ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI

FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ

SPECIALIZAREA CIBERNETICĂ ECONOMICĂ

PROIECT

= INTELIGENȚĂ COMPUTAȚIONALĂ ÎN ECONOMIE =

Student : Dadu Maria Alexandra

1. Descrierea datelor, sursa si statistici descriptive.

Obiectivul principal al acestei analize este explorarea și modelarea factorilor care influențează performanța școlară a elevilor din două școli din Portugalia, măsurată prin nota finală (G3). Utilizând tehnici moderne de analiză a datelor și învățare automată, mi-am propus să identific modelele ascunse și relațiile dintre variabilele socio-demografice, familiale și școlare ale elevilor.

Voi grupa elevii în funcție de caracteristici similare prin clusterizare fuzzy, pentru a evidenția tipologii comportamentale sau educaționale. În continuare,voi prezice performanța școlară (nota finală) folosind metode precum: regresie logistică binomială și multinomială, arbori de decizie și regresie, KNN de clasificare, rețele neuronale si voi evalua impactul unor factori precum: timpul de studiu (studytime), numărul de absențe (absences), eșecuri anterioare (failures), educația părinților (Medu, Fedu), genul (sex) sau ocupația mamei (Mjob).

Această analiză urmărește să ofere o înțelegere mai profundă a influenței mediului familial și școlar asupra reușitei academice, cu potențial aplicabil în politicile educaționale sau intervențiile pedagogice.

Nume atribut	**Tip**	**Descriere**
school	Categorial, binar	Școala la care învață elevul GP = elevul învață la școala Gabriel Pereira MS = elevul învață la școala Mousinho da Silveira
G3	Țintă (Target), întreg	Nota finală (numeric: de la 0 la 20) – variabilă de ieșire
absences	Numeric, întreg	Numărul de absențe de la școală (de la 0 la 93)
traveltime	Numeric, ordinal	Timpul de călătorie de acasă la școală: 1 – <15 min., 2 – 15-30 min., 3 – 30-60 min., 4 – >1 oră
studytime	Numeric, ordinal	Timpul de studiu săptămânal: 1 – <2 ore, 2 – 2-5 ore, 3 – 5-10 ore, 4 – >10 ore
failures	Numeric, întreg	Numărul de materii repetate anterior: n dacă $1 \leq n < 3$, altfel 4
Medu	Numeric, ordinal	Nivelul de educație al mamei: 0 – niciunul, 1 – școala primară, 2 – 5–9 clase, 3 – liceu, 4 – studii superioare
Fedu	Numeric, ordinal	Nivelul de educație al tatălui: 0 – niciunul, 1 – școala primară, 2 – 5–9 clase, 3 – liceu, 4 – studii superioare
Mjob	Categorial, nominal	Ocupația mamei: 'profesor', 'sănătate', 'servicii' (administrație, poliție), 'acasă', 'altceva'
sex	Categorial binar	Sexul elevului: 'F' – feminin, 'M' – masculin

age	Numeric, întreg	Vârsta elevului: de la 15 la 22 de ani
Pstatus	Categorial, binar	Statutul de coabitare al părinților: 'T' – trăiesc împreună, 'A' – separați

Table 1 Descrierea atributelor

Sursa datelor: UC Irvine Machine Learning Repository
https://archive.ics.uci.edu/dataset/320/student%2Bperformance

```
> # 1.5 Statistici descriptive pentru variabilele numerice
> describe(select(data, age, Medu, Fedu, traveltime, studytime, failures, absences, G3))
           vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
age           1 139 15.58 0.72     15   15.48 0.00  15  19     4  1.49     3.46 0.06
Medu          2 139  3.01 1.04      3    3.13 1.48   0   4     4 -0.63    -0.73 0.09
Fedu          3 139  2.74 1.09      3    2.81 1.48   0   4     4 -0.25    -1.14 0.09
traveltime    4 139  1.36 0.68      1    1.21 0.00   1   4     3  2.01     3.78 0.06
studytime     5 139  2.07 0.90      2    1.97 1.48   1   4     3  0.69    -0.16 0.08
failures      6 139  0.24 0.67      0    0.06 0.00   0   3     3  2.86     7.49 0.06
absences      7 139  4.86 6.68      2    3.65 2.97   0  54    54  3.58    20.20 0.57
G3            8 139 11.21 4.34     11   11.42 4.45   0  20    20 -0.57     0.34 0.37
```

- **G3:** Nota medie este puțin peste pragul de trecere → indică o distribuție echilibrată, dar cu cazuri de eșec . Asimetria negativă sugerează mai multe note mari decât note mici extreme.

- **Medu (Educația mamei):** Media este 3.01, indicând că mamele au în general între 5–9 clase și liceu. Distribuția negativ asimetrică sugerează un nivel educațional relativ ridicat în rândul mamelor.

- **Fedu (Educația tatălui):** Media este 2.74, puțin mai mic decât la mame, cu educația tatălui distribuită echilibrat. Skew-ul ușor negativ indică prezența predominantă a educației medii și superioare.

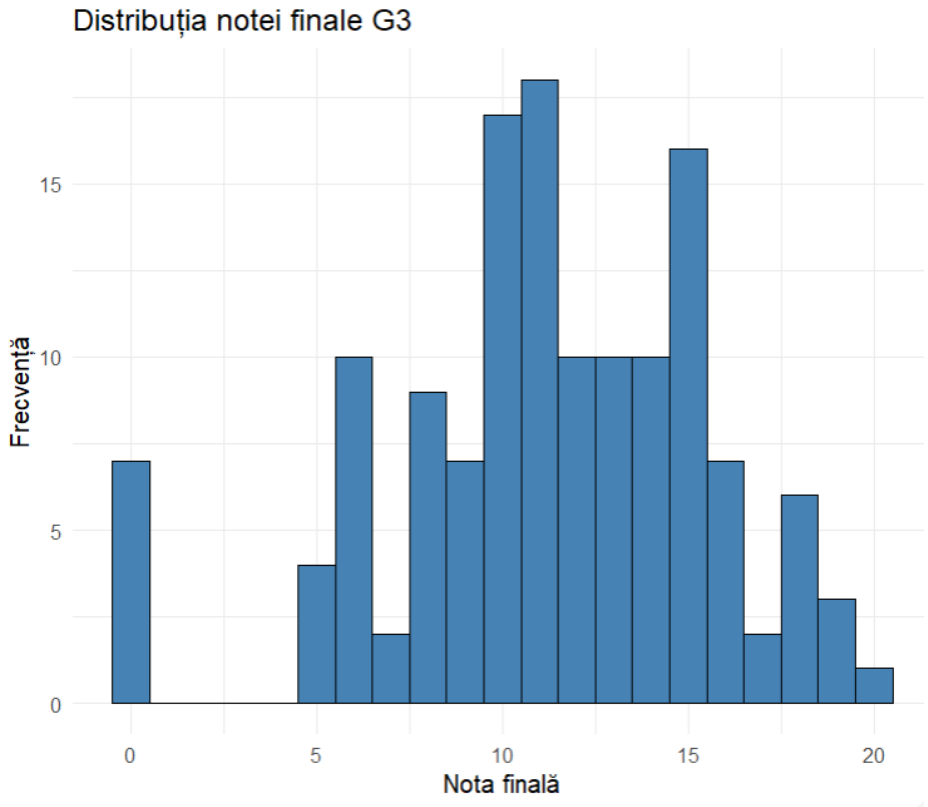
- **Studytime (Timp de studiu):** Media este 2.07, adică majoritatea elevilor studiază 2–5 ore/săptămână. Distribuția pozitiv asimetrică arată că puțini elevi depășesc pragul de 5 ore.

