**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI**

**FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**

**SPECIALIZAREA CIBERNETICĂ ECONOMICĂ**

**PROIECT LA DISCIPLINA INTELIGENŢĂ COMPUTAŢIONALĂ ȊN ECONOMIE**

**Profesori coordonatori: Student:**

CONF.UNIV.DR. GEORGESCU IRINA ALEXANDRA LUNGU VANESA-DENISA

**București**

**2025**

**CUPRINS**

**[CAPITOLUL 1: DESCRIEREA DATELOR ŞI STATISTICI DESCRIPTIVE](#_Toc199291962)**

**[2](#_Toc199291962)**

**[CAPITOLUL 2: CLUSTERIZARE FUZZY](#_Toc199291963)**

**[6](#_Toc199291963)**

**[CAPITOLUL 3: REGRESIE LOGISTICĂ MULTINOMIALĂ](#_Toc199291964)**

**[15](#_Toc199291964)**

**[CAPITOLUL 4: ARBORI DE DECIZIE ŞI DE CLASIFICARE. CURĂŢAREA ARBORILOR](#_Toc199291965)**

**[21](#_Toc199291965)**

**[CAPITOLUL 5: KNN DE CLASIFICARE](#_Toc199291966)**

**[27](#_Toc199291966)**

**[CAPITOLUL 6: REŢELE NEURONALE PENTRU CLASIFICARE](#_Toc199291967)**

**[33](#_Toc199291967)**

[**ANEXE 35**](#_Toc199291968)

# **CAPITOLUL 1**

**DESCRIEREA DATELOR ŞI STATISTICI DESCRIPTIVE**

1.Obiectivul acestei analize este acela de a observa indicatorii din domeniul educational al fiecărei

regiuni la nivel global, pentru a evidenția diferențele dintre:

• Numărul persoanelor cu studii superioare și cele fără studii;

• Persoanele analfabete;

• Rata abandonului școlar.

De asemenea, analiza urmărește să identifice eventualele cauze care contribuie la aceste diferențe

și tendințe. Indicatorii analizați, cum ar fi cheltuielile guvernamentale pentru educație, gradul de

înscriere, rata de finalizare a învățământului primar și secundar, și durata așteptată a școlarizării,

pot oferi perspective valoroase asupra factorilor care influențează succesul sau eșecul sistemelor

educaționale.

Sursa datelor este World Bank, iar perioada analizată este anul 2017.

Variabila de clasificare utilizată în acest context este „Nivel educațional general al țării”, iar scopul său este de a grupa țările în funcție de numărul anilor aşteptaţi de şcolarizare.

1-Nivel scăzut (numărul anilor aşteptaţi de şcolarizare<10)

2-Nivel mediu (numărul anilor aşteptaţi de şcolarizare ȋntre 10 şi 13)

3-Nivel ridicat (numărul anilor aşteptaţi de şcolarizare>10)

A screenshot of a computer screen

AI-generated content may be incorrect.

1. PA (Populația analfabetă, 25-64 ani)

• Media: Numărul mediu de persoane analfabete este de 1.831.097.

• Mediana: Jumătate dintre țări au o populație analfabetă mai mică de 815.456, iar cealaltă

jumătate o populație mai mare.

• Minimum: Cea mai mică populație analfabetă raportată este de 3.026 persoane.

• Maximum: Cea mai mare populație analfabetă raportată este de 3.983.1098 persoane.

2. AAS (Ani așteptați de școlarizare)

• Media: În medie, se așteaptă ca un copil să fie școlarizat timp de 12,12 ani.

• Mediana: Jumătate dintre țări au o durată așteptată a școlarizării mai mică de 12,9 ani, iar

cealaltă jumătate mai mare.

• Minimum: Țara cu cea mai scurtă durată așteptată a școlarizării are doar 4,95 ani.

• Maximum: Cea mai mare durată așteptată a școlarizării este de 13,96 ani.

3. GIIS (Gradul de înscriere în învățământul secundar)

• Media: Gradul mediu de înscriere în învățământul secundar este de 99,90%.

• Mediana: Jumătate dintre țări au un grad de înscriere sub 102,94%, iar cealaltă jumătate

peste.

• Minimum: Cel mai mic grad de înscriere în învățământul secundar este de 20,68%.

• Maximum: Cel mai mare grad de înscriere este de 158,54%.

4. GIIP (Gradul de înscriere în învățământul primar)

• Media: Gradul mediu de înscriere în învățământul primar este de 102,29%.

• Mediana: Jumătate dintre țări au un grad de înscriere mai mic de 101,82%, iar cealaltă

jumătate mai mare.

• Minimum: Cel mai mic grad de înscriere este de 74,74%.

• Maximum: Cel mai mare grad de înscriere este de 143,92%.

5. CGIS (Cheltuielile guvernamentale pentru învățământul secundar)

• Media: Cheltuielile guvernamentale medii pentru învățământul secundar sunt de 10.175,9

milioane USD.

• Mediana: Jumătate dintre țări cheltuie mai puțin de 3.636,1 milioane USD, iar cealaltă

jumătate mai mult.

• Minimum: Cele mai mici cheltuieli raportate sunt de 50,8 milioane USD.

• Maximum: Cele mai mari cheltuieli raportate sunt de 325.625,6 milioane USD.

6. CGIP (Cheltuielile guvernamentale pentru învățământul primar)

• Media: Cheltuielile guvernamentale medii pentru învățământul primar sunt de 8.110,21

milioane USD.

• Mediana: Jumătate dintre țări cheltuie mai puțin de 2.777,78 milioane USD, iar cealaltă

jumătate mai mult.

• Minimum: Cele mai mici cheltuieli raportate sunt de 32,41 milioane USD.

