Universitatea Babeș-Bolyai

Facultatea de Științe Economice și Gestiunea Afacerilor

FUNDAMENTE DE BIG DATA

Analiza factorilor care influențează performanța școlară

Studenți: Radu Teodora

teodora.radu1@stud.ubbcluj.ro

Stratu Mihaela

mihaela.stratu@stud.ubbcluj.ro

Specializarea: Informatică Economică

grupa 6

Introducere

În contextul actual al educației, identificarea factorilor care influențează performanța academică a elevilor este esențială pentru îmbunătățirea procesului de învățare și pentru asigurarea succesului în parcursul lor școlar. Una dintre variabilele cheie care poate afecta performanța academică este timpul dedicat studiului săptămânal. Prin urmare, acest studiu se concentrează pe investigarea relației dintre timpul dedicat studiului săptămânal și performanța academică a elevilor.

Astfel, întrebarea principală a analizei noastre este: "Care sunt factorii care influențează performanța academică a elevilor?". Ca întrebare suplimentară în cadrul cercetării noastre, ne-am angajat să investigăm dacă: "Există o corelație semnificativă între numărul de ore de activități extracurriculare și performanța academică a studenților?"

Acest studiu are o importanță crucială în peisajul educațional actual, deoarece furnizează date esențiale pentru profesori, părinți și organele de decizie din acest domeniu. Prin înțelegerea relației dintre timpul dedicat studiului săptămânal, gen și performanța academică, se pot dezvolta strategii educaționale mai eficiente și mai adaptate nevoilor individuale ale elevilor. Cadrele didactice pot folosi rezultatele studiului pentru a adapta metodele lor de predare și evaluare în funcție de nevoile specifice ale elevilor. Părinții pot beneficia de aceste informații pentru a sprijini în mod corespunzător studiul copiilor lor. De asemenea, cei din conducerea sistemului educațional pot utiliza rezultatele pentru a lua decizii informate cu privire la politici și programe educaționale.

Conform articolului, Chee, K. L., Pino, N., & Smith, W. L. (2005). Gender differences in the academic ethic and academic achievement.College Student Journal,39(3), 604–619 a fost dovedit faptul că în medie, femeile au rezultate academice mai bune decât bărbații.

Studiul "The Relationship between Students' Study Time and Academic Performance and its Practical Significance" realizat de Mukun Liu de la Universitatea Queen's din Kingston, Canada, evidențiază o relație pozitivă liniară între timpul investit de studenți în învățare și notele academice obținute. Totuși, cercetarea relevă și o limitare: după un anumit prag de timp dedicat studiului, creșterile ulterioare nu îmbunătățesc semnificativ notele studenților.

Setul de date

Setul de date ales este relevant pentru întrebările de cercetare de mai sus deoarece conține informații esențiale despre factorii care pot influența performanța academică a elevilor. Prin intermediul acestui set de date, putem investiga relația dintre timpul dedicat studiului săptămânal și performanța academică, precum și alte aspecte importante legate de absente, activități extracurriculare și scorurile la diferite materii. Având în vedere structura bogată a setului de date, care include detalii despre peste 2000 de studenți, putem examina în profunzime variabilele relevante și putem trage concluzii robuste în legătură cu întrebările noastre de cercetare.