- **Absences (Absențe):** Media este 4.86, dar cu un maxim de 54, ceea ce sugerează cazuri de absențe excesive. Distribuția extrem de asimetrică și kurtosis mare indică prezența unor outlieri ce pot influența analiza.

- **Failures (Eșecuri anterioare):** Media este 0.24, adică majoritatea elevilor nu au avut eșecuri, dar unii au repetat mai multe materii. Skew-ul ridicat (2.86) arată că distribuția este platicurtică.

- **Traveltime (Timp de navetă):** Media este 1.36, ceea ce sugerează că majoritatea elevilor locuiesc aproape de școală. Distribuția asimetrică pozitiv indică puțini elevi care au de parcurs distanțe lungi.

- **Age (Vârsta elevilor):** Vârsta medie este 15–16 ani, specifică nivelului liceal, dar există și elevi mai mari (18–19 ani). Distribuția este asimetrică pozitiv, reflectând cazuri atipice cauzate probabil de repetenție.



Interpretare:

Cele mai frecvente note se situează între 10 și 15, cu un vârf vizibil în zona 11–12. Aceasta indică un nivel mediu de performanță generală în rândul elevilor. Există elevi care au obținut nota 0, semnalând cazuri de eșec total (absență, abandon sau nepromovare), dar și elevi cu nota 20, ceea ce arată performanță maximă. Se observă o ușoară aglomerare spre notele mari, ceea ce înseamnă că mai mulți elevi tind spre succes decât spre eșec total, dar cu o variabilitate semnificativă.

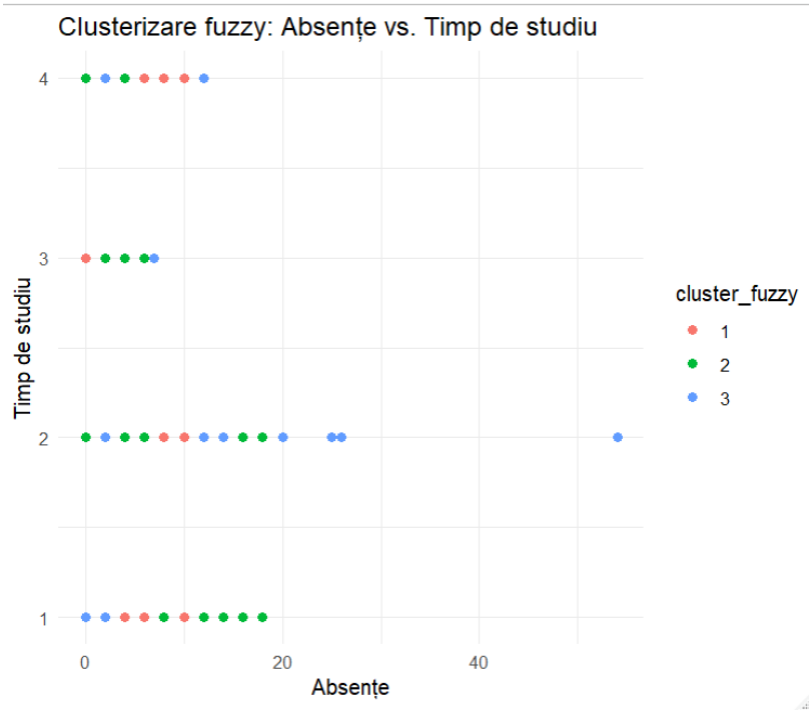
2. Clusterizare fuzzy

```
> set.seed(123)
> fcm_result <- cmeans(data_scaled, centers = 3,139, m = 2,method="cmeans")
> fcm_result
Fuzzy c-means clustering with 3 clusters

Cluster centers:
         age    absences   studytime traveltime   failures
1 -0.5877965 -0.28947036 -0.11163454 -0.3301897 -0.2339607
2  0.4103012  0.04765152  0.06308303 -0.2044603 -0.1012984
3  0.3730015  0.28285499  0.02007316  0.7125737  0.4040146

Memberships:
              1          2          3
  [1,] 0.2200235 0.37949223 0.40048422
  [2,] 0.2058259 0.51387983 0.28029424
  [3,] 0.3127043 0.31611110 0.37118462
  [4,] 0.5581818 0.28343934 0.15837887
  [5,] 0.1374232 0.77831723 0.08425957
  [6,] 0.1831551 0.62478755 0.19205734
  [7,] 0.2896271 0.56492903 0.14544386
  [8,] 0.1726322 0.37657191 0.45079590
  [9,] 0.8358741 0.10938856 0.05473735
 [10,] 0.8358741 0.10938856 0.05473735
 [11,] 0.8358741 0.10938856 0.05473735
 [12,] 0.2923244 0.27821908 0.42945654
 [13,] 0.6214850 0.23231332 0.14620166
 [14,] 0.4521755 0.24703462 0.30078987
 [15,] 0.5583735 0.28051619 0.16111028
 [16,] 0.3199665 0.46098983 0.21904363
 [17,] 0.2247663 0.56967145 0.20556221
 [18,] 0.2132718 0.27217714 0.51455104
```

Cluster 1 : Elevi mai tineri, cu puține absențe, timp redus de studiu, traseu scurt spre școală, fără eșecuri → profil de elev regulat, dar pasiv.

Cluster 2 : Elevi de vârstă medie, cu valori apropiate de medie la toți indicatorii → profil mixt, poate reprezenta elevul "mediu", echilibrat.

Cluster 3 : Elevi mai în vârstă, cu mai multe absențe, timp de navetă mai lung și mai multe eșecuri → posibil profil de risc educațional.



Axa X (Absences) și axa Y (Studytime) sunt analizate în funcție de apartenența fuzzy la clustere.

- Clusterul 3 (albastru) are puncte clar dispersate în zona cu absențe mari și timp de studiu mic (1–2) → elevi cu risc clar.
- Clusterul 1 (roșu) se concentrează pe zona cu puține absențe și valori diferite ale timpului de studiu.
- Clusterul 2 (verde) este intermediar, răspândit uniform → susține ideea unui profil generalist.