• Maximum: Cele mai mari cheltuieli raportate sunt de 293.636,2 milioane USD.

7. CGE (Cheltuielile guvernamentale totale pentru educație)

• Media: Cheltuielile guvernamentale totale medii pentru educație sunt de 17.448,84

milioane USD.

• Mediana: Jumătate dintre țări cheltuie mai puțin de 4.780,55 milioane USD, iar cealaltă

jumătate mai mult.

• Minimum: Cele mai mici cheltuieli raportate sunt de 2.389,00 milioane USD.

• Maximum: Cele mai mari cheltuieli raportate sunt de 179.812,03 milioane USD.

8. NPIS (Numărul de persoane înscrise în învățământul secundar)

• Media: Numărul mediu de persoane înscrise în învățământul secundar este de 4.996.317.

• Mediana: Jumătate dintre țări au mai puțin de 1.216.214 persoane înscrise, iar cealaltă

jumătate mai mult.

• Minimum: Cel mai mic număr de înscriși este de 3.240 persoane.

• Maximum: Cel mai mare număr de înscriși este de 12.982.194 persoane.

9. NPIP (Numărul de persoane înscrise în învățământul primar)

• Media: Numărul mediu de persoane înscrise în învățământul primar este de 5.335.395.

• Mediana: Jumătate dintre țări au mai puțin de 889.102 persoane înscrise, iar cealaltă

jumătate mai mult.

• Minimum: Cel mai mic număr de înscriși este de 2.005 persoane.

• Maximum: Cel mai mare număr de înscriși este de 14.322.742 persoane.

10. RFIP (Rata de finalizare a învățământului primar)

• Media: Rata medie de finalizare a învățământului primar este de 95,75%.

• Mediana: Jumătate dintre țări au o rată de finalizare sub 96,30%, iar cealaltă jumătate

peste.

• Minimum: Cea mai mică rată de finalizare raportată este de 50,0%.

• Maximum: Cea mai mare rată de finalizare este de 99,7%.

# **CAPITOLUL 2**

**CLUSTERIZARE FUZZY**

A number on a white background

AI-generated content may be incorrect.

1. **Cluster 1** (probabil cel mai puțin dezvoltat):
   * Afghanistan are cea mai mare apartenență (43.8%)
   * Azerbaijan are cea mai mică apartenență (7.1%)
2. **Cluster 2** (probabil țări dezvoltate):
   * Belgium (47.3%) și Australia (40.2%) au cele mai mari scoruri
   * Azerbaijan are doar 27.1% apartenență
3. **Cluster 3** (probabil țări în curs de dezvoltare):
   * Azerbaijan (65.8%) și Albania (50.4%) domină
   * Australia are cea mai mică apartenență (28.8%)



**0.427 sugerează o separare moderată între clustere**, dar nu excelentă

0.141 indică o structură slabă a clusterelor

* Ambele versiuni ale coeficientului indică o **separare sub-optimă** a clusterelor
* Clusterii au o anumită structură (valori > 0), dar nu foarte bine definită

A map of countries/regions with different names

AI-generated content may be incorrect.

**Dim1 (27%)** și **Dim2 (20.8%)** reprezintă primele două componente principale care captează împreună 47.8% din variația datelor.

Clusterizarea fuzzy (probabil Fuzzy C-Means) asociază fiecărei țări o apartenență parțială la mai multe grupuri, dar în grafic este afișată doar **cea mai probabilă** apartenență:

* **Albastru (cluster 1)**: țări precum SUA, China, India, majoritatea țărilor africane și asiatice.
* **Galben (cluster 2)**: țări dezvoltate din Europa de Vest, America de Sud și câteva din Asia (ex: Japonia, Australia, Germania, Canada).
* **Gri (cluster 3)**: multe țări europene, din Orientul Mijlociu și America Latină, poziționate central în grafic (valori moderate pe ambele dimensiuni).
* **Clusterul 1 (Albastru):**
  + Conține țări mari ca suprafață și populație, unele în curs de dezvoltare (India, China) și țări africane. Acest cluster ar putea indica:
    - Țări cu valori mari pe o dimensiune (Dim1 pozitiv), posibil asociată cu populație mare, urbanizare sau alte metrici structurale.
* **Clusterul 2 (Galben):**
  + Țări dezvoltate economic și social (ex. Norvegia, Elveția, Japonia, Canada, Australia).
  + Poziționate în stânga jos, semn al unor valori negative pe Dim1 și Dim2. Probabil acest cluster indică:
    - Țări cu standard de viață ridicat, stabilitate economică, scoruri înalte în educație, sănătate etc.
* **Clusterul 3 (Gri):**
  + Țări din Europa de Est, America Latină, și Asia Centrală.
  + Poziționate central, ceea ce poate sugera că aceste țări au valori moderate față de celelalte două grupuri.

A graph of different colored triangles

AI-generated content may be incorrect.

**Lățimea medie Silhouette de 0.11** indică o calitate slabă a clusterizării, sugerând că:

* Structura clusterelor este **minimală sau aproape aleatoare**
* Punctele de date nu sunt bine grupate
* Granițele dintre clustere sunt neclare

Clusterul 1 (presupunând din layout):

Probabil cel mai bine definit

Clusterul 2 și 3:

Probabil suprapunere semnificativă

Multe puncte cu valori Silhouette apropiate de 0.



Clusterul 1 (61 de observații)

* **Lățime medie Silhouette**: -0.03
  + Valori negative indică o **clusterizare slabă**
  + Multe puncte sunt probabil atribuite greșit acestui cluster
  + Structura internă este **necoerentă**

Clusterul 2 (39 de observații)

* **Lățime medie Silhouette**: 0.50
  + Valori > 0.25 indică o **structură rezonabilă**
  + Punctele sunt bine grupate și separate de alte clustere
  + Cluster **bine definit și omogen**

Calitate Inegală:

Un cluster bun (2) și un cluster problematic (1)

Discrepanța mare sugerează posibile probleme în algoritm sau date

Distribuție Asimetrică:

Clusterul slab este mai mare (61 vs 39 observații)

Posibilă nevoie de reechilibrare

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A list of text on a white background

AI-generated content may be incorrect.

