

Universitatea Babeș-Bolyai
Facultatea de Științe Economice și Gestiunea Afacerilor

PROIECT

**Analiza factorilor socio-economici determinanți
ai averii miliardarilor**

Studente: Olteanu Carla-Antonia, Sîrbu Iulia
Informatică Economică
An 3, grupa 5, 6

Introducere

Un subiect de interes global, atât pentru economiști, cât și pentru sociologi, politologi și alți oameni de rând, îl reprezintă studiul averii miliardarilor și factorii influenți care contribuie la aceasta. Există numeroase întrebări cu privire la factorii determinanți care stau la baza obținerii averii.

Acest studiu explorează diversele atribute și contexte ale miliardarilor din întreaga lume prin utilizarea unui set de date cuprinzător care încapsulează diferite aspecte ale vieții și intereselor lor de afaceri. Setul de date include informații relevante precum câștigul net, genul și vârsta individului, categoria industriei în care aceștia activează, nivelul de educație al fiecăruia, precum și indici economici care contribuie la bunăstarea acestora.

Pentru a înțelege cât mai bine analiza acestui studiu, dorim să răspundem unor întrebări de cercetare specifice care să ne clarifice anumite aspecte de interes.

1. Există o legătură semnificativă între averea miliardarilor și genul, vârsta, educația universitară(de nivel superior) și industria în care activează?
2. În cazul existenței acestei legături, putem afirma că este o legătură puternică față de toate variabilele sau există diferențe de importanță între acestea?

Rezultatele acestui studiu de cercetare ar putea avea un impact major nu doar pentru indivizii interesați de profilul economic și social al miliardarilor, dar și pentru anumite strategii de investiții și politici economice la nivel macro.

Spre exemplu, dacă în urma analizei se constată că educația universitară, pe lângă cea primară (de bază în viața oricărui individ), are o influență semnificativă asupra atingerii unui statut mai înalt și acumularea averii, acest fapt ar putea încuraja investițiile în educația superioară și reformele în sistemele educaționale pentru a stimula și a susține dezvoltarea economică.

Simultan, examinarea impactului pe care diverse industrii îl au asupra acumulării de bogăție ar putea oferi perspective vitale pentru avansarea politicilor economice și strategice. Aceste date au potențialul de a direcționa factorii de decizie politică către promovarea și sprijinirea sectoarelor economice care au un potențial semnificativ pentru crearea de locuri de muncă. Prin urmare, guvernele și agențiile economice ar putea să aloce resurse și să prioritizeze investițiile în domenii specifice industriei, cum ar fi tehnologia informației, sănătatea sau energia regenerabilă, care au un impact semnificativ asupra creșterii economice și bunăstării generale.

Setul de date

Setul de date care ni s-a părut cel mai relevant pentru cercetarea noastră și pe care am decis să îl utilizăm în cadrul acestui proiect este **Billionaires Statistics Dataset (2023)**, preluat din Kaggle este disponibil la adresa web: <https://www.kaggle.com/datasets/nelgiriyeewithana/billionaires-statistics-dataset>.

Autorul acestui set de date a oferit informații specifice despre anumite caracteristici ale indivizilor, însă nu a furnizat detalii specifice despre miliardarii care dețin aceste date.