Clusterizarea fuzzy reușește să evidențieze trei tipologii educaționale:

1. Elevi tineri, stabili, dar pasivi (roșu).
2. Elevi echilibrați, cu profil neutru (verde).
3. Elevi vulnerabili: navetă mare, absențe, eșecuri (albastru).

3. Problema de regresie logistica binomială

Se creează un model de regresie logistică binomială pentru a prezice probabilitatea ca un elev să fie „cu risc educațional” în funcție de: timpul de studiu, educația părinților, sexul și școala

frecventată. Se consideră că un elev este „cu risc” dacă: a avut cel puțin 2 eșecuri sau are mai mult de 9 absențe.

```
> data$high_risk <- ifelse(data$failures >= 2 | data$absences > 9, 1, 0)
> data$high_risk_f <- factor(data$high_risk)
> model_risk <- glm(high_risk_f ~ studytime + Medu + Fedu + sex + school, data = data, family = binom
ial)
> summary(model_risk)

Call:
glm(formula = high_risk_f ~ studytime + Medu + Fedu + sex + school,
    family = binomial, data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.08604    0.92392   0.093  0.92580
studytime   -0.58039    0.27836  -2.085  0.03707 *
Medu        -0.23129    0.27402  -0.844  0.39863
Fedu        -0.01899    0.26399  -0.072  0.94265
sexM        -0.69346    0.47405  -1.463  0.14351
schoolMS     1.28586    0.47961   2.681  0.00734 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 142.37  on 138  degrees of freedom
Residual deviance: 126.63  on 133  degrees of freedom
AIC: 138.63

Number of Fisher Scoring iterations: 5
```

Timpul de studiu (studytime) este un predictor semnificativ: Cu cât elevul alocă mai mult timp studiului, cu atât scade probabilitatea de a fi considerat "cu risc educațional".

Școala (school) este semnificativă:Elevii de la școala MS au o probabilitate semnificativ mai mare de a fi în risc comparativ cu elevii de la GP.

Educația părinților și sexul elevului nu sunt predictori semnificativi în acest model.

Scăderea de la 142.37 (null deviance) la 126.63 (residual deviance) sugerează că modelul explică o parte din variația din date – este mai bun decât modelul fără predictori, dar nu perfect.

```
> exp(coef(model_risk))  # rapoartele de șanse
(Intercept)   studytime        Medu        Fedu        sexM    schoolMS
  1.0898513   0.5596813   0.7935089   0.9811885   0.4998458   3.6177920
```

Variabilă	Rata de șansă (exp(coef))	Interpretare
(Intercept)	1.0898	Valoare de bază fără predictori
studytime	0.5597	Pentru fiecare unitate în plus la timpul de studiu, șansa de a fi "cu risc" scade cu 44.03% (1 - 0.5597) – efect protectiv

Medu	0.7935	Educația mamei mai mare reduce ușor riscul, dar efectul este slab
Fedu	0.9812	Aproape 1 → educația tatălui nu are impact semnificativ asupra riscului
sexM	0.4998	Elevii băieți au o șansă de aproape 50% mai mică de a fi "cu risc" comparativ cu fetele
schoolMS	3.6178	Elevii de la școala MS au o șansă de 3,6 ori mai mare de a fi în risc comparativ cu cei de la GP → efect clar semnificativ

- **Acuratețea și matricea de confuzie pe setul de antrenare**

```
> pred <- rep("0", dim(setantrenare)[1])
> pred[predict(model_risk, setantrenare, type = "response") > 0.5] <- "1"
> table(pred, setantrenare$high_risk_f)

pred  0  1
   0 74 19
   1  1  3
> acc_train <- mean(pred == setantrenare$high_risk_f)
> print(paste("Acuratețea pe setul de antrenare:", round(acc_train * 100, 2), "%"))
[1] "Acuratețea pe setul de antrenare: 79.38 %"
```

- 74 elevi au fost corect clasificați corect ca „fără risc”
- 19 elevi au fost greșit clasificați ca „fără risc”
- 1 elev a fost fals clasificat ca „cu risc”
- 3 elevi „cu risc” au fost clasificați corect

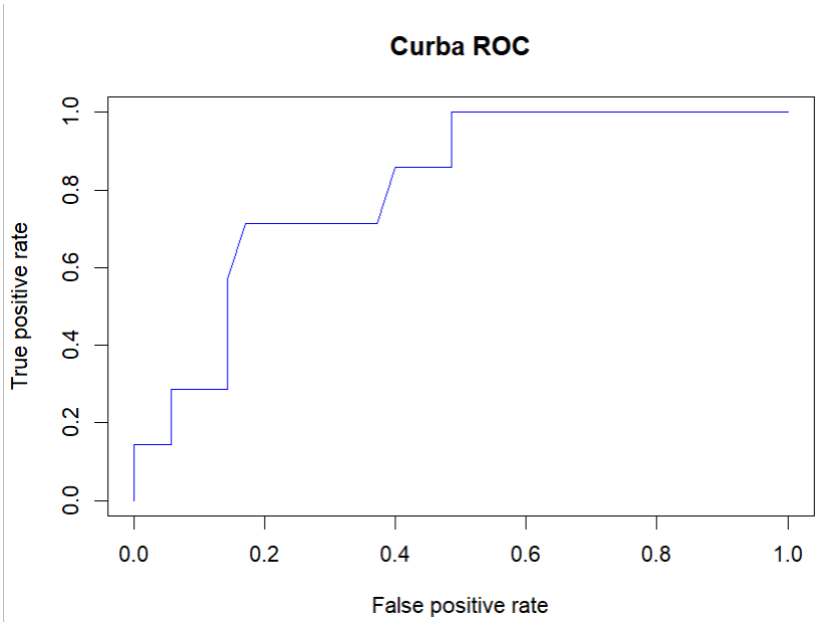
Modelul clasifică corect aproximativ 79.4% dintre elevii din setul de antrenare.

- **Acuratețea și matricea de confuzie pe setul de testare**

```
> pred1 <- rep("0", dim(settestare)[1])
> pred1[prob > 0.5] <- "1"
> table(pred1, settestare$high_risk_f)

pred1  0  1
    0 35  6
    1  0  1
> acc_test <- mean(pred1 == settestare$high_risk_f)
> print(paste("Acuratețea pe setul de testare:", round(acc_test * 100, 2), "%"))
[1] "Acuratețea pe setul de testare: 85.71 %"
```

- 35 elevi „fără risc” au fost corect clasificați.
- 6 elevi „cu risc” au fost greșit clasificați ca „fără risc”
- 1 elev „cu risc” a fost corect identificat

Modelul are o acuratețe foarte bună pe test: 85.71%.