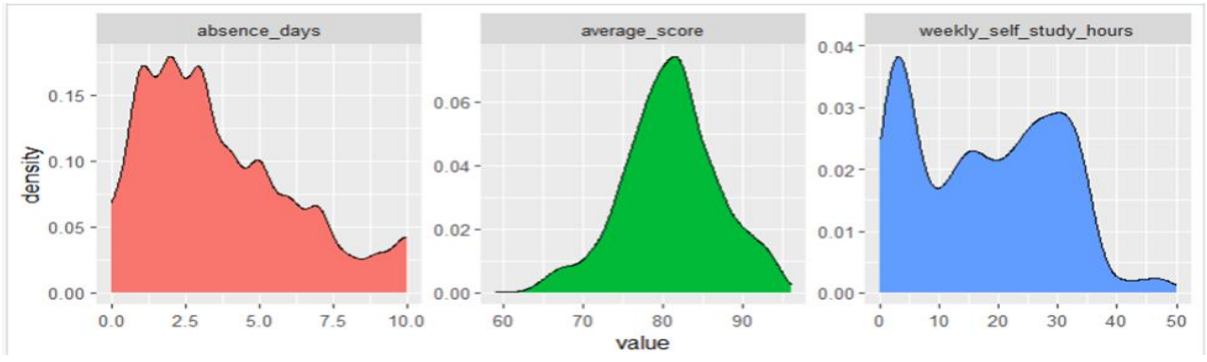
Setul de date pe baza căruia am făcut cercetarea a fost preluat de pe site-ul https://www.kaggle.com/datasets/mexwell/student-scores?resource=download.Autorul setului de date nu a oferit informații referitoare la școlile de unde au fost prelevate datele. Structura setului de date este reprezentată de 17 coloane: ”ID”- este un număr unic pentru fiecare student, ”first_name” și ”last_name”- compun numele studenților, ”email”- este adresa de email a fiecărui student, ”gender”- reprezintă sexul feminin sau masculin, ”part_time_job” - conține valoarea ”TRUE” dacă studentul lucrează part-time și ”FALSE” dacă nu, ”absence_days” – reprezintă numărul de zile în care studentul a absentat, ”extracurricular_activities” - conține valoarea ”TRUE” dacă studentul are alte activități în afara celor specifice studiului și ”FALSE” dacă nu, ”weekly_self_study_hours” – reprezintă numărul de ore saptamanale alocate studiului individual, ”career-aspiration” – cariera pe care studenul și-o dorește, ”math_score”, ”history_score”, ”physics_score”, ”chemistry_score”, ”biology_score”, ”english_score”, ”geography_score” – reprezintă nota finală la fiecare materie.

Curățarea datelor

Odată ce am încărcat setul de date în RStudio, a fost crucial să efectuăm o filtrare a coloanelor care nu erau relevante pentru cercetarea noastră. Spre exemplu, coloanele id, nume, prenume, email au fost eliminate deoarece în analiza noastra contează genul și acestea nu aveau un impact asupra cercetării. Coloanele ”math_score”, ”history_score”, ”physics_score”, ”chemistry_score”, ”biology_score”, ”english_score”, ”geography_score” au fost prelucrate rezultând media generală a fiecărui student și coloana cu denumirea Average_Scores.

```
studenti_clear    2000 obs. of 7 variables
    $ gender                          : Factor w/ 2
    $ part_time_job                   : logi   FALSE
    $ absence_days                    : int   3 2 9 5
    $ extracurricular_activities      : logi   FALSE
    $ weekly_self_study_hours         : int   27 47 1
    $ career_aspiration               : Factor w/ 17
    $ average_score                   : num   82 91.4
```

Vizualizarea datelor



Graficul de mai sus reprezintă o vizualizare a distribuției datelor numerice ce cuprinde numărul de zile absentate, scorul mediu obținut de elevi și numărul de ore de studiu individual saptamanal.

Pentru a putea aplica diverse metode asupra setului de date am stabilit variabila dependentă average_score. Pentru fiecare metodă utilizată am împărțit setul de date astfel: set de antrenament și set de test, unde am ales o proporție 70%-30%. Procesul de învățare a avut loc pe setul de antrenament, iar procesul testării s-a realizat pe setul de test. Pentru a putea studia și interpreta setul de date am ales să utilizăm metodele:

-Predicție numerică (regresii)

-Arbori de decizie (metoda bagging și randomForest)

Rezultate și discuții

Regresie liniara simpla

Calculând regresiile simple pentru variabilele independente din setul de date, se vor alege acelea care îndeplinesc următoarele criterii. Căutăm variabilele care au un RSE minim deoarece cu cât valoarea e mai mică cu atât modelul se potrivește mai bine pe setul de date. Apoi, căutăm variabile cu p-value-uri mici ce indica o asociere statistică considerabilă între variabila independentă și variabila dependentă. În plus, ne orientăm către variabile care generează R^2 mai mari pentru că exprimă o capacitate mai mare a modelului de a explica variația variabilei dependente. Prin aplicarea acestor criterii, putem identifica variabilele care sunt mai susceptibile să prezică eficient variabila dependentă în cadrul modelului nostru.