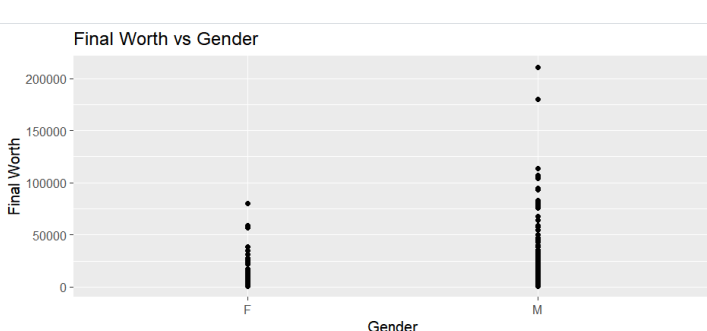
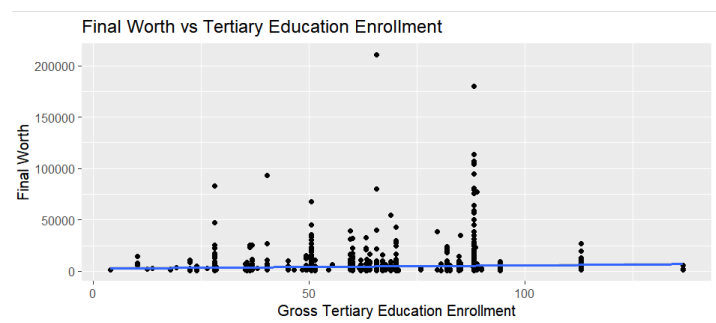
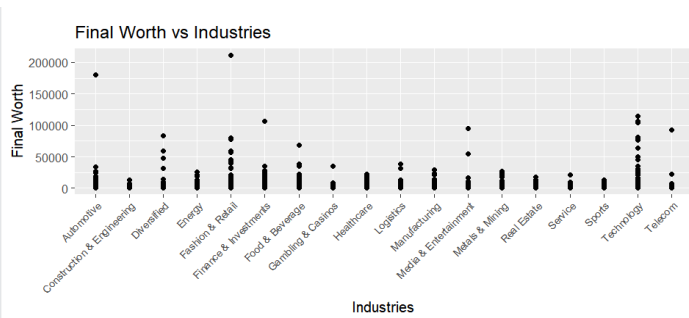
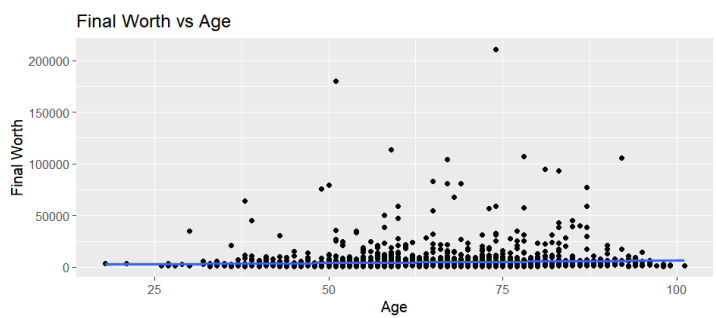
Structura setului de date este formată din 35 de coloane, fiecare reprezentând o variabilă specifică. A fost necesară preprocesarea acestor date pentru a ne asigura că sunt adecvate pentru întrebările noastre de cercetare, astfel am trecut acest set de date prin Tableau Prep Builder pentru a le curăța.

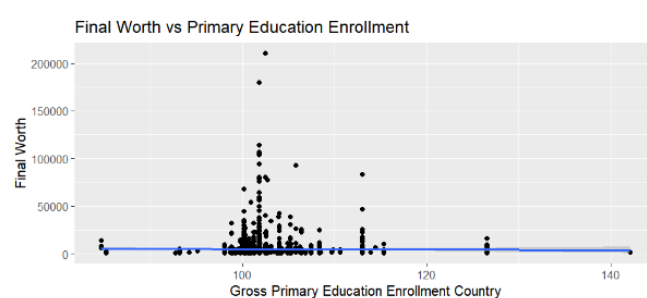
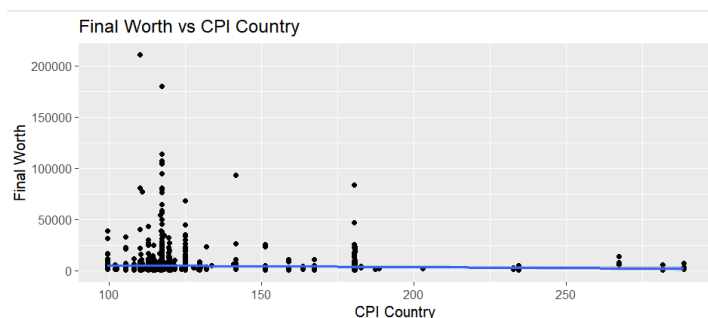
Așadar, am rămas cu 9 coloane relevante pentru studiul nostru: finalWorth (valoarea finală a averii miliardarului), age (vârsta miliardarului), industries (industriile în care activează miliardarul), gender (genul miliardarului), cpi_country (indicele prețurilor de consum al țării de proveniență), cpi_change_country (schimbarea indicelui prețurilor de consum al țării de proveniență), gdp_country (PIB-ul țării de proveniență), gross_tertiary_education_enrollment (rata de înscriere în învățământul terțiar a țării de proveniență), gross_primary_education_enrollment_country (rata de înscriere în învățământul primar a țării de proveniență).

Coloanele gender, industries sunt de tip caracter, deci a trebuit să le convertim în factori. Coloana gdp_country a fost transformată în coloană numerică. De asemenea, variabila independentă finalWorth a fost transformată într-o variabilă binară pentru buna funcționare a regresiei logistice și într-o variabilă categorică pentru clasificare la arbori.

Acest set de date este relevant pentru întrebările de cercetare alese deoarece include variabile esențiale pentru a înțelege factorii care influențează averea miliardarilor. Totodată, cu ajutorul acestui set de date, putem explora legăturile dintre caracteristicile personale și economice ale miliardarilor și averea lor.

Mai departe vom analiza datele pe care le avem la dispoziție, realizând grafice pentru fiecare variabilă importantă.