Curba este departe de diagonala → modelul are putere predictivă reală. Cu cât curba este mai aproape de colțul stânga-sus, cu atât este mai performant. Forma este aproape concavă, crescând rapid la început , ceea ce semnalează un echilibru bun între sensibilitate și specificitate pentru unele praguri de decizie.

```
> auc <- performance(pr, measure = "auc")
> auc_value <- auc@y.values[[1]]
> print(paste("AUC =", round(auc_value, 4)))
[1] "AUC = 0.8041"
```

AUC > 0.8 => modelul este bun.
Modelul logistic reușește să discrimineze eficient între elevii cu risc și cei fără, conform curbei ROC.

➢ Regresia logistică multinomiala

Am aplicat un model de regresie logistică multinomială (multinom din pachetul nnet) pentru a prezice ocupația mamei (Mjob) în funcție de:

- absences (absențele elevului)
- studytime (timpul de studiu)
- failures (eșecuri anterioare)
- age (vârsta elevului)

```
> summary(logit_multi)
Call:
multinom(formula = Mjob ~ absences + studytime + failures + age,
    data = data)

Coefficients:
         (Intercept)   absences  studytime    failures        age
health     17.356250 0.03105777 -0.4806423  -1.0761460 -1.0344637
other       7.735102 0.04141361  0.1218342  -0.4614424 -0.4222784
services   14.013963 0.02309438 -0.0451414  -0.0123788 -0.8173442
teacher    16.770533 0.04730760 -0.1887860 -11.0422016 -1.0150319

Std. Errors:
         (Intercept)   absences studytime     failures       age
health      9.703185 0.07091372 0.4863510 0.9158441937 0.6200858
other       6.610242 0.05784604 0.3859407 0.4306399428 0.4117725
services    6.923238 0.05985970 0.3955689 0.4095315176 0.4332571
teacher     9.014996 0.06425015 0.4395942 0.0009556589 0.5729170

Residual Deviance: 385.9903
AIC: 425.9903
```

Coeficienții modelului:

Există coeficienți pentru fiecare clasă (health, other, services, teacher), raportați la categoria de referință (at_home).

- Pentru health:
    - studytime = -0.4806 → Elevii care studiază mai mult au șanse mai mici ca mama lor să lucreze în domeniul sănătății (vs. a fi casnică).
    - failures = -1.0716 → Elevii cu mai multe eșecuri au șanse reduse de a avea mame în domeniul sănătății.
- Pentru services:

- failures = -0.2137, studytime = -0.0451 – efecte slabe, dar tot în direcția negativă (elevii slabi tind să provină mai rar din familii cu mamă care să lucreze).

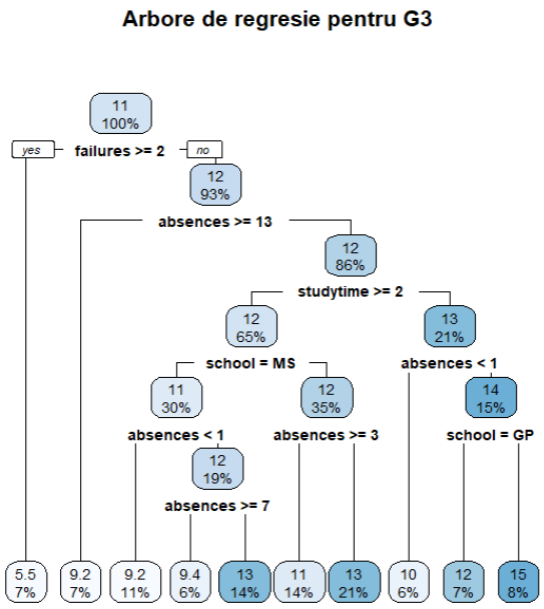
- Pentru teacher:
  - failures = -0.8786, age = -1.015 – elevii cu mame profesoare par să aibă mai puține eșecuri și să fie mai mici.

Mamele care lucrează (în orice domeniu) sunt asociate cu copii cu mai puține eșecuri, mai puține absențe și mai mult timp de studiu, în comparație cu mamele care stau acasă. Deci, ocupația mamei se corelează indirect cu implicarea școlară a elevului.

Concluzionând, elevii cu părinți activi profesional (mai ales în domenii ca sănătate, educație) tind să aibă performanțe și implicare școlară mai bună.

```
> # Matrice de confuzie
> table(Predicted = pred_mjob, Actual = data$Mjob)
          Actual
Predicted  at_home health other services teacher
  at_home        1      0     0        1       0
  health         0      0     0        0       0
  other          7     10    39       24      14
  services       5      5    11       16       6
  teacher        0      0     0        0       0
```

➢ Arbori de regresie si de clasificare. Curatarea arborelui



Rădăcina arborelui :

- Se începe cu întreaga populație (100% din date).
- Primul criteriu de divizare este failures >= 2.

Căile posibile:

- Dacă failures >= 2 ⇒ G3 mediu este 5.5, cu 7% din observații.
- Dacă failures < 2, se merge mai departe și se evaluează absences.

Noduri interioare:

- Fiecare nod reprezintă o decizie de tip „dacă… atunci…”, bazată pe o variabilă (ex: absences >= 13, studytime >= 2).
- Procentul de sub nod (ex. 14%) indică proporția observațiilor care ajung în acel nod.
- Valoarea din cerc (ex. 13) este media valorii G3 pentru acel nod (adică predicția pentru cazurile care ajung acolo).

Frunzele (noduri terminale):

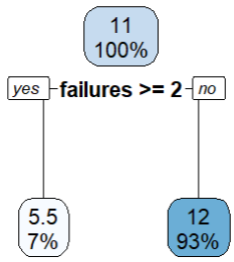
- Sunt predicțiile finale: de exemplu, dacă un elev are failures < 2, absences < 1, și school = GP, atunci se prezice nota 15, cu 8% din cazuri în acea frunză.

Exemplu de traseu în arbore:

Dacă un elev are: failures = 0, absences = 0, merge la școala GP => Va fi clasificat în frunza cu valoarea 15, ceea ce înseamnă că predicția notei finale G3 este 15.

- Mai multe absențe și eșecuri duc la note mai mici: elevii cu failures >= 2 și absences >= 13 au G3 în jur de 5–9.
- Mai puține absențe, școală GP și timp de studiu mai mare sunt asociate cu note mai mari (13–15).

Arbore de regresie curățat pentru G3



Comparativ cu arborele inițial, acesta este mult simplificat, având doar o singură variabilă de decizie.