1.Media notelor în funcție de numărul de ore alocate studiului săptămânal

Rezultate

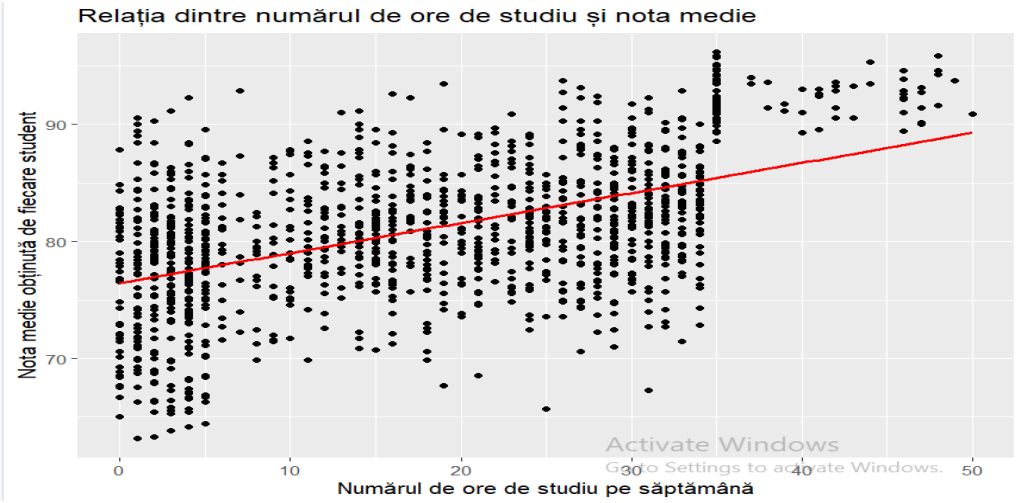
Estimările coeficienților indică modificarea medie a variabilei dependente atunci când variabila independentă se schimbă cu o unitate. În acest caz, interceptul este 76.44331, semnificand că se anticipează ca nota medie a unui student care nu studiază deloc este de aproximativ 76.44. Coeficientul pentru variabila (weekly_self_study_hours) este 0.25635, reliefând că, în medie, nota medie crește cu aproximativ 0.25635 pentru fiecare oră suplimentară de studiu pe săptămână.

R^2 are valoarea 0.2685, generand faptul că aproximativ 26.85% din variația notei medii este explicată de numărul de ore de studiu pe săptămână.Avem un p-value foarte mic ceea ce indică o asociere semnificativă între variabila noastră independentă și nota fiecărui elev(variabila dependentă).

Conform intervalului de confidență, pentru fiecare oră suplimentară de studiu pe săptămână, nota medie este așteptată să crească cu aproximativ între 0.2341 și 0.2786.

```
                              2.5 %     97.5 %
(Intercept)             75.9607524 76.9258678
weekly_self_study_hours  0.2341373  0.2785533
```

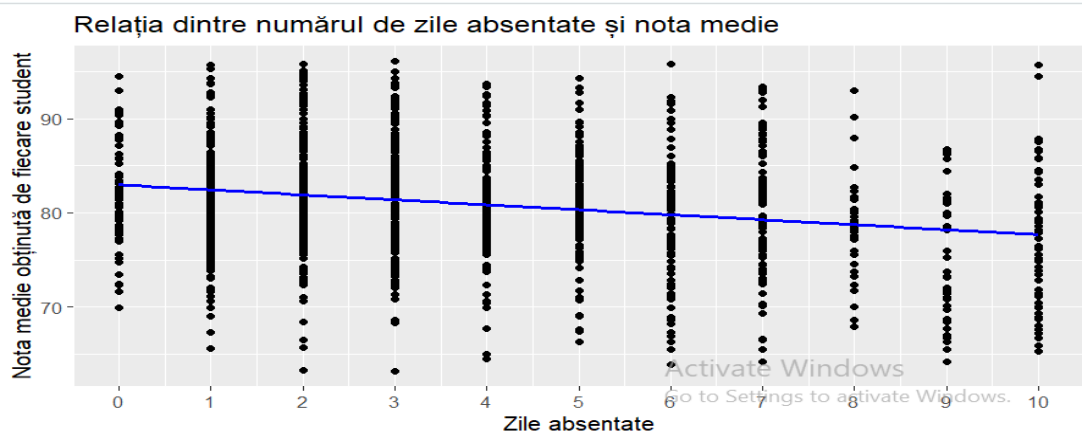
```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             76.44331    0.24599  310.75   <2e-16 ***
weekly_self_study_hours  0.25635    0.01132   22.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.19 on 1397 degrees of freedom
Multiple R-squared:  0.2685,    Adjusted R-squared:  0.268
F-statistic: 512.7 on 1 and 1397 DF,  p-value: < 2.2e-16
```