1. **Clusterul 1 are probleme grave**:
   * Toate aceste țări ar aparține mai bine altor clustere (2 sau 3)
   * Clusterul 1 pare a fi un "cluster rezidual" neomogen
2. **Țări cu probleme majore**:
   * Bangladesh are cea mai slabă apartenență (-0.21)
   * China și Pakistan au probleme semnificative
3. **Modelul de clusterizare**:
   * Probabil număr neoptim de clustere
   * Metrică de distanță inadecvată
   * Variabile nesemnificative

| **Țară** | **Silhouette Width** | **Semnificație** |
| --- | --- | --- |
| United States | -0.0187 | Ușor atribuită greșit |
| India | -0.0213 | Ușor atribuită greșit |
| Senegal | -0.0488 | Atribuire moderat nepotrivită |
| Pakistan | -0.0767 | Atribuire nepotrivită clară |
| China | -0.1028 | Atribuire nepotrivită semnificativă |
| Bangladesh | -0.2100 | Atribuire profund nepotrivită |

Secvenţa de cod:

install.packages("cluster")

install.packages("readxl")

install.packages("factoextra")

library(cluster)

library(readxl)

library(factoextra)

# 2. Încărcare date

date <- read\_excel("ICE.xlsx")

tari <- date$Tara

date1 <- date[ , sapply(date, is.numeric)]

rownames(date1) <- tari

# 4.Standardizare

date1 <- scale(date1)

# 5. Clusterizare fuzzy cu `fanny()` (alegem 3 clustere)

res.fanny <- fanny(date1, k = 3, metric = "euclidean", stand = TRUE, memb.exp = 1.5)

# 6. Afișare rezultate

head(res.fanny$membership, 5) # Primele 5 grade de apartenență

res.fanny$coeff # Coeficientul Dunn (calitatea clusterizării)

# 7. Vizualizare clustere (cu etichete)

fviz\_cluster(res.fanny,

ellipse.type = "norm",

repel = TRUE,

palette = "jco",

ggtheme = theme\_minimal(),

legend = "right")

# 8. Analiza siluetei (calitatea fiecărui obiect și globală)

fviz\_silhouette(res.fanny,

palette = "jco",

ggtheme = theme\_minimal())

# 9. Rezultatul clusterizării (țară + cluster)

rezultate <- data.frame(Tara = tari, Cluster = res.fanny$clustering)

print(rezultate)

# 10. Silueta individuală (pentru interpretări avansate)

head(res.fanny$silinfo$widths)

# **CAPITOLUL 3**

**REGRESIE LOGISTICĂ MULTINOMIALĂ**

A screenshot of a computer

AI-generated content may be incorrect.

initial value: 88.987595 — valoarea inițială a funcției de cost (log-likelihood negativ) înainte de antrenare.

iter XX value YY.YYYYY — valoarea funcției de cost după fiecare 10 iterații.

final value: 4.892482 — valoarea finală a funcției de cost, după optimizare (convergență).

* Modelul a pornit de la o funcție de cost mare (88.987595), ceea ce indică o potrivire slabă inițială.
* După **100 de iterații**, a reușit să minimizeze semnificativ funcția de cost până la 4.892482.
* Această scădere **accentuată și stabilizată** arată că **optimizarea a convergat** eficient și că modelul s-a antrenat corect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

Coeficienți (efecte asupra log-odd-urilor)

Modelul prezice probabilitățile pentru trei niveluri educaționale (presupunând că nivelul 1 este categoria de referință):

1. **Pentru nivelul 2 vs nivelul 1**:
   * Fiecare an suplimentar de școlarizare așteptată crește log-odd-urile pentru nivelul 2 cu 22.69
   * Inscrierea în învățământul primar scade log-odd-urile (-0.82)
   * Inscrierea în învățământul secundar scade log-odd-urile (-0.33)
   * Rata de finalizare a învățământului primar scade log-odd-urile (-1.41)
2. **Pentru nivelul 3 vs nivelul 1**:
   * Efectul anilor de școlarizare așteptați este și mai mare (44.46)
   * Toți ceilalți coeficienți au efect negativ, cel mai puternic fiind pentru rata de finalizare a învățământului primar (-3.64)

Erorile standard

Erorile standard sunt relativ mari pentru coeficienții anilor de școlarizare așteptați, sugerând o anumită incertitudine în estimări.

Măsuri de ajustare

* **Devianța reziduală**: 9.78 (o măsură a bunătății de potrivire, valori mai mici indică un model mai bun)
* **AIC**: 29.78 (util pentru compararea modelelor, valori mai mici sunt preferate)

A screenshot of a computer

AI-generated content may be incorrect.