Rădăcina arborelui :

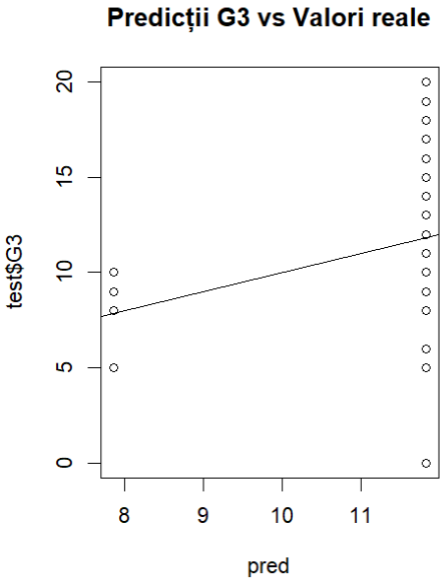
- Modelul ia în considerare doar o singură caracteristică importantă: failures (numărul de eșecuri anterioare).
- Dacă failures >= 2, atunci predicția pentru G3 este 5.5 (probabil un elev cu dificultăți).
- Dacă failures < 2, predicția este 12, deci o performanță școlară decentă.

Distribuția:

- Doar 7% dintre elevi au failures >= 2, iar 93% au mai puține eșecuri.

Doar o singură variabilă (failures) este suficientă pentru a face o predicție rezonabilă despre G3. Celelalte variabile (absences, studytime etc.) nu au mai fost păstrate în modelul curățat deoarece nu au redus semnificativ eroarea de predicție.

Modelul este mai simplu, mai interpretabil și evită supraajustarea (overfitting).



Pe axa X avem predicțiile modelului (pred) — observăm că majoritatea valorilor sunt concentrate între 8 și 11.

Pe axa Y avem valorile reale din setul de testare (test$G3) — care variază mult mai larg (de la 0 la peste 20).

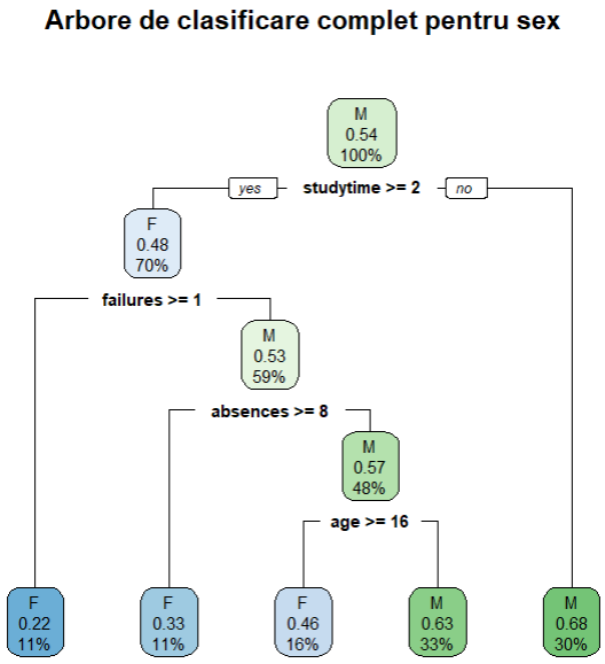
Punctele sunt dispersate vertical, ceea ce arată că modelul oferă predicții similare pentru mulți elevi, indiferent de valoarea reală.

Linia de regresie este aproape orizontală ⇒ slabă corelație între valorile prezise și cele reale, ceea ce sugerează că modelul nu captează bine variația din date ⇒ modelul este subajustat.

```
> # Calculăm eroarea medie pătratică
> mse <- mean((pred - test$G3)^2)
> print(paste("Eroarea de previziune (MSE):", round(mse, 3)))
[1] "Eroarea de previziune (MSE): 14.399"
```

MSE măsoară diferența medie pătratică între predicțiile modelului și valorile reale. Un MSE de 14.399 este destul de mare având în vedere că notele G3 sunt într-un interval de obicei între 0 și 20. Aceasta confirmă vizual că predicțiile sunt departe de valorile reale în multe cazuri.

Arbore de clasificare



Nodul rădăcină :

- Modelul pornește de la întreaga populație: 100% din date.
- Proporția bărbaților (M) este 54%, deci modelul prezice inițial că un elev tipic este mai probabil bărbat.

Dacă studytime >= 2:

- Trecem în partea stângă a arborelui.
- Aici, sunt mai multe fete (F: 70%).

Dacă și failures >= 1:

- Modelul revine la prezicerea bărbaților (M: 59%).

Dacă absences >= 8:

- Predicție: M (57%) – dar nu e o clasificare foarte sigură (aproape 50–50).

Dacă absences < 8 și age >= 16:

- Predicție: M (63%).

Dacă absences < 8 și age < 16:

- Predicție: F (46%) ⇒ aproape echilibrat.

Dacă failures < 1:

- Predicție: F (22%) – foarte feminin acest grup.

Dacă studytime < 2:

- Trecem în partea dreaptă a arborelui.
- Predicție clară: M (68%) – deci puțin timp de studiu e asociat mai frecvent cu băieții în datele tale.

Fetele par să aibă în general: mai mult timp de studiu (studytime >= 2), mai puține eșecuri (failures < 1) și sunt mai frecvent în grupurile cu note mai bune/absențe mai mici. Băieții sunt asociați cu: studytime < 2, mai multe eșecuri și absențe.

```
> pred_class <- predict(tree_class, test, type = "class")
> mean(pred_class == test$sex)  # acuratețea
[1] 0.6964286
```

```
> conf_matrix <- table(Predicted = pred_class, Actual = test$sex)
> conf_matrix
         Actual
Predicted  F  M
        F 24  8
        M  9 15
```

Acuratețea modleului este 69% , ceea ce sugerează că modelul este bun.

Dintre cele 33 de fete 24 au fost clasificate corect, iar 9 au fost clasificate greșit. Dintre cei 23 de baieți, 15 au fost clasificați corect , iar restul greșit.

Arbore de clasificare curățat pentru sex

```
> # Predicții cu arborele curățat
> pred_pruned <- predict(tree_class_pruned, test, type = "class")
> mean(pred_pruned == test$sex)  # acuratețea arborelui curățat
[1] 0.4107143
>
> # Matrice de confuzie pentru arborele curățat
> conf_matrix_pruned <- table(Predicted = pred_pruned, Actual = test$sex)
> conf_matrix_pruned
         Actual
Predicted  F  M
        F  0  0
        M 33 23
```

Acuratețea a scăzut drastic la 41.1%, față de 69.6% în arborele complet.

Matricea de confuzie arată că 23 de elevi de gen masculin au fost clasificați corect, iar 33 de elevi de gen masculin au fost clasificați greșit.