2. Media notelor în funcție de numărul de absențe

Rezultate

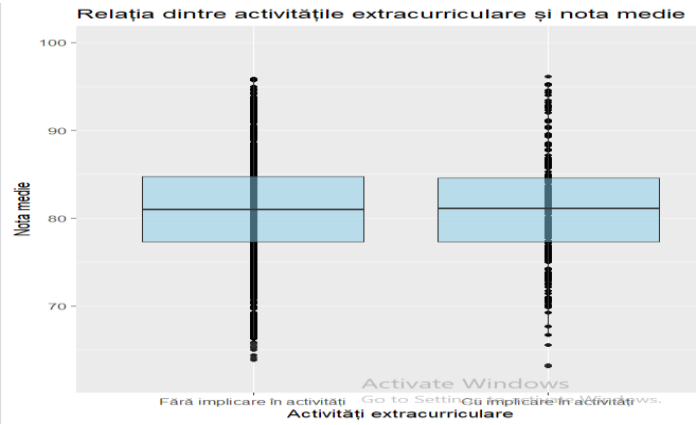
Conform datelor rezultate din regresie, se anticipează că nota medie a unui student care nu a lipsit deloc (0 zile de absență) este de aproximativ 82.98. Coeficientul pentru variabila "absence_days" este -0.52967, indicând că pentru fiecare zi absentată, se anunță o scădere de aproximativ 0.52967 la nota medie. În această regresie, R^2 este de 0.05119, simbolizand că aproximativ 5.119% din variația notei medii este explicată de zilele de absență. Interpretarea intervalului de confidență ar fi că pentru fiecare zi de absență în plus, nota medie este așteptată să scadă cu o valoare cuprinsa între 0.649 și 0.410.



3.Media notelor în funcție de participarea la activități extrașcolare sau nu

Rezultate

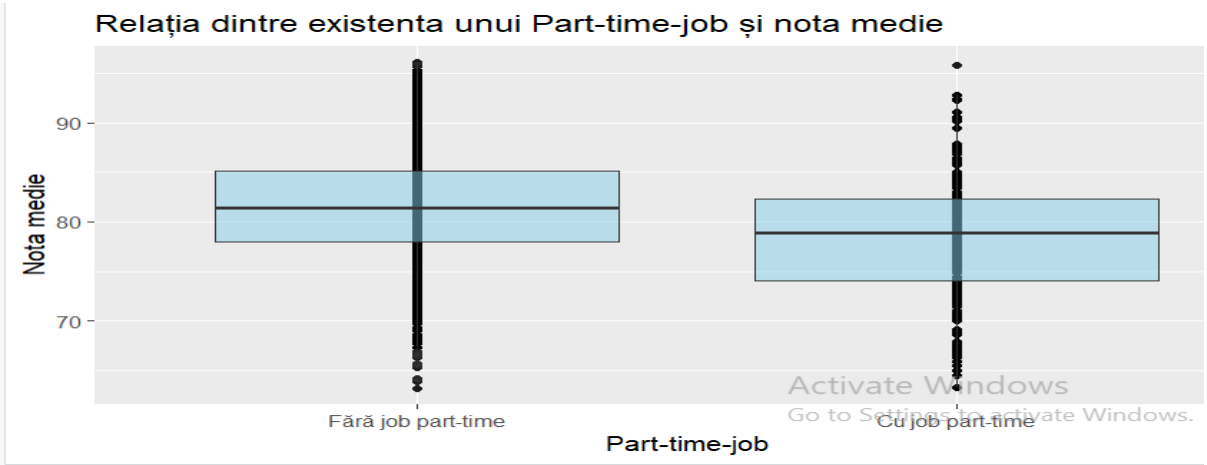
Coeficientul pentru variabila extracurricular_activities este -0.1832, fiind negativ, acesta sugerează că, în medie, elevii care nu participă la activități extracurriculare au o notă medie ușor mai mică decât cei care participă, însă această diferență nu este semnificativă statistic, având în vedere p-value-ul mare (0.65). RSE este 6.068, evidențiind faptul că deviația medie a observațiilor față de linia de regresie este de aproximativ 6.068 puncte. R-squared este foarte mic, 0.0001473 subliniind ca variabila extracurricular_activities nu are un efect semnificativ asupra notei medii.