* **Clasa 1**:
  + Corect prezisă de 2 ori
  + 2 predicții greșite (clasificate ca 2)
* **Clasa 2**:
  + Corect prezisă de 7 ori
  + 1 predicție greșită (clasificată ca 3)
* **Clasa 3**:
  + Corect prezisă de 7 ori
  + 1 predicție greșită (clasificată ca 2)

Metricile Globale

1. **Acuratețe (Accuracy)**: 84.21% (16/19 predicții corecte)
   * Interval de încredere 95%: (60.42%, 96.62%)
   * Semnificativ mai bun decât un model aleatoriu (p-value = 0.001)
2. **Coeficientul Kappa**: 0.7467
   * Acord bun între predicții și realitate (valori peste 0.7 sunt considerate bune)

Metricile pe Clase

Clasa 1

* **Sensitivitate:** 100% (detectează toate cazurile reale)
* **Specificitate**: 88.24% (identifică corect non-cazurile)
* **Precizie**: 50% (jumătate din predicțiile pentru clasa 1 sunt corecte)

Clasa 2

* **Sensitivitate**: 77.78%
* **Specificitate**: 90%
* **Precizie**: 87.5%

Clasa 3

* **Sensitivitate**: 87.5%
* **Specificitate**: 100% (nicio falsă identificare ca 3)
* **Precizie**: 100% (toate predicțiile pentru 3 sunt corecte)

A white background with black numbers

AI-generated content may be incorrect.

1. **Predicții clare**: Observațiile 1, 2, 4 și 5 au probabilități foarte apropiate de 0 sau 1, indicând încredere mare în predicție.
2. **Predicții cu incertitudine**: Observațiile 3 și 6 arată probabilități dominante pentru clasa 2, dar cu o oarecare probabilitate (7.08% și respectiv 2.30%) pentru clasa 3.
3. **Model decisiv**: În majoritatea cazurilor, modelul atribuie probabilități foarte apropiate de 0 sau 1, sugerând că este foarte sigur în predicțiile sale.
4. **Nicio ambiguitate majoră**: Nu există cazuri unde modelul să fie puternic indecis între două clase (de exemplu, 45%-55%).

A graph of a graph

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.

 **Roșu (Clasa 1)** și **Albastru (Clasa 3):**

* Se suprapun și ating colțul stânga-sus.
* Asta indică un **model perfect sau aproape perfect** pentru aceste două clase (AUC ≈ 1.0).
* Regresia logistică face **clasificări extrem de precise** pentru aceste clase.

 **Verde (Clasa 2):**

* Are o curbă în trepte, care nu atinge colțul din stânga sus.
* Asta sugerează că regresia logistică **nu separă la fel de bine** această clasă de celelalte.
* Posibil modelul are dificultăți în a identifica trăsături distinctive pentru clasa 2, sau clasa este dezechilibrată.



* **Clasa 1: AUC = 1.000000** → Modelul clasifică perfect această clasă.
* **Clasa 2: AUC = 0.955556** → Modelul clasifică foarte bine această clasă, dar nu perfect.
* **Clasa 3: AUC = 1.000000** → Clasificare perfectă și pentru această clasă.

Modelul de regresie logistică are o performanță excelentă, cu o ușoară scădere pentru clasa 2. Acest lucru poate indica fie că datele pentru clasa 2 sunt mai greu de separat, fie că sunt mai puține mostre sau mai mult zgomot.

Dacă dorești, pot analiza mai în detaliu contextul sau să-ți ofer sugestii de îmbunătățire în funcție de date.

Secvenţa de cod:

library(readxl)

library(nnet)

library(caret)

install.packages("tidyverse")

library(tidyverse)

# 2. Încarcă datele

date <- read\_excel("ICE.xlsx")

# 3. Selectează coloanele utile

date\_filtrat <- date %>%

select(`Ani asteptati de scolarizare`,

`Gradul de inscriere in invatamantul primar`,

`Gradul de inscriere in invatamantul secundar`,

`Rata de finalizare a invatamantului primar`,

`Nivel educațional general al țării`)

# 4. Elimină rândurile incomplete

date\_filtrat <- na.omit(date\_filtrat)

# 5. Transformă variabila țintă într-un factor

date\_filtrat$`Nivel educațional general al țării` <- as.factor(date\_filtrat$`Nivel educațional general al țării`)

# 6. Împarte setul în antrenare/test

set.seed(123)

ind <- createDataPartition(date\_filtrat$`Nivel educațional general al țării`, p = 0.8, list = FALSE)

train\_data <- date\_filtrat[ind, ]

test\_data <- date\_filtrat[-ind, ]

# 7. Antrenează modelul de regresie logistică multinomială

model\_multinom <- multinom(`Nivel educațional general al țării` ~ ., data = train\_data)

# 8. Rezumă modelul

summary(model\_multinom)

# 9. Preziceri pe setul de testare

predictii <- predict(model\_multinom, newdata = test\_data)

# 10. Matrice de confuzie

conf\_matrix <- confusionMatrix(predictii, test\_data$`Nivel educațional general al țării`)

print(conf\_matrix)

# 11. Probabilități pentru fiecare clasă

predictii\_prob <- predict(model\_multinom, newdata = test\_data, type = "probs")

head(predictii\_prob)

library(pROC)

library(ggplot2)

# 1. Obține probabilitățile prezise pentru fiecare clasă

predictii\_prob <- predict(model\_multinom, newdata = test\_data, type = "probs")

# 2. Listă pentru a stoca curbele ROC

roc\_list <- list()

# 3. Generare curbe ROC One-vs-Rest pentru fiecare clasă

for (clasa in colnames(predictii\_prob)) {

roc\_obj <- roc(

response = as.numeric(test\_data$`Nivel educațional general al țării` == clasa),

predictor = predictii\_prob[, clasa]