6. KNN de clasificare

```
> knn_pred <- knn(train = train_norm, test = test_norm, cl = train_knn$sex_knn, k = 5)
> confusionMatrix(knn_pred, test_knn$sex_knn)
Confusion Matrix and Statistics

          Reference
Prediction  F  M
         F 14  5
         M 10 13

               Accuracy : 0.6429
                 95% CI : (0.4803, 0.7845)
    No Information Rate : 0.5714
    P-Value [Acc > NIR] : 0.2189

                  Kappa : 0.2953

 Mcnemar's Test P-Value : 0.3017

            Sensitivity : 0.5833
            Specificity : 0.7222
         Pos Pred Value : 0.7368
         Neg Pred Value : 0.5652
             Prevalence : 0.5714
         Detection Rate : 0.3333
   Detection Prevalence : 0.4524
      Balanced Accuracy : 0.6528

       'Positive' Class : F
```

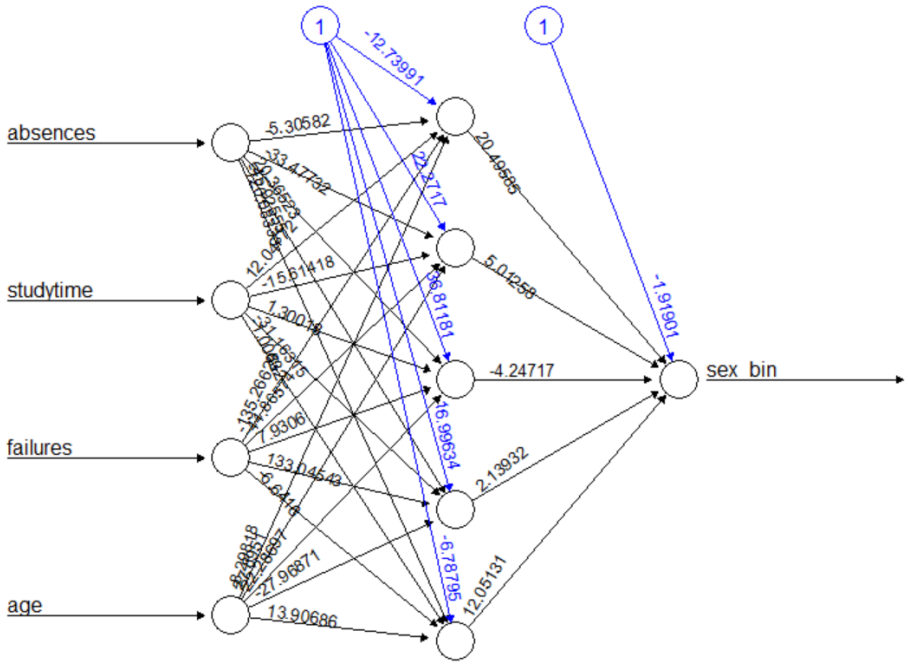
Modelul a prezis:

- 14 femei corect (True Positives)
- 13 bărbați corect (True Negatives)
- 10 femei greșit clasificate ca bărbați (False Negatives)

    - 5 bărbați greșit clasificați ca femei (False Positives)

- Accuracy: 0.6429 → Modelul a prezis corect ~64.3% din cazuri.
- Kappa: 0.2953 → Acord slab spre moderat între predicții și realitate.
- Balanced Accuracy: 0.6528 → Medie între:
    - Sensitivity (Recall pentru F): 0.5833 → Modelul identifică ~58% din femei.
    - Specificity (Recall pentru M): 0.7222 → Modelul identifică ~72% din bărbați.
- Positive class este F (setată implicit de caret, pe primul nivel al factorului).
- Valori PPV și NPV:
    - PPV pentru F (Precizia pentru F): 0.7368 → Din toate predicțiile F, ~73.7% au fost corecte.
    - NPV: 0.5652 → Din toate predicțiile M, ~56.5% au fost corecte.
- Modelul KNN cu k = 5 oferă performanță acceptabilă, dar nu remarcabilă.
- Prezice ceva mai bine bărbații (Specificity > Sensitivity).

7. Retele neuronale pentru clasificare



Această rețea neuronală încearcă să învețe o relație între caracteristicile personale ale elevilor (absences, studytime, failures, age) și sexul lor (sex_bin). Fiecare conexiune și greutate arată cum contribuie o variabilă sau un neuron la decizia finală.

Rețeaua neuronală folosește următoarele variabile de intrare introduse în rețea pentru a ajuta la prezicerea variabilei de ieșire:

- absences (absențe)
- studytime (timp de studiu)
- failures (număr de corigențe/eșecuri)
- age (vârstă)

Rețeaua are un strat ascuns format din 5 neuroni. Fiecare variabilă de intrare este conectată la fiecare neuron din stratul ascuns printr-o linie care are asociată o greutate numerică:

- Greutățile indică cât de puternic influențează fiecare variabilă activarea neuronului.
- Valori pozitive contribuie pozitiv la activare, iar cele negative o reduc.

Ieșirea rețelei este sex_bin – o variabilă binară. Neuronii din stratul ascuns sunt conectați la ieșirea finală, iar aceste conexiuni au și ele greutăți (indicând influența fiecărui neuron asupra rezultatului final).

Sunt prezente și neuroni de tip bias (indicați cu cercuri marcate „1”), care ajută la ajustarea activării neuronilor. Aceștia sunt conectați cu săgeți albastre, iar valorile lor pot fi negative sau pozitive.

Eroare: 8.399209 – indică cât de departe sunt predicțiile rețelei față de valorile reale. Cu cât este mai mică, cu atât rețeaua a învățat mai bine.

Pași: 1361 – numărul de pași/iterații necesari pentru antrenarea rețelei.

Valoarea erorii este relativ mare, ceea ce poate însemna că modelul nu este suficient de complex și are nevoie de mai multe date.

```
> confusionMatrix(as.factor(nn_pred_class), as.factor(test_nn$sex_bin))
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 12  8
         1 10 12

               Accuracy : 0.5714
                 95% CI : (0.4096, 0.7228)
    No Information Rate : 0.5238
    P-Value [Acc > NIR] : 0.3225

                  Kappa : 0.1448

 Mcnemar's Test P-Value : 0.8137

            Sensitivity : 0.5455
            Specificity : 0.6000
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.5455
             Prevalence : 0.5238
         Detection Rate : 0.2857
   Detection Prevalence : 0.4762
      Balanced Accuracy : 0.5727

       'Positive' Class : 0
```

Matricea de confuzie prezice:

- 12 elevi au fost corect clasificați (TN)
- 12 elevi au fost corect clasificate (TP)
- 10 elevi au fost clasificați greșit ca fete (FN)
- 8 elevi au fost clasificate greșit ca băieți (FP)

Rețeaua neuronală face destule erori în ambele direcții, ceea ce sugerează că nu există tipare clare sau consistente între sexul elevilor și comportamentele lor școlare.