Așadar, observăm că este o probabilitate să existe o legătură între variabilele age, gender, gross_tertiarity_education_enrollment și industries. Pe de altă parte, variabilele gross_primary_education_enrollment și cpi_country nu par să influențeze în mod semnificativ averea miliardarilor.

Rezultate și discuții

Se observă că datele pe care le conține setul de date sunt atât numerice, cât și categorice. Dorind să stabilim dacă genul, vârsta, educația universitară(de nivel superior) și industria influențează sau nu câștigul net al milionarilor, alegem pentru început să previzionăm cu ajutorul regresiei logistice.

Datele au fost împărțite în seturi de antrenament și test, pentru a evalua performanța modelelor de regresie logistică. Prin această împărțire, am antrenat modelul pe subsetul train și l-am validat pe subsetul test.

Înainte de a trece la interpretarea rezultatelor modelării logistice pentru întrebarea noastră, am considerat necesară o prelucrare a datelor. Astfel, pentru a prezice logistic finalWorth-ul, trebuie să îl transformăm într-o variabilă binară, iar restul variabilelor (age, gender, gross_tertiarity_education_enrollment, industries) trebuie să le transformăm în factori.

```
> data %>% mutate_if(is.character, as.factor)
> data$finalWorth <- as.numeric(data$finalWorth)
> data$finalWorth_binary <- ifelse(data$finalWorth > median(data$finalWorth, na.rm = TRUE), 1, 0)
```

Mai departe am creat 4 modele pentru a evalua impactul vârstei, genului, industriei și educației terțiare asupra probabilității ca un miliardar să aibă o avere peste mediana dataset-ului.

```
> mod_age <- glm(finalworth_binary ~ age, data, family = binomial)
> summary(mod_age)

Call:
glm(formula = finalworth_binary ~ age, family = binomial, data = data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.280976    0.210904  -6.074 1.25e-09 ***
age           0.019080    0.003181   5.999 1.99e-09 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3321.9  on 2396  degrees of freedom
Residual deviance: 3285.2  on 2395  degrees of freedom
AIC: 3289.2

Number of Fisher Scoring iterations: 4
```

Se poate observa că vârsta are un efect pozitiv și semnificativ asupra probabilității ca averea să fie peste mediană deoarece $p = 1.99e-09 < 0.001$, deci este foarte mic.

```
> mod_gender <- glm(finalworth_binary ~ gender, data, family = binomial)
> summary(mod_gender)
```

```
Call:
glm(formula = finalworth_binary ~ gender, family = binomial,
    data = data)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.1061    0.1191    0.891   0.373
genderM       -0.1667    0.1268   -1.315   0.189
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 3321.9 on 2396 degrees of freedom
Residual deviance: 3320.2 on 2395 degrees of freedom
AIC: 3324.2
```

```
Number of Fisher Scoring iterations: 3
```

Pentru impactul genului, $p = 0.189$, indicând că genul nu este un predictor semnificativ pentru averea finală.

```
> mod_industries <- glm(finalworth_binary ~ industries, data, family = binomial)
> summary(mod_industries)
```

```
Call:
glm(formula = finalworth_binary ~ industries, family = binomial,
    data = data)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.3314    0.2477    1.338   0.1810
industriesConstruction & Engineering -0.7776    0.4048   -1.921   0.0547
industriesDiversified    -0.5623    0.2909   -1.933   0.0533
industriesEnergy         -0.2668    0.3231   -0.826   0.4090
industriesFashion & Retail -0.2165    0.2789   -0.776   0.4377
industriesFinance & Investments -0.1900    0.2705   -0.702   0.4826
industriesFood & Beverage  -0.1296    0.2876   -0.451   0.6523
industriesGambling & Casinos -0.3314    0.4931   -0.672   0.5016
industriesHealthcare     -0.6315    0.2883   -2.190   0.0285 *
industriesLogistics      -0.1795    0.4412   -0.407   0.6842
industriesManufacturing  -0.6475    0.2750   -2.354   0.0186 *
industriesMedia & Entertainment -0.2361    0.3303   -0.715   0.4748
industriesMetals & Mining  -0.2170    0.3445   -0.630   0.5289
industriesReal Estate    -0.5715    0.2948   -1.939   0.0525
industriesService        -0.0915    0.3853   -0.237   0.8123
industriesSports         -0.4367    0.4086   -1.069   0.2851
industriesTechnology     -0.5024    0.2741   -1.833   0.0668
industriesTelecom        -0.1883    0.4527   -0.416   0.6775
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 3321.9 on 2396 degrees of freedom
Residual deviance: 3291.9 on 2379 degrees of freedom
AIC: 3327.9
```

Pentru predicatul industries, însă, este o discuție mai elaborată deoarece coeficienții pentru diferitele industrii variază.