)

roc\_list[[clasa]] <- roc\_obj

}

# 4. Plotare toate curbele ROC într-un singur grafic

plot(roc\_list[[1]], col = "red", main = "Curbe ROC One-vs-Rest")

for (i in 2:length(roc\_list)) {

lines(roc\_list[[i]], col = i + 1) # Culori diferite

}

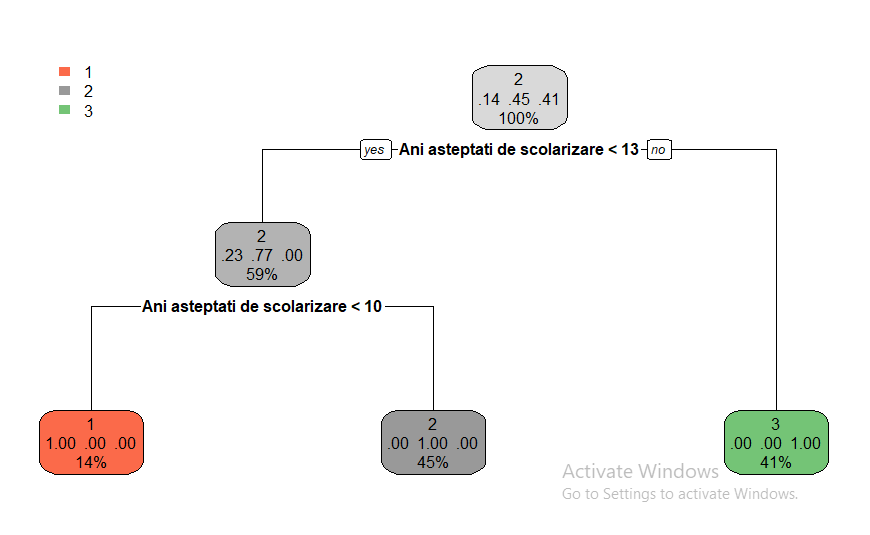
legend("bottomright", legend = names(roc\_list), col = 2:(length(roc\_list)+1), lty = 1)

# 5. Afișare AUC pentru fiecare clasă

sapply(roc\_list, auc)

# **CAPITOLUL 4**

**ARBORI DE DECIZIE ŞI DE CLASIFICARE. CURĂŢAREA ARBORILOR**



**Nodul 1 (rădăcina arborelui)**

* În nodul 1 sunt **100 de observații** în total.
* Distribuția pe clase este:
  + Clasa 1: 14 observații (14%)
  + Clasa 2: 45 observații (45%)
  + Clasa 3: 41 observații (41%)
* **Vectorul de probabilitate**: (0.14, 0.45, 0.41)
* **Clasa dominantă**: Clasa **2** – nivel educațional **mediu**

**Nodul 2 (stânga, condiția: Ani așteptați < 13)**

* În nodul 2 sunt **59 de observații**.
* Distribuția:
  + Clasa 1: 23 observații (39%)
  + Clasa 2: 36 observații (61%)
  + Clasa 3: 0 observații (0%)
* **Vectorul de probabilitate**: (0.39, 0.61, 0.00)
* **Clasa dominantă**: Clasa **2**

**Nodul 3 (sub-nod stânga: Ani așteptați < 10)**

* În nodul 3 sunt **14 observații**.
* Distribuția:
  + Clasa 1: 14 observații (100%)
  + Clasa 2: 0
  + Clasa 3: 0
* **Vectorul de probabilitate**: (1.00, 0.00, 0.00)
* **Clasa dominantă**: Clasa **1** – nivel educațional **scăzut**

**Nodul 4 (sub-nod dreapta: Ani așteptați ≥ 10 și < 13)**

* În nodul 4 sunt **45 de observații**.
* Distribuția:
  + Clasa 1: 0
  + Clasa 2: 45 observații (100%)
  + Clasa 3: 0
* **Vectorul de probabilitate**: (0.00, 1.00, 0.00)
* **Clasa dominantă**: Clasa **2**

**Nodul 5 (dreapta: Ani așteptați ≥ 13)**

* În nodul 5 sunt **41 de observații**.
* Distribuția:
  + Clasa 1: 0
  + Clasa 2: 0
  + Clasa 3: 41 observații (100%)
* **Vectorul de probabilitate**: (0.00, 0.00, 1.00)
* **Clasa dominantă**: Clasa **3** – nivel educațional **ridicat**

A number grid with black text

AI-generated content may be incorrect.

* **Clasa 1**: toate cele 7 instanțe reale au fost clasificate corect ca fiind clasa 1.
* **Clasa 2**: 21 din 22 instanțe au fost clasificate corect, 1 a fost clasificată greșit ca fiind clasa 3.
* **Clasa 3**: toate cele 20 de instanțe au fost clasificate corect.

Acuratețea=(7+21+20)/(7+22+20)=48/49≈97.96

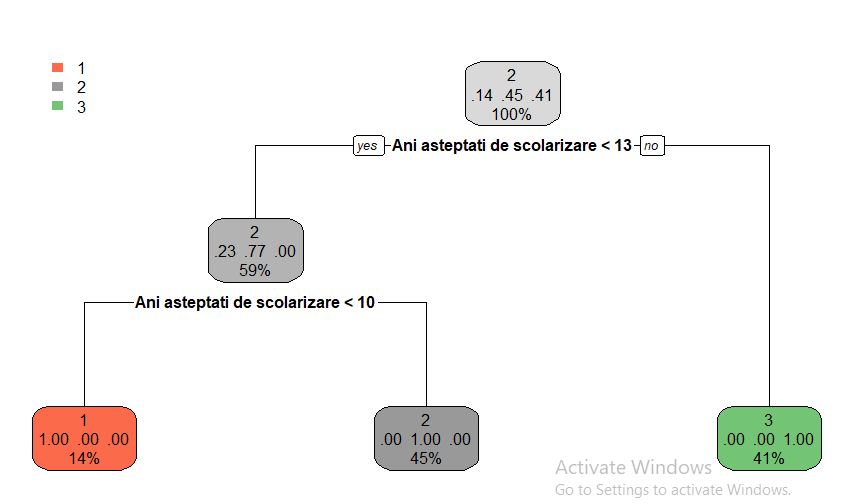


**2.04%** din observațiile din setul de testare **au fost clasificate greșit** de modelul de arbore de decizie.

A graph of a tree

AI-generated content may be incorrect.