- Accuratețe: 57.1% → Puțin mai bună decât o presupunere întâmplătoare.
- Kappa: 0.14 → Concordanță slabă între predicție și realitate.
- Sensibilitate (pentru băieți - 0): 54.5%
- Specificitate (pentru fete - 1): 60.0%

Aceste valori indică faptul că modelul nu reușește să identifice în mod constant și corect diferențele de sex în baza comportamentelor școlare.

Modelul de rețea neuronală, aplicat asupra datelor elevilor din două școli, arată că sexul elevului nu poate fi prezis cu acuratețe pe baza comportamentelor și performanțelor școlare. Acest lucru sugerează că, în aceste școli, nu există diferențe educaționale semnificative între băieți și fete din punctul de vedere al indicatorilor analizați.

CONCLUZII

Analiza realizată în cadrul acestui proiect a avut ca scop investigarea factorilor care influențează performanța școlară a elevilor din două unități de învățământ din Portugalia, utilizând metode moderne de inteligență computațională. Studiul a fost orientat atât spre identificarea profilurilor educaționale ale elevilor, cât și spre dezvoltarea unor modele predictive pentru estimarea riscului școlar și a notelor finale.

Din analiza descriptivă reiese că majoritatea elevilor au o frecvență bună și timp de studiu moderat (2–5 ore pe săptămână). Totuși, există variații semnificative, unele cazuri prezentând absențe excesive și performanțe slabe. Media notei finale G3 se situează în jurul pragului de promovare, cu o ușoară concentrare spre notele mari, semnalând o tendință generală pozitivă în rândul elevilor.

Prin aplicarea clusterizării fuzzy, au fost identificate trei tipologii clare de elevi: cei stabili dar pasivi, elevii echilibrați și cei vulnerabili – cu multe absențe și rezultate slabe. Această segmentare oferă o bază utilă pentru politici educaționale diferențiate.

Modelul de regresie logistică binomială a demonstrat că timpul de studiu este un factor protector important împotriva riscului educațional. De asemenea, elevii de la școala MS au o

probabilitate de 3,6 ori mai mare de a fi în risc comparativ cu cei de la GP. Educația părinților și sexul nu au avut un impact semnificativ statistic în acest model, ceea ce sugerează o relativă echitate în șansele educaționale între grupuri.

În ceea ce privește predicția ocupației mamei, regresia logistică multinomială a indicat că elevii cu performanțe mai bune și mai puține eșecuri tind să aibă mame active profesional, în special în domeniul sănătății sau educației. Aceasta reflectă o legătură indirectă între mediul socio-profesional al familiei și implicarea școlară a elevului.

Arborii de decizie au evidențiat în mod constant că variabila „failures" (numărul de eșecuri anterioare) este cel mai puternic predictor al performanței școlare. Chiar și arborii simplificați au obținut rezultate satisfăcătoare folosind doar această variabilă, subliniind impactul major al eșecurilor în traiectoria academică.

Modelele de clasificare aplicate pentru a prezice sexul elevului (arbori, KNN, rețele neuronale) nu au reușit să obțină acuratețe ridicată, sugerând că nu există diferențe comportamentale sau educaționale semnificative între băieți și fete în acest eșantion.

În concluzie, proiectul evidențiază faptul că principalii factori predictivi ai performanței academice sunt de natură comportamentală și educațională (absențe, eșecuri, timp de studiu), în timp ce variabilele demografice (sexul, ocupația părinților) au un rol secundar. Modelele predictive implementate pot constitui un punct de plecare pentru intervenții țintite în scopul reducerii riscului de eșec școlar și creșterii performanței educaționale în rândul elevilor.