Avem unele industrii în care coeficientul este mic, cum ar fi la Construction&Engineering, Diversified, Healthcare sau Manufacturing. La alte industrii, însă, coeficienții sunt nesemnificativi.

```
> mod_gross_tertiary_education <- glm(finalworth_binary ~ gross_tertiary_education_
enrollment, data, family = binomial)
> summary(mod_gross_tertiary_education)
```

```
Call:
glm(formula = finalworth_binary ~ gross_tertiary_education_enrollment,
    family = binomial, data = data)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.710250    0.136788   -5.192 2.08e-07 ***
gross_tertiary_education_enrollment  0.009911    0.001929    5.137 2.80e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 3321.9 on 2396 degrees of freedom
Residual deviance: 3295.2 on 2395 degrees of freedom
AIC: 3299.2
```

```
Number of Fisher Scoring iterations: 4
```

P-value pentru acest predicat este egal cu $2.80e-07$, indicând că există o asociere semnificativă între gross_tertiary_education_enrollment și averea finală.

Acum vom trece la evaluarea modelului complet pentru toți predictorii, rezultatul oferit indicând un model dezechilibrat, care tinde să clasifice majoritatea observațiilor în clasa negativă, 367 de cazuri fiind true negatives, iar 353 de false negatives.

Mai departe, am făcut cross-validation pentru a ne asigura că modelul generalizează bine pe date noi. Acuratețea medie a modelului pe cele 10 folduri este de 0.554. Coeficientul Kappa este 0.106, sugerând un acord ușor, dar nu foarte puternic.

```

> conf_matrix <- confusionMatrix(pred_test_cv, test$finalworth_binary)
> print(conf_matrix)
Confusion Matrix and Statistics

      Reference
Prediction 0      1
0      235 169
1      132 184

      Accuracy : 0.5819
      95% CI   : (0.5449, 0.6183)
      No Information Rate : 0.5097
      P-Value [Acc > NIR] : 5.934e-05

      Kappa : 0.1619

McNemar's Test P-Value : 0.03799

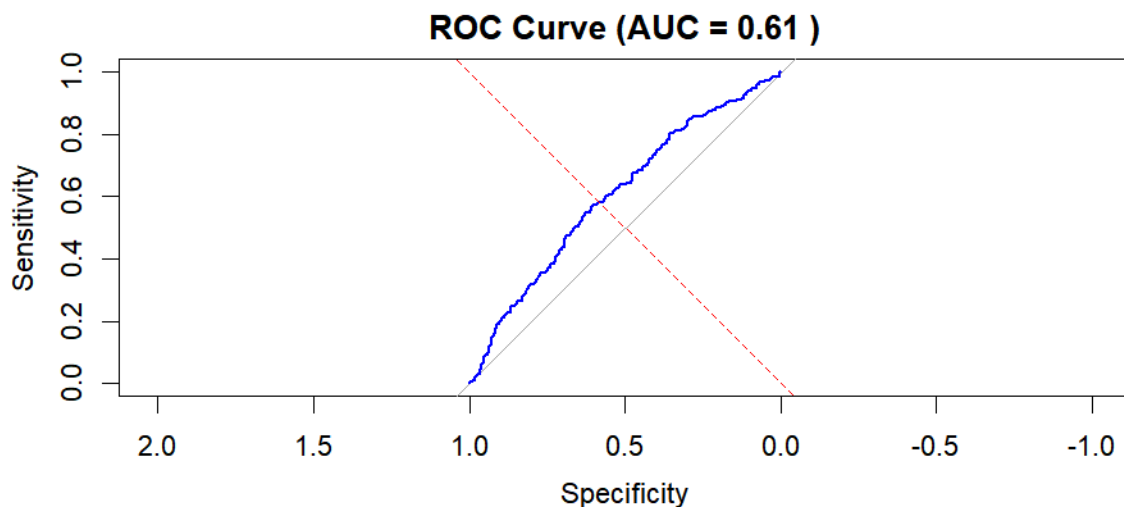
      Sensitivity : 0.6403
      Specificity : 0.5212
      Pos Pred Value : 0.5817
      Neg Pred Value : 0.5823
      Prevalence : 0.5097
      Detection Rate : 0.3264
      Detection Prevalence : 0.5611
      Balanced Accuracy : 0.5808

      'Positive' Class : 0

```

Modelul prezice corect 58.19% din cazuri, detectând 64.03% din cazurile pozitive. Coeficientul Kappa sugerează un grad moderat de acord între predicțiile modelului și valorile reale.

Folosind sensibilitatea și specificitatea pentru diferitele praguri de decizie ale modelului, am creat graficul ROC și aria AUC. În cazul nostru, AUC este egal cu 0.61, având deci o capacitate modestă de a diferenția instanțele pozitive și cele negative. Deoarece $AUC > 0.5$ modelul are o performanță mai bună decât clasificarea aleatorie. Forma curbei ROC arată că există loc de îmbunătățire.