Modelul cu cp = 0.05 și 3 frunze are cea mai mică eroare relativă de validare încrucișată și pare a fi cea mai bună alegere.



A number with black text

AI-generated content may be incorrect.



Arborele curăţat este identic cu cel iniţial.

A graph of a graph

AI-generated content may be incorrect.

 Toate cele 3 curbe se apropie foarte mult de colțul stânga-sus al graficului (unde **Sensibilitatea ≈ 1** și **1 - Specificitate ≈ 0**), ceea ce indică **o performanță excelentă**.

 Aceasta sugerează că modelul:

* Are **puține erori de clasificare** (false pozitive și false negative).
* Este **foarte bun în a distinge** între cele 3 clase.



 **AUC = 1.0 (Clasa 1):**  
Performanță perfectă. Modelul distinge perfect clasa 1 de celelalte — **nu există erori de clasificare** pentru această clasă.

 **AUC ≈ 0.977 (Clasa 2):**  
Performanță foarte bună. Modelul are o **capacitate excelentă de a diferenția** clasa 2 de celelalte, cu foarte puține greșeli.

 **AUC ≈ 0.983 (Clasa 3):**  
De asemenea, o performanță **foarte ridicată**. Modelul identifică corect clasa 3 în aproape toate cazurile.

Secvenţa de cod:

library(readxl)

library(rpart)

library(rpart.plot)

library(pROC)

library(caret)

# Citirea fișierului Excel

educatie <- read\_excel("ICE.xlsx")

# Preprocesare

educatie$`Nivel educațional general al țării` <- as.factor(educatie$`Nivel educațional general al țării`)

educatie <- na.omit(educatie)

educatie <- educatie[ , !names(educatie) %in% c("Tara")]

# Împărțirea în set de antrenare și testare

set.seed(123)

split <- createDataPartition(y = educatie$`Nivel educațional general al țării`, p = 0.5, list = FALSE)

train <- educatie[split, ]

test <- educatie[-split, ]

# Construirea arborelui de decizie

arbore <- rpart(`Nivel educațional general al țării` ~ ., data = train, method = "class")

# Vizualizarea arborelui

rpart.plot(arbore, extra = 104)

# Predictii și matrice de confuzie

predictii <- predict(arbore, test, type = "class")

print(paste("Eroare clasificare:", mean(predictii != test$`Nivel educațional general al țării`)))

#matricea de confuzie

confuzie<-table(test$`Nivel educațional general al țării`, predictii)

confuzie

predictie1<-predict(arbore, test, type="prob")

predictie1

# Curățarea arborelui (Pruning)

plotcp(arbore)

cp\_min <- arbore$cptable[which.min(arbore$cptable[,"xerror"]),"CP"]

cp\_min

arbore\_curatat <- prune(arbore, cp = cp\_min)

# Vizualizarea arborelui curatat

rpart.plot(arbore\_curatat, extra = 104)

# Evaluare pe arborele curățat

predictii2 <- predict(arbore\_curatat, test, type = "class")

confuzie2<-(table(test$`Nivel educațional general al țării`, predictii2))

print(paste("Eroare clasificare arbore curățat:", mean(predictii2 != test$`Nivel educațional general al țării`)))

library(pROC)

library(ggplot2)

# Obținem probabilitățile pentru toate clasele

predictii\_prob <- predict(arbore, test, type = "prob")

# Listă pentru a stoca curbele ROC

roc\_list <- list()

# Generăm curbe ROC pentru fiecare clasă (One-vs-Rest)

for (clasa in colnames(predictii\_prob)) {

roc\_obj <- roc(

response = as.numeric(test$`Nivel educațional general al țării` == clasa),

predictor = predictii\_prob[, clasa]

)

roc\_list[[clasa]] <- roc\_obj

}

# Plotăm toate curbele ROC

plot(roc\_list[[1]], col = "red", main = "Curbe ROC One-vs-Rest")

for (i in 2:length(roc\_list)) {

lines(roc\_list[[i]], col = i + 1) # Culori diferite pentru fiecare clasă

}

legend("bottomright", legend = names(roc\_list), col = 2:(length(roc\_list)+1), lty = 1)

# Afișăm AUC pentru fiecare clasă

sapply(roc\_list, auc)

# **CAPITOLUL 5**

**KNN DE CLASIFICARE**

A screenshot of a computer

AI-generated content may be incorrect.

* **Tip model**: k-Nearest Neighbors (k-NN) - algoritm de învățare automată bazat pe similaritate
* **Dimensiune set de date**: 81 de observații (eșantioane)
* **Predictori**: 4 variabile independente
* **Clase țintă**: 3 categorii ('1', '2', '3')
* **Preprocesare**: Datele au fost centrate și scalate (standardizate) pentru toți cei 4 predictori
* **Validare**: Cross-validation pe 10 fold-uri

1. **Performanța Optimă**:
   * Cea mai bună acuratețe (86.88%) și Kappa (0.7739) s-au obținut pentru **k=13**
   * Acest lucru indică faptul că folosirea celor 13 vecini cei mai apropiați oferă cele mai bune rezultate
2. **Tendințe în Performanță**:
   * Performanța crește de la k=5 până la k=13
   * După k=13, performanța scade treptat
   * Pentru k>19, performanța se deteriorează semnificativ
3. **Stabilitatea Modelului**:
   * Performanța rămâne relativ stabilă între k=5 și k=17 (acuratețe peste 85%)
   * Scăderea bruscă pentru k≥19 sugerează că modelul devine prea generalist (underfitting)
4. **Metrici de Evaluare**:
   * **Acuratețe**: Procentul de predicții corecte (86.88% pentru k=13)
   * **Kappa**: Măsoară acordul dintre predicții și realitate (0.7739 pentru k=13), unde:
     + 0.61-0.80 = Acord substanțial
     + 0.81-1.00 = Acord aproape perfect

A graph with lines and numbers

AI-generated content may be incorrect.

