+*

*geom\_text(label=date$`Country Name`,nudge\_x = 0.25, nudge\_y = 0.25, check\_overlap = T)*

*hist(date$TT, col="blue") lines(density(date$TT),col="blue")*

*hist(date$SET)*

*lines(density(date$SET),col="red")*

*a<-date$LR*

*hist(a,freq=F, xlab="LR", main="Histograma Ratei de alfabetizare", col="pink") lines(density(a), col="red", lwd=3)*

*#prelucrarea datelor*

*date\_std<-scale(date[,2:12], scale=TRUE) View(date\_std)*

*apply(date\_std,2,sd) round(apply(date\_std,2,mean),5)*

*#matricea de corelatie round(cor(date\_std),3) #matricea de covarianta round(cov(date\_std),3)*

*functie<-function(x){ a<-x/mean(x) return(a)*

*}*

*functie(date$LR)*

*a<-apply(date[,2:12],2,functie) a*

*round(cor(date[,2:12]),3) round(cor(a),3)*

*round(cov(date[,2:12]),3) round(cov(a),3)*

*round(crossprod(b),2) library(Hmisc)*

*M<-rcorr(as.matrix(date[-1])) M*

*library(corrplot) corrplot(M$r)*

*corrplot(M$r, type="upper", method = "square")*

*corrplot(M$r, type="lower", method = "number", p.mat=M$P)*

*corrplot(M$r, type="lower", method = "number", p.mat=M$P, insig = "blank") corrplot(M$r, type="lower", method = "number", p.mat=M$P, sig.level=0.1, insig = "blank")*

*install.packages("PerformanceAnalytics") library(PerformanceAnalytics) chart.Correlation(date[-1],histogram=TRUE, pch=19)*

*round(cor(date\_std),3) round(cov(date\_std),3)*

*#observam ca pe datele standardizate matricile de corelatie si covarianta au aceleasi valori*

*pca<-princomp(date\_std, cor=TRUE) sdev<-pca$sdev*

*valp<-sdev\*sdev*

*#val proprii ale matr de covarianta e varianta componentelor principale procent\_info<-(valp\*100)/11*

*procent\_cumulat<-cumsum(procent\_info)*

*X<-round(data.frame(sdev, valp, procent\_info, procent\_cumulat),3) X*

*#observam ca primele 4 componente au deviatie standard mai mare decat 1.*

*#vom lua 4 componente principale care explica modelul ales in proportie de 82,294%*

*scree\_plot<-prcomp(date\_std) plot(scree\_plot, type="l", main="screeplot")*

*#cu cat e mai abrupt in capat cu atat pastram mai putine comp in analiza*

*a<-pca$loadings a*

*#vect proprii ai matr de covarianta*

*round(eigen(cov(date\_std))$values,3) round(eigen(cov(date\_std))$vectors, 3) write.table(round(a,3))*

*View(date\_std)*

*c<-pca$scores[,1:4] c*

*rownames(c)<-date$`Country Name` c*

*plot(cos(cerc),sin(cerc),type="l",col="blue",xlab="W3",ylab="W4")*

*text(matricea\_factor[,3],matricea\_factor[,4],rownames(matricea\_factor),col="red",cex=0.7)*