În concluzie, vom formula următoarele răspunsuri la întrebările noastre de cercetare.

1. Există o legătură între vârsta, industria în care activează miliardarii și educația de nivel superior și averea acestora. Genul nu are o relevanță semnificativă.
2. Legătura dintre aceste variabile și averea miliardarilor nu este una puternică, deși vârsta și educația de nivel superior au arătat semnificație statistică în modelul nostru.

Având în vedere că modelele de regresie logistică nu au indicat o legătură puternică între predictorii studiați și averea miliardarilor, vom explora metode alternative. Una dintre aceste metode este utilizarea arborilor de decizie care vor oferi o abordare mai puternică pentru analiza datelor noastre, îmbunătățind acuratețea predicțiilor.

În cele ce urmează, pentru a răspunde mai relevant și precis la întrebările de cercetare, am utilizat 2 metode de modelare, și anume arborii de decizie și random forest. Fiecare dintre aceste metode de modelare are avantaje și limitări specifice.

Arborii de decizie constituie o metoda prin care ni se permite identificarea criteriilor esențiale ce influențează rezultatele. Aceștia sprijină și facilitează înțelegerea rezultatelor de către persoane care nu au cunoștințe avansate în domeniul statisticii spre exemplu.

Random forest, pe de altă parte, este o metoda mult mai complexă și puternică ce constă în construirea și combinarea mai multor arbori de decizie pentru a putea obține o predicție finală. Prin folosirea acestei metode, reducem riscul de overfitting ce ar putea interveni prin utilizarea metodei simple a arborilor de decizie și de asemenea putem ajunge la predicții mai precise și mai solide.

În cadrul scriptului din R Studio în care am realizat modelarea prin intermediul arborilor, am realizat conversia variabilelor necesare într-un format adecvat pentru analiză. În mod specific, am convertit variabila **gdp_country** într-o variabilă numerică prin eliminarea simbolurilor speciale, am transformat variabilele **gender** și **finalWorth** în variabile categorice, iar **finalWorth** am împărțit-o în doua categorii: Low și High, în funcție de mediana valorii averii nete finale.

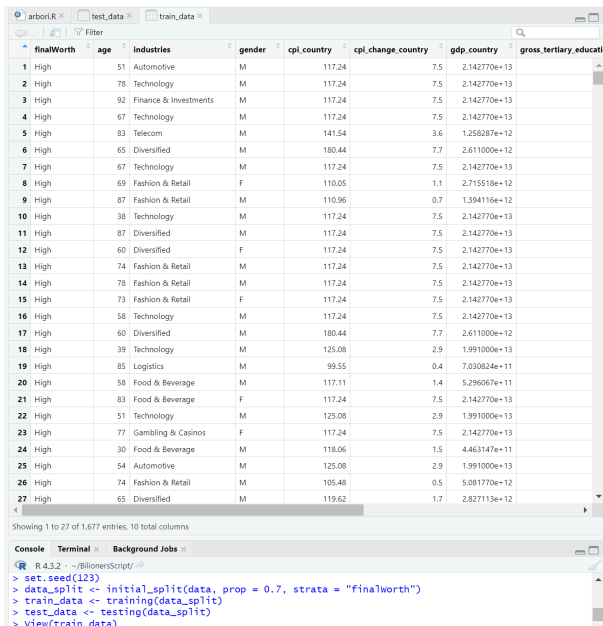
```
> data$gdp_country <- as.numeric(gsub('[$,]', '', data$gdp_country))
> data$gender <- as.factor(data$gender)
> data <- data %>% mutate(finalworth = ifelse(finalworth <= median(finalworth, na.rm = TRUE), "Low", "High"))
```

Observăm că avem 1174 valori pentru High și 1223 pentru Low, ceea ce înseamnă că distribuția e aproape egală între valorile totale, ceea ce va permite modelelor de învățare să preia date dintr-un set care nu este puternic dezechilibrat între clase, ceea ce ar putea afecta acuratețea predicțiilor.

```
> data <- data %>% mutate(finalworth = factor(finalworth))
> table(data$finalworth)
```

```
High Low
1174 1223
```

Am continuat analiza printr-un demers esențial pentru a asigura o evaluare corectă a modelelor utilizate, și anume împărțirea setului de date. Am împărțit setul de date inițial într-un set de antrenament (70%) și un set de test (30%).