1. **Performanța Maximă**:
   * Cea mai mare acuratețe (~0.85) se obține la valorile mici ale lui k (probabil k=5 sau 7)
   * Acuratețea scade treptat pe măsură ce k crește
2. **Tendința Generală**:
   * Există o relație inversă între k și acuratețe în acest interval
   * Scăderea de la ~0.85 la ~0.75 când k crește de la 5 la 15
3. **Implicații Practice**:
   * Alegerea optimă ar fi un k mic (5-7) pentru acest set de date
   * Valorile mai mari ale lui k duc la underfitting (model prea generalist)

A screenshot of a computer

AI-generated content may be incorrect.

Modelul de clasificare multiclasă demonstrează o **acuratețe globală excelentă de 89.47%**, semnificativ mai bună decât un clasificator aleatoriu (p-value = 0.0001592). Coeficientul Kappa de 0.8137 indică un **acord excelent** între predicții și valorile reale.

Clasa 1 (Prevalență 10.53%):

* **Corect clasificate**: 1 (TP)
* **Greșeli**: 1 fals negativ (clasificat ca 2)
* **Sensitivitate**: 50% - cea mai slabă performanță
* **Specificitate**: 100% - nicio falsă pozitivă
* **Precizie**: 100% - toate predicțiile pozitive sunt corecte

Clasa 2 (Prevalență 47.37%):

* **Performanță perfectă în detectare**: Sensitivitate 100%
* **2 erori**:
  + 1 fals pozitiv pentru clasa 1
  + 1 fals pozitiv pentru clasa 3
* **Specificitate**: 80%
* **Precizie**: 81.82%

Clasa 3 (Prevalență 42.11%):

* **Performanță excelentă**:
  + Sensitivitate 87.5%
  + Specificitate 100%
  + Precizie perfectă (100%)

Interval de Încredere 95% pentru Acuratețe: (66.86%, 98.7%)

Performanța este consistent bună și în condiții de incertitudine

Rata de Detectare:

Clasa 2: 47.37% (reflectă prevalența sa ridicată)

Clasa 3: 36.84% (performanță solidă dată prevalența)

Acuratețe Echilibrată:

Clasa 1: 75% (media dintre sensitivitate și specificitate)

Clasa 2: 90%

Clasa 3: 93.75%

Secvenţa de cod:

library(readxl)

library(caret)

library(class)

library(tidyverse)

# 2. Încarcă datele

date <- read\_excel("ICE.xlsx")

# 3. Selectează doar coloanele relevante

date\_filtrat <- date %>%

select(`Ani asteptati de scolarizare`,

`Gradul de inscriere in invatamantul primar`,

`Gradul de inscriere in invatamantul secundar`,

`Rata de finalizare a invatamantului primar`,

`Nivel educațional general al țării`)

# 4. Elimină rânduri cu valori lipsă

date\_filtrat <- na.omit(date\_filtrat)

# 5. Transformă coloana țintă într-un factor (multi-clasă)

date\_filtrat$`Nivel educațional general al țării` <- as.factor(date\_filtrat$`Nivel educațional general al țării`)

# 6. Împarte datele în seturi de antrenare/test

set.seed(123)

ind <- createDataPartition(date\_filtrat$`Nivel educațional general al țării`, p = 0.8, list = FALSE)

train\_data <- date\_filtrat[ind, ]

test\_data <- date\_filtrat[-ind, ]

# 7. Creează modelul KNN

control <- trainControl(method = "cv", number = 10)

model\_knn <- train(`Nivel educațional general al țării` ~ .,

data = train\_data,

method = "knn",

trControl = control,

preProcess = c("center", "scale"),

tuneLength = 10)

# 8. Rezultate model

print(model\_knn)

plot(model\_knn)

# 9. Preziceri și evaluare

predictii <- predict(model\_knn, newdata = test\_data)

conf\_matrix <- confusionMatrix(predictii, test\_data$`Nivel educațional general al țării`)

print(conf\_matrix)

# **CAPITOLUL 6**

**REŢELE NEURONALE PENTRU CLASIFICARE**

**A diagram of a network

AI-generated content may be incorrect.**

**1. Date Demografice de Bază**

* **Populația analfabetă: 25.64% (presupunând că este procentajul)**
* **Ani așteptați de școlarizare: Indicator important al duratei medii de educație**

**2. Indicatorii de Acces la Educație**

* **Înscrieri în învățământ:**
  + **Primar: Gradul de înscriere**
  + **Secundar: Gradul de înscriere**
* **Numărul de persoane înscrise:**
  + **În învățământul primar**
  + **În învățământul secundar**

**3. Investiții în Educație**

* **Cheltuieli guvernamentale:**
  + **Pentru învățământul primar (în milioane USD)**
  + **Pentru învățământul secundar**
  + **Total educație (în milioane USD)**

**4. Eficiența Sistemului Educațional**

* **Rata de finalizare a învățământului primar: Indicator cheie al succesului sistemului**

**Clusterizare Preliminară (C1, C2, C3)**

**Trei clustere sunt sugerate, probabil grupând țări/teritorii după:**

1. **Performanță educațională**
2. **Nivel de investiții**
3. **Rate de acces și finalizare**

**A screenshot of a computer

AI-generated content may be incorrect.**

* + **Majoritatea observațiilor au o probabilitate >98% pentru o singură clasă**
  + **Exemple:**
    - **Randul 5: 98.4% clasa 1**
    - **Randul 10: 98.2% clasa 1**
    - **Randul 15: 98.5% clasa 1**