*c2=data.frame(c)*

*dev.new()*

*plot(c2[,1],c2[,2],main="Plot componente W1 si W2",xlab="W1",ylab="W2")*

*text(c2[,1],c2[,2],labels=rownames(c2),col="red",pos=3,cex=0.8)*

*c2=data.frame(c)*

*dev.new()*

*plot(c2[,2],c2[,3],main="Plot componente W2 si W3",xlab="W2",ylab="W3")*

*text(c2[,2],c2[,3],labels=rownames(c2),col="red",pos=3,cex=0.8)*

*windows()*

*biplot(c2[,1:2],pca$loadings[,1:2],cex=c(0.6,0*

*ANEXA 1- Codul R Analiza cluster*

*#tema analiza cluster*

*temaf<-tema[,2:12]*

*date\_std <- scale(temaf, scale=T)*

*row.names(date\_std)= tema$`Country Name`*

*View(date\_std)*

*# calculam distata*

*d\_std <- dist(as.matrix(date\_std),method="euclidian")*

*# method poate fi manhattan maximum canberra binar*

*d\_std*

*date\_std[1:2,]*

*d\_euclid\_2forme <- sqrt((date\_std[1,1]-date\_std[2,1])^2+*

*(date\_std[1,2]-date\_std[2,2])^2+*

*(date\_std[1,3]-date\_std[2,3])^2+*

*(date\_std[1,4]-date\_std[2,4])^2+*

*(date\_std[1,5]-date\_std[2,5])^2+*

*(date\_std[1,6]-date\_std[2,6])^2+*

*(date\_std[1,7]-date\_std[2,7])^2+*

*(date\_std[1,8]-date\_std[2,8])^2+*

*(date\_std[1,9]-date\_std[2,9])^2 )*

*d\_euclid\_2forme*

*library(ggplot2)*

*library(reshape2)*

*m <- melt(as.matrix(d\_std))*

*View(m)*

*# aici vedem distantele dintre fiecare 2 tari in mod mai clar; ia toate combinatiile posibile*

*ggplot(data=m,aes(x=Var1, y=Var2, fill=value))+*

*geom\_tile()+*

*theme(axis.text.x = element\_text(angle = 45))+*

*scale\_fill\_gradient(low="white",high = "black")*

*# acolo unde culoarea e ft deschisa, tarile sunt ft asemanatoare = similitudine mare*

*#2.metode ierarhice*

*#metoda ward*

*cls\_std <- hclust(d\_std,method = "ward.D2")*

*cbind(cls\_std$merge,cls\_std$height)*

*clust\_std = hclust(d\_std, method = "ward.D2")*

*cbind(clust\_std$merge,clust\_std$height)*

*plot(clust\_std,labels=rownames(date\_std))*

*library(ggplot2)*

*library(factoextra)*

*fviz\_nbclust(date\_std, hcut, method = "wss") +*

*geom\_vline(xintercept = 3, linetype = 2)+*

*labs(subtitle = "Elbow method - STD")*

*#aici cautam un punct de la care varianta suplimentara sa fie redusa*

*#linia punctata e pusa pt 3 clase*

*#am pus 4 noi pt ca ne-au reiesit 3 clustere din taietura*

*#axa oy este variabilitatea intre clase*

*#k este nr de clase*

*install.packages("NbClust")*

*library(NbClust)*

*res<-NbClust(date\_std, distance = "euclidean", min.nc=2, max.nc=6,*

*method = "ward.D2", index = "all")*

*#alta met de a identifica nr optim de clase*

*#de la 2 la 6 clase=>calculeaza toti indicii*

*library(cluster)*

*si4\_std <- silhouette(cutree(clust\_std, k = 3), d\_std)*

*plot(si4\_std, cex.names = 0.5)*

*si4\_std*

*library(MASS)*

*centroizi\_std <- tapply(as.matrix(date\_std), list(rep(cutree(clust\_std, 3), ncol(date\_std)), col(date\_std)), mean)*

*colnames(centroizi\_std)=colnames(date\_std)*

*round(centroizi\_std,3)*

*plot(clust\_std,labels=rownames(date\_std))*

*rect.hclust(clust\_std,k=3, border=2:5)*

*#within cluster s of sq= variabilitatea intra clasa pt fiecare clasa*

*#between s of sq= variabilitatea totala intra-clasa*

*#total sum of squares=variabilitatea totala*

*#daca facem suma variab interclasa obtinem variabilitatea totala*

*clasa\_std=k\_std$cluster*

*c\_std=cbind(clasa\_std,round(date\_std,3))*

*c\_std*

*m\_std=data.frame(c\_std)*

*plot(m\_std[,3], m\_std[,6], col=c("red","blue","green","black","magenta","orange")*

*[m\_std$clasa\_std], main="Reprezentarea claselor", xlab=colnames(m\_std[3]), ylab=colnames(m\_std[6]))*

*text(m\_std[,3],m\_std[,6],labels=rownames(m\_std),col="black",pos=3,cex=0.7)*

*#cerculetele sunt in functie de clasa de apartenenta*

*#de ex cea cu buline rosu, rentabilitati mai mici si grad de indatorare mai mic fata de cele cu verde*