	finalWorth	age	industries	gender	cpi_country	cpi_change_country	gdp_country	gross_tertiary_education_enrollment
1	High	51	Automotive	M	117.24	7.5	2.142770e+13	
2	High	78	Technology	M	117.24	7.5	2.142770e+13	
3	High	92	Finance & Investments	M	117.24	7.5	2.142770e+13	
4	High	67	Technology	M	117.24	7.5	2.142770e+13	
5	High	83	Telecom	M	141.54	3.6	1.258287e+12	
6	High	65	Diversified	M	180.44	7.7	2.611000e+12	
7	High	67	Technology	M	117.24	7.5	2.142770e+13	
8	High	69	Fashion & Retail	F	110.05	1.1	2.715518e+12	
9	High	87	Fashion & Retail	M	110.96	0.7	1.394116e+12	
10	High	38	Technology	M	117.24	7.5	2.142770e+13	
11	High	87	Diversified	M	117.24	7.5	2.142770e+13	
12	High	60	Diversified	F	117.24	7.5	2.142770e+13	
13	High	74	Fashion & Retail	M	117.24	7.5	2.142770e+13	
14	High	78	Fashion & Retail	M	117.24	7.5	2.142770e+13	
15	High	73	Fashion & Retail	F	117.24	7.5	2.142770e+13	
16	High	58	Technology	M	117.24	7.5	2.142770e+13	
17	High	60	Diversified	M	180.44	7.7	2.611000e+12	
18	High	39	Technology	M	125.08	2.9	1.991000e+13	
19	High	85	Logistics	M	99.55	0.4	7.030824e+11	
20	High	58	Food & Beverage	M	117.11	1.4	5.296067e+11	
21	High	83	Food & Beverage	F	117.24	7.5	2.142770e+13	
22	High	51	Technology	M	125.08	2.9	1.991000e+13	
23	High	77	Gambling & Casinos	F	117.24	7.5	2.142770e+13	
24	High	30	Food & Beverage	M	118.06	1.5	4.463147e+11	
25	High	54	Automotive	M	125.08	2.9	1.991000e+13	
26	High	74	Fashion & Retail	M	105.48	0.5	5.081770e+12	
27	High	65	Diversified	M	119.62	1.7	2.827113e+12	

```

R 4.3.2 -- R61095Script --
> set.seed(123)
> data_split <- initial_split(data, prop = 0.7, strata = "finalWorth")
> train_data <- training(data_split)
> test_data <- testing(data_split)
> view(train_data)

```

Modelul de decizie al arborilor prezice valoarea averii finale (**finalWorth**) pe baza mai multor variabile predictive, antrenat folosind datele din train_data obtinute la pasul anterior, având ca scop clasificarea observațiilor în funcție de valoarea **finalWorth**-ului.

```

> m1 <- rpart(
+   formula = finalWorth ~ gender + age + gross_tertiary_education_enrollment +
+   gross_primary_education_enrollment_country + gdp_country +
+   cpi_change_country + cpi_country,
+   data = train_data,
+   method = "class"
+ )
> summary(m1)
Call:
rpart(formula = finalWorth ~ gender + age + gross_tertiary_education_enrollment +
  gross_primary_education_enrollment_country + gdp_country +
  cpi_change_country + cpi_country, data = train_data, method = "class")
n= 1677

      CP nsplit rel error   xerror   xstd
1 0.14494519    0 1.0000000 1.0000000 0.02493439
2 0.03288672    1 0.8550548 0.8757613 0.02468525
3 0.01461632    2 0.8221681 0.8672351 0.02465437
4 0.01000000    3 0.8075518 0.8745432 0.02468095

Variable importance
gross_primary_education_enrollment_country      gross_tertiary_education_enrollment
                                38                                18
                                cpi_country                                cpi_change_country
                                18                                15
                                age                                gdp_country
                                9                                2

```

```

> m1
n= 1677

```

```

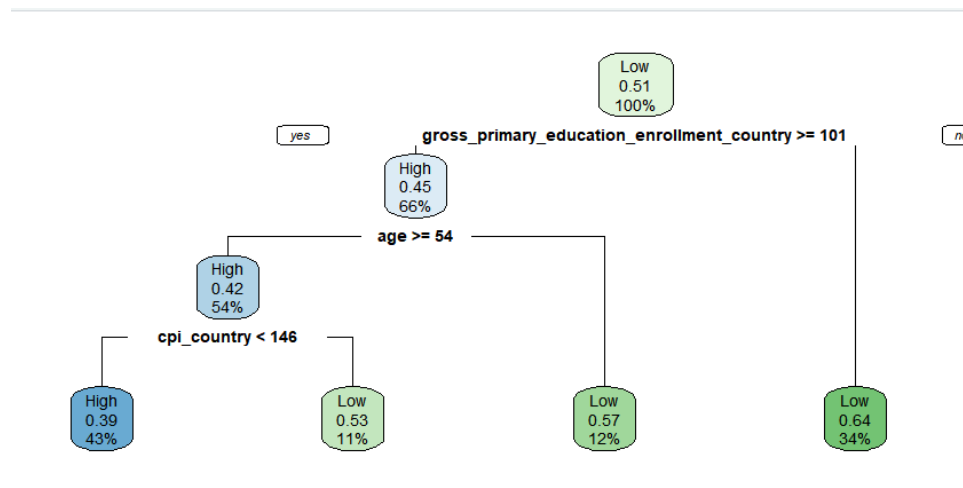
node), split, n, loss, yval, (yprob)
* denotes terminal node