**Distribuții tipice:**

* + **Clasa 1: Probabilități ~0.98-0.99**
  + **Clasa 2: Probabilități ~0.01-0.02 (când nu este clasa dominantă)**
  + **Clasa 3: Probabilități ~0.005-0.01 (când nu este clasa dominantă)**

**Variabilele predicted\_class și actual\_class:**

* **Perfectă aliniere: Toate predicțiile corespund exact cu valorile reale**
* **Exemplu:**
  + **ID 7: Predicție 2 ↔ Real 2**
  + **ID 8: Predicție 3 ↔ Real 3**
  + **ID 21: Predicție 1 ↔ Real 1**

**Evaluarea Performanței Modelului**

**1. Acuratețe Perfectă:**

* **100% corectitudine pe setul afișat (25 de observații)**
* **Modelul a identificat perfect toate clasele**

**2. Încredere Înaltă:**

* **Probabilități foarte apropiate de 1 pentru clasele corecte**
* **Valori neglijabile (<1%) pentru clasele incorecte**

**3. Tipare de Predicție:**

* **Clasele 2 și 3 apar mai frecvent în date**
* **Clasa 1 apare mai rar (doar la ID 21, 42, 60, 66)**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

* **Clasa 1:**
  + **4 predicții corecte**
  + **0 erori (nici fals negativ, nici fals pozitiv)**
* **Clasa 2:**
  + **11 predicții corecte**
  + **0 erori**
* **Clasa 3:**
  + **11 predicții corecte**
  + **0 erori**

**Metrici Cheie Calculate:**

1. **Acuratețe Globală: 100%**
   * **(4 + 11 + 11) / (4 + 11 + 11) = 26/26**
2. **Precizie pe Clase:**
   * **Toate clasele: 100%**
3. **Recall (Sensitivitate) pe Clase:**
   * **Toate clasele: 100%**
4. **Scor F1 pe Clase:**
   * **Toate clasele: 100%**

**1. Echilibrul Datelor:**

* **Clasa 1: 4 observații (15.4%)**
* **Clasa 2: 11 observații (42.3%)**
* **Clasa 3: 11 observații (42.3%)**
* **Setul este moderat dezechilibrat, cu Clasa 1 subreprezentată**

**2. Semnificația Performanței Perfecte:**

* **Modelul a reușit să:**
  + **Identifice corect toate cazurile rare (Clasa 1)**
  + **Nu a confundat nicio clasă cu alta**
  + **Menținut performanță egală pentru clasele majoritare**

****

**Secvenţa de cod:**

**install.packages("neuralnet")**

**install.packages("readxl")**

**install.packages("caTools")**

**library(neuralnet)**

**library(readxl)**

**library(caTools)**

**# 2. Citirea și curățarea datelor**

**data <- read\_excel("ICE.xlsx")**

**data <- na.omit(data)**

**# 3. Verificare existență coloană**

**if (!"Nivel educațional general al țării" %in% colnames(data)) stop("Coloana 'Nivel educațional general al țării' nu există în fișier!")**

**# 4. Transformare clasă multiclasa în one-hot encoding**

**data$C1 <- ifelse(data$`Nivel educațional general al țării` == 1, 1, 0)**

**data$C2 <- ifelse(data$`Nivel educațional general al țării` == 2, 1, 0)**

**data$C3 <- ifelse(data$`Nivel educațional general al țării` == 3, 1, 0)**

**# 5. Selectăm doar coloane numerice utile (excludem Tara și nivelul educational original)**

**data\_num <- data[sapply(data, is.numeric)]**

**data\_num <- subset(data\_num, select = -c(`Nivel educațional general al țării`))**

**# 6. Standardizare**

**minim <- apply(data\_num, 2, min)**

**maxim <- apply(data\_num, 2, max)**

**data\_scaled <- as.data.frame(scale(data\_num, center = minim, scale = maxim - minim))**

**names(data\_scaled) <- make.names(names(data\_scaled))**

**data\_ready <- cbind(data\_scaled, C1 = data$C1, C2 = data$C2, C3 = data$C3)**

**# 8. Împărțim în train/test**

**set.seed(123)**

**split <- sample.split(data\_ready$C1, SplitRatio = 0.75)**

**train\_data <- subset(data\_ready, split == TRUE)**

**test\_data <- subset(data\_ready, split == FALSE)**

**# 9. Construim formula rețelei neuronale**

**names(data\_scaled) <- make.names(names(data\_scaled))**

**input\_vars <- names(data\_scaled)**

**formula <- as.formula(paste("C1 + C2 + C3 ~", paste(input\_vars, collapse = " + ")))**

**# 10. Antrenare rețea neuronală**

**model <- neuralnet(formula, data = data\_ready, hidden = 5, linear.output = FALSE, threshold = 0.01)**

**plot(model)**

**# 11. Predicții**

**pred <- compute(model, test\_data[, input\_vars])$net.result**

**pred**

**predicted\_class <- apply(pred, 1, which.max)**

**predicted\_class**

**actual\_class <- apply(test\_data[, c("C1", "C2", "C3")], 1, which.max)**

**actual\_class**

**# 12. Evaluare**

**conf\_matrix <- table(Actual = actual\_class, Predicted = predicted\_class)**

**accuracy <- sum(diag(conf\_matrix)) / sum(conf\_matrix)**

**print(conf\_matrix)**

**cat(" Acuratețea rețelei pentru clasificare în 3 clase (nivel educațional):", round(accuracy \* 100, 2), "%\n")**

# **ANEXE**

A table of numbers and letters

AI-generated content may be incorrect.

A table of numbers and letters

AI-generated content may be incorrect.

A table of numbers

AI-generated content may be incorrect.