*library(factoextra)*

*fviz\_cluster(list(data = date\_std, cluster = clasa\_std))*

*spat\_std=k\_std$totss*

*spaw\_std=k\_std$tot.withinss*

*spab\_std=k\_std$betweenss*

*r\_cls\_std=spab\_std/spaw\_std*

*variab\_std=cbind(spat\_std,spaw\_std,spab\_std,r\_cls\_std)*

*variab\_std*

*k\_std$withinss*

*#spat=suma patratelor abaterilor totale= variab totala(nu se tine cont de impartirea pe clase)*

*#spaw=variab intra clase=suma patratelor abaterilor within*

*#spab=variabilitatea intraclase*

*#vectorul 2 ala cu k\_std withinss*

*#reprez suma acestor elem e variabilitatea totala interclasa*

*#raportul r este variabilitatea spab/spaw*

*#trb sa fie cat mai mare raportul*

*#criteriul general al clasificarii*

*#criteriul general al clasificarii: variab intraclasa sa fie cat mai mica si cea interclasa cat mai mare*

*library(psych)*

*library(ggplot2)*

*#evaluarea puterii de discriminare a variabilelor*

*data*

*ggplot(data) +*

*geom\_segment( aes(x=rownames(data), xend=rownames(data), y=a, yend=b), color="black") +*

*geom\_point( aes(x=rownames(data), y=c1), color=rgb(0.9,0.1,0.1,0.5), size=5 ) +*

*geom\_point( aes(x=rownames(data), y=c2), color=rgb(0.1,0.9,0.1,0.5), size=5 ) +*

*geom\_point( aes(x=rownames(data), y=c3), color=rgb(0.1,0.1,0.9,0.9), size=5 )*

*k=kmeans(date\_std,3)*

*k*

*cls=k$cluster*

*set\_date=cbind(date\_std,cls)*

*df2=data.frame(set\_date)*

*round(df2,3)*

*#setul de date cu var originale standardizate*

*nr = round(nrow(df2)\*.70)*

*a <- sample(seq\_len(nrow(df2)), size = nr)*

*train <- df2[a, ]*

*test <- df2[-a, ]*

*round(train,3)*

*round(test,3)*

*df=data.frame(train)*

*df$cls[df$cls ==1] <- "clasa1"*

*df$cls[df$cls ==2] <- "clasa2"*

*df$cls[df$cls ==3] <- "clasa3"*

*cbind(round(df[,1:11],3),df[,12])*

*library(e1071)*

*model <- naiveBayes(as.factor(df[,12]) ~., data=df[,-12])*

*#ne arata clasele*

*pred <- predict(model, df[,-12], type="class")*

*pred*

*#cu pred facem predictii cu privire la setul de natrenare*

*#reclasificam formele din setul de antrenare*

*pred\_2 <- predict(model, df[,-12], type="raw")*

*pred\_2*

*#obtinem probabilitatile aposteriorice*

*table(pred, df[,12],dnn=c("Prediction","Actual"))*

*#matrice de confuzie=>actual clasa 1 si 2*

*#cu ajutorul acestei matrici putem det gradul de acuratete*

*#gradul de eroare=1- gradul de acuratete*

*pred\_test <- predict(model, test[,-12], type="class")*

*pred\_test*

*pred\_test2 <- predict(model, test[,-12], type="raw")*

*pred\_test2*

*table(pred\_test, test[,12],dnn=c("Prediction","Actual"))*

*#gradul de clasificare/ acuratete=(3+6)/10*

*library(class)*

*#knn vine de la cel mai apropiat k vecin*

*#noi avem nr de observatii din setul de antrenare=24=>radical din 24= 4, cv*

*#am considerat cel mai apropiat k vecin, 4 si 5*

*#verificam gradul de acuratete pt acest clasificator*

*pr <- knn(train[,-12],test[,-12],cl=train[,12],k=3)*

*pr*

*pr2 <- knn(train[,-12],test[,-12],cl=train[,12],k=4)*

*pr2*

*c1 <- table(pr,test[,12])*

*c1*

*#gradul de acuratete va fi suma pe diag principala supra tuturor elem*

*c2 <- table(pr2,test[,12])*

*c2*

*acc <- function(x){sum(diag(x)/(sum(rowSums(x)))) \* 100}*

*acc(c1)*

*acc(c2)*