```

- 1) root 1677 821 Low (0.4895647 0.5104353)
- 2) gross_primary_education_enrollment_country>=101.05 1107 494 High (0.5537489 0.4462511)
- 4) age>=53.5 912 383 High (0.5800439 0.4199561)
- 8) cpi_country< 146.36 720 281 High (0.6097222 0.3902778) *
- 9) cpi_country>=146.36 192 90 Low (0.4687500 0.5312500) *
- 5) age< 53.5 195 84 Low (0.4307692 0.5692308) *
- 3) gross_primary_education_enrollment_country< 101.05 570 208 Low (0.3649123 0.6350877) *

CP-ul, parametrul de complexitate controlează împărțirea arborelui. Valorile mai mici ale acestuia indică un model mai complet, iar pe masura ce crește numărul de splituri al arborelui, eroarea relativă pe setul de antrenament(rel error) scade. Totuși, xerror (eroarea de validare încrucișată) nu scade semnificativ rezultând faptul că modelul poate fi supus unui

overfitting ușor. Importanța variabilelor ne arată ca **gross_primary_education_enrollment_country** este cea mai importantă variabilă, urmată de **gross_tertiary_education_enrollment**, **cpi_country** și **cpi_change_country**. Acestea au cea mai mare influență asupra predicțiilor modelului. Variabilele **age** și **gdp_country** nu sunt relevante în acest caz, iar **gender** nu a avut o importanță semnificativă în acest model. Așadar, modelul pare să aibă un potențial de overfitting, deoarece eroarea de validare încrucișată nu scade semnificativ cu creșterea complexității modelului.



Vom prezenta numărul de predicții corecte și greșite făcute de model în comparație cu valorile reale. Valori true positive sunt 186, false positive 127, false negative 167, true negative 240. Deci aceste erori indica necesitatea îmbunătățirii modelului pentru a reduce numărul de predicții incorecte.

```
> pred_m1 <- predict(m1, newdata = test_data, type = "class")
> confusionMatrix(pred_m1, test_data$finalWorth)
Confusion Matrix and Statistics
```

```

      Reference
Prediction High Low
High      186 127
Low       167 240

    Accuracy : 0.5917
    95% CI   : (0.5548, 0.6278)
  No Information Rate : 0.5097
  P-Value [Acc > NIR] : 6.078e-06

    Kappa : 0.1813

  Mcnemar's Test P-Value : 0.02293

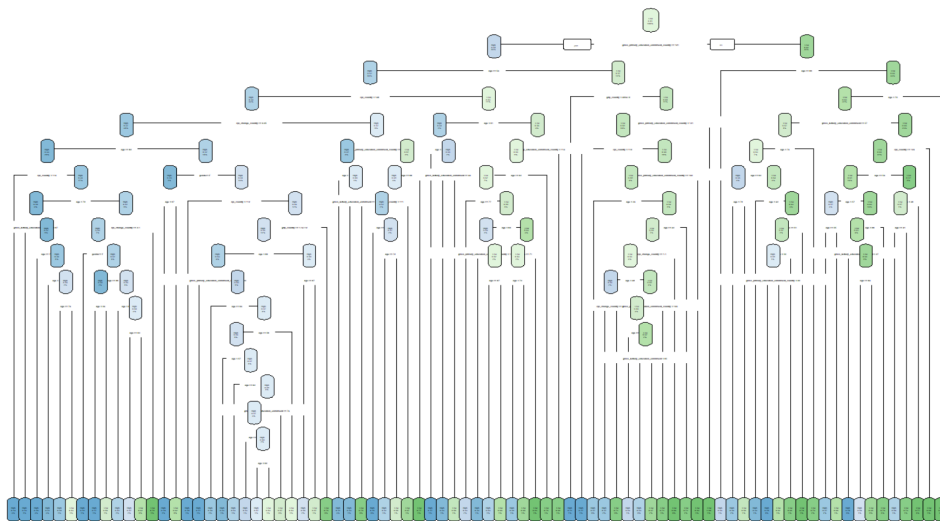
    Sensitivity : 0.5269
    Specificity : 0.6540
   Pos Pred Value : 0.5942
   Neg Pred Value : 0.5897
    Prevalence : 0.4903
  Detection Rate : 0.2583
  Detection Prevalence : 0.4347
   Balanced Accuracy : 0.5904

 'Positive' Class : High

```

Putem trage următoarele concluzii și anume faptul că acuratețea modelului este de 59.17%, ceea ce este decent, dar există loc de îmbunătățiri. Sensibilitatea și specificitatea arată că modelul are o performanță moderată în detectarea corectă a ambelor clase, cu specificitate puțin mai ridicată. Valoarea Kappa indică un acord scăzut între predicții și valori reale, ajustat pentru acordul întâmplător. Valoarea P pentru acuratețea comparată cu NIR arată că modelul performează semnificativ mai bine decât ghicitul aleatoriu.