COD R:

```
data <- ice

str(data)

summary(data)

cor(data)

cor(data[sapply(data, is.numeric)])

# Conversia variabilelor categoriale

data$school <- as.factor(data$school)

data$sex <- as.factor(data$sex)
```

```
data$Mjob <- as.factor(data$Mjob)
data$Pstatus <- as.factor(data$Pstatus)

# Statistici descriptive pentru variabilele numerice
describe(select(data, age, Medu, Fedu, traveltime, studytime, failures, absences, G3))

# Histogramă pentru nota finală (G3)
ggplot(data, aes(x = G3)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black") +
  labs(title = "Distribuția notei finale G3", x = "Nota finală", y = "Frecvență") +
  theme_minimal()

# Selectarea variabilelor numerice pentru clusterizare
data_num <- select(data, age, absences, studytime, traveltime, failures)

# Normalizarea datelor
data_scaled <- scale(data_num)

# Aplicarea algoritmului Fuzzy C-Means (k = 3 clustere)
set.seed(123)
fcm_result <- cmeans(data_scaled, centers = 3,139, m = 2,method="cmeans")
fcm_result

# Vizualizarea rezultatelor

head(fcm_result$membership)
```

```
# Afișarea etichetelor de cluster
table(fcm_result$cluster)

# Adăugarea clusterului în setul de date
data$cluster_fuzzy <- as.factor(fcm_result$cluster)

# Vizualizare clustere în funcție de 2 variabile
ggplot(data, aes(x = absences, y = studytime, color = cluster_fuzzy)) +
  geom_point(size = 2) +
  labs(title = "Clusterizare fuzzy: Absențe vs. Timp de studiu", x = "Absențe", y = "Timp de studiu")
+
  theme_minimal()

#Se ordoneaza crescator  rangul genului pe clustere:
o<-order(fcm_result$cluster)
o
data.frame(data$sex[o],fcm_result$cluster[o])



# Regresie logistică binomială

library(ggplot2)
library(caTools)
library(ROCR)

data$high_risk <- ifelse(data$failures >= 2 | data$absences > 9, 1, 0)
data$high_risk_f <- factor(data$high_risk)
```

```
model_risk <- glm(high_risk_f ~ studytime + Medu + Fedu + sex + school, data = data, family = binomial)
summary(model_risk)



set.seed(123)
ind <- sample(1:nrow(data), size = 0.7 * nrow(data))  # 70% antrenare
setantrenare <- data[ind, ]
settestare <- data[-ind, ]

# Interpretare coeficienți
exp(coef(model_risk))  # rapoartele de șanse

# Predictii probabilități
prob <- predict(model_risk, settestare, type = "response")
prob

# Clasificare pe setul de antrenare
pred <- rep("0", dim(setantrenare)[1])
pred[predict(model_risk, setantrenare, type = "response") > 0.5] <- "1"
table(pred, setantrenare$high_risk_f)

# Acuratețe pe setul de antrenare
acc_train <- mean(pred == setantrenare$high_risk_f)
print(paste("Acuratețea pe setul de antrenare:", round(acc_train * 100, 2), "%"))
```

```
# Clasificare pe setul de testare
pred1 <- rep("0", dim(settestare)[1])
pred1[prob > 0.5] <- "1"
table(pred1, settestare$high_risk_f)

# Acurateţe pe setul de testare
acc_test <- mean(pred1 == settestare$high_risk_f)
print(paste("Acurateţea pe setul de testare:", round(acc_test * 100, 2), "%"))

#Curba ROC
install.packages("ROCR")
library(ROCR)
p <- predict(model_risk, newdata = settestare, type = "response")
pr <- prediction(p, settestare$high_risk_f)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf, col = "blue", main = "Curba ROC")

# AUC
auc <- performance(pr, measure = "auc")
auc_value <- auc@y.values[[1]]
print(paste("AUC =", round(auc_value, 4)))

# 3.2 Regresie logistică multinomială

install.packages("nnet")
library(nnet)
```

```
# Eliminăm NA din Mjob dacă există
data <- data[!is.na(data$Mjob), ]

# Model multinomial
logit_multi <- multinom(Mjob ~ absences + studytime + failures + age, data = data)

# Rezumat
summary(logit_multi)

# Prezicerea etichetelor
pred_mjob <- predict(logit_multi, newdata = data)

# Matrice de confuzie
table(Predicted = pred_mjob, Actual = data$Mjob)

library(rpart)
library(rpart.plot)

# Arbore de regresie pentru variabila G3 (nota finală)

# Se construiește un arbore de regresie folosind câteva variabile relevante
tree_reg <- rpart(G3 ~ absences + studytime + failures + age + school,
                  data = data,
                  method = "anova")

# Se afișează arborele de regresie
rpart.plot(tree_reg, main = "Arbore de regresie pentru G3")
```

```
# Alegem complexitatea optimă (cp) care minimizează eroarea estimată (xerror)
cp_opt <- tree_reg$cptable[which.min(tree_reg$cptable[,"xerror"]), "CP"]

# Se construiește arborele curățat (pruned) pe baza cp-ului optim
tree_reg_pruned <- prune(tree_reg, cp = cp_opt)

# Se afișează arborele de regresie curățat
rpart.plot(tree_reg_pruned, main = "Arbore de regresie curățat pentru G3")

# Predicția valorilor G3 pe un set de test (vom crea o împărțire în train/test)
set.seed(123)
sample_index <- sample(1:nrow(data), nrow(data)/2)
train <- data[sample_index, ]
test <- data[-sample_index, ]

# Reconstruim arborele pe setul de antrenare
tree_train <- rpart(G3 ~ absences + studytime + failures + age + school,
                    data = train,
                    method = "anova")

# Pruning pe arborele antrenat
cp_train <- tree_train$cptable[which.min(tree_train$cptable[,"xerror"]), "CP"]
tree_pruned <- prune(tree_train, cp = cp_train)

# Predicția pe setul de test
pred <- predict(tree_pruned, newdata = test)
```

```
# Comparația dintre valorile prezise și cele reale
plot(pred, test$G3, main = "Predicții G3 vs Valori reale")
abline(0, 1)

# Calculăm eroarea medie pătratică
mse <- mean((pred - test$G3)^2)
print(paste("Eroarea de previziune (MSE):", round(mse, 3)))

# Arbori de clasificare pentru sex + pruning

data$sex <- as.factor(data$sex)
set.seed(123)

# Împărțim datele în set de antrenament (60%) și testare (40%)
split <- sample(1:nrow(data), nrow(data) * 0.6)
train <- data[split, ]
test <- data[-split, ]

# Construim arborele de clasificare (rpart)
tree_class <- rpart(sex ~ absences + studytime + failures + age + school,
                    data = train, method = "class", cp = 0.01)

# Afișăm arborele complet
rpart.plot(tree_class, main = "Arbore de clasificare complet pentru sex")

# Rată de acuratețe pe testare (arbore complet)
```

```
pred_class <- predict(tree_class, test, type = "class")
mean(pred_class == test$sex)  # acuratețea

# Matrice de confuzie pentru arborele complet
conf_matrix <- table(Predicted = pred_class, Actual = test$sex)
conf_matrix

# Alegem complexitatea optimă (cp) corespunzătoare celei mai mici erori de cross-validare
cp_opt_class <- tree_class$cptable[which.min(tree_class$cptable[, "xerror"]), "CP"]

# Curățăm arborele (pruning)
tree_class_pruned <- prune(tree_class, cp = cp_opt_class)

# Afișăm arborele curățat
rpart.plot(tree_class_pruned, main = "Arbore de clasificare curățat pentru sex")

# Predicții cu arborele curățat
pred_pruned <- predict(tree_class_pruned, test, type = "class")
mean(pred_pruned == test$sex)  # acuratețea arborelui curățat

# Matrice de confuzie pentru arborele curățat
conf_matrix_pruned <- table(Predicted = pred_pruned, Actual = test$sex)
conf_matrix_pruned

# 5. KNN de clasificare (sex)
install.packages("dplyr")
library(dplyr)
```

```
# Pregătirea datelor
data_knn <- data %>%
  mutate(sex_knn = as.factor(sex)) %>%
  select(sex_knn, absences, studytime, failures, age)

# Împărțire train/test
set.seed(123)
train_idx <- sample(1:nrow(data_knn), 0.7 * nrow(data_knn))
train_knn <- data_knn[train_idx, ]
test_knn <- data_knn[-train_idx, ]

# Normalizare
normalize <- function(x) {(x - min(x)) / (max(x) - min(x))}
train_norm <- as.data.frame(lapply(train_knn[,-1], normalize))
test_norm <- as.data.frame(lapply(test_knn[,-1], normalize))

# KNN
install.packages("class")
library(class)
install.packages("caret")
library(caret)
library(ggplot2)

knn_pred <- knn(train = train_norm, test = test_norm, cl = train_knn$sex_knn, k = 5)
confusionMatrix(knn_pred, test_knn$sex_knn)
```

```
# 6. Rețele neuronale pentru clasificare (sex)
data_nn <- data %>%
  mutate(sex_bin = ifelse(sex == "M", 1, 0)) %>%
  select(sex_bin, absences, studytime, failures, age)

# Normalizare
data_nn_norm <- as.data.frame(lapply(data_nn, normalize))

# Împărțire train/test
train_idx_nn <- sample(1:nrow(data_nn_norm), 0.7 * nrow(data_nn_norm))
train_nn <- data_nn_norm[train_idx_nn, ]
test_nn <- data_nn_norm[-train_idx_nn, ]

# Rețea
install.packages("neuralnet")
library(neuralnet)

set.seed(123)
nn_model <- neuralnet(sex_bin ~ absences + studytime + failures + age, data = train_nn, hidden =
c(5), linear.output = FALSE)

# Plot
plot(nn_model)

# Predictii
nn_pred <- compute(nn_model, test_nn[,-1])$net.result
nn_pred_class <- ifelse(nn_pred > 0.5, 1, 0)
```

```
# Evaluare
confusionMatrix(as.factor(nn_pred_class), as.factor(test_nn$sex_bin))
```