Acesta e arborele m2 ce arată modelul rpart fără factorul CP. Se poate observa complexitatea sportita.



În continuare, am aplicat tehnica de Random Forest pentru a îmbunătăți performanța modelului de predicție a valorii finale a miliardarilor.

```
> # Modelul random forest
> set.seed(123)
> m1_rf <- randomForest(
+   formula = finalworth ~ gender + age + gross_tertiary_education_enrollment +
+   gross_primary_education_enrollment_country + gdp_country +
+   cpi_change_country + cpi_country,
+   data = train_data
+ )
> plot(m1_rf)
> print(m1_rf)
```

Call:

```
randomForest(formula = finalworth ~ gender + age + gross_tertiary_education_enrollment +
  gross_primary_education_enrollment_country + gdp_country + cpi_change_country + cpi_country, data = train_data)
```

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 2

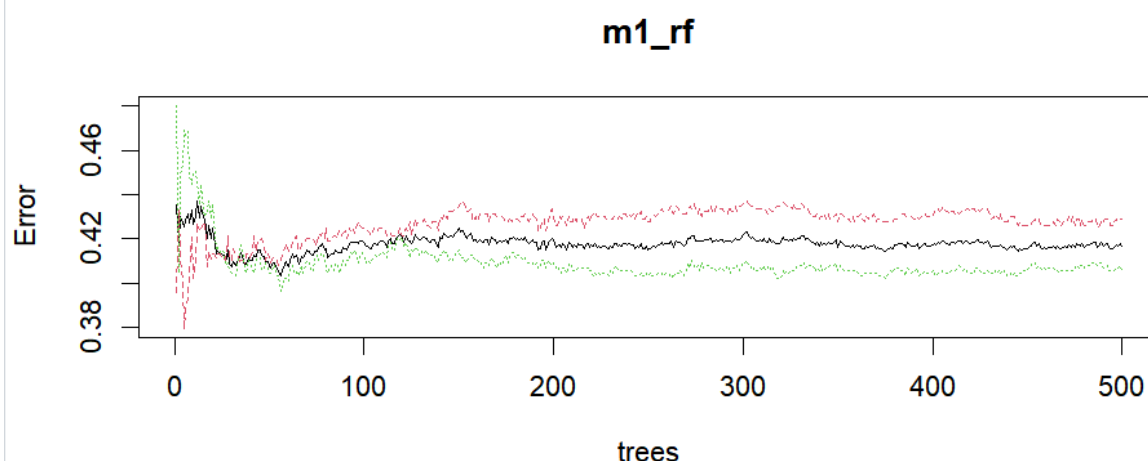
OOB estimate of error rate: 42.22%

Confusion matrix:

	High	Low	class.error
High	467	354	0.4311815
Low	354	502	0.4135514

> |

Modelul nostru utilizează 500 de arbori. Din matricea de confuzie observăm că modelul a clasificat corect 467 de observații High și 502 Low. Totuși, există un număr semnificativ de observații clasificate greșit, respectiv 354 pentru fiecare clasa. Așadar, rata de eroare OOB este de 42.22%.



La început, rata de eroare scade rapid pe măsură ce numărul de arbori crește. Linia neagră arată rata de eroare generală a modelului care converge către o valoare în jur de 0.42. Linia roșie reprezintă rata de eroare pentru clasa High, respectiv linia verde reprezintă rata de eroare pentru clasa Low. Graficul indică faptul că modelul Random Forest devine stabil după aproximativ 100 de arbori.

În concluzie, deși vârsta și educația de nivel superior au avut o influență notabilă, legătura generală între variabilele studiate și averea miliardarilor nu este foarte puternică. Totuși, modelul Random Forest a avut o rată de eroare OOB de 42.22%, fiind mult mai eficient față de restul modelelor aplicate în această analiză.

Concluzie

Analiza a avut ca obiectiv răspunsul la întrebările privind existența unei legături semnificative între averea miliardarilor și factori precum genul, vârsta, educația universitară (de nivel superior) și industria în care activează miliardarii.

Aplicând ambele metode (regresia logistică și arborele de decizie), rezultatele au arătat că metodele aplicate pe arborii de decizie sunt mai eficiente, în special random forest. Metoda random forest a prezentat valori mai bune ale sensibilității și specificității, ceea ce indică o clasificare mai bună. Astfel, putem concluziona că acea legătură există, chiar dacă este de o intensitate moderată.