4.Media notelor în funcție de statul de angajat sau nu

Rezultate

Nota medie a unui student care nu are loc de munca part-time este de 81.5722.Coeficientul pentru variabila part_time_job este -3.3517, indicând că studenții cu un loc de muncă part-time au o notă medie mai mică cu 3.35 puncte decât cei fără loc de muncă part-time. R-squared este 0.04064, simbolizand faptul că aproximativ 4.06% din variația notei medii a studentului este explicată de variabila mentionata. Dacă studentul are job part-time, atunci nota sa medie este de așteptat să scadă cu o valoare din intervalul [2.49703, 4.2064] .

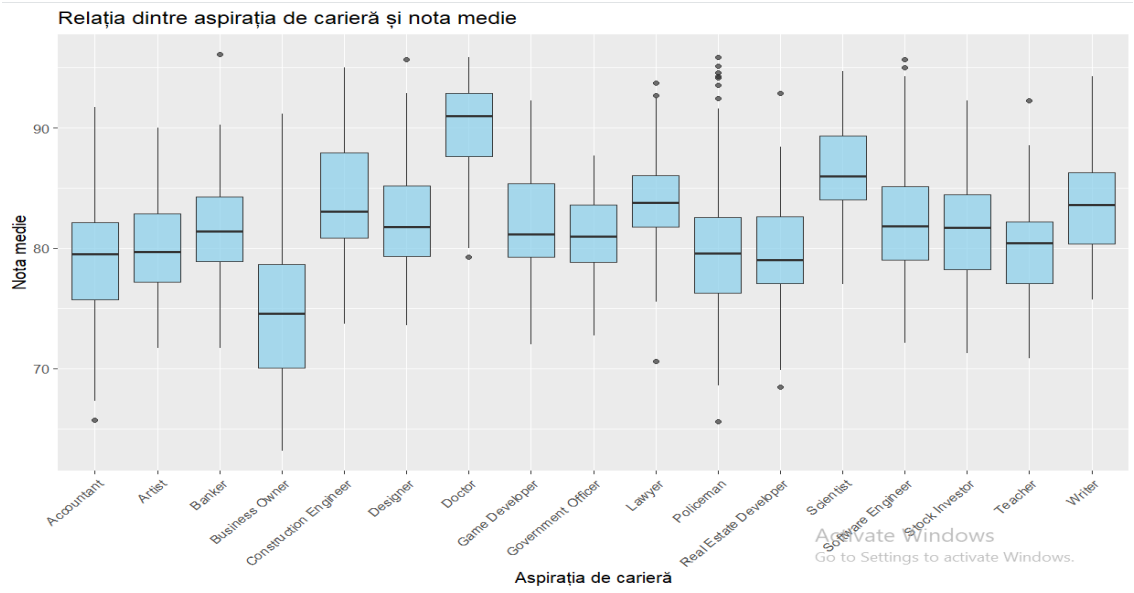


5.Media notelor în funcție de profesia dorită

Rezultate

Coeficienții pentru fiecare aspirație de carieră reprezintă schimbarea mediei notelor asociate cu o unitate de schimbare în aspirația de carieră respectivă, menținând celelalte variabile constante. De exemplu, coeficientul pentru "Doctor" este aproximativ 10.76, semnificand ca, în medie, aspirația de a deveni medic este asociată cu o creștere de aproximativ 10.76 puncte în media notelor. In schimb, elevii care doresc sa devina antreprenori au în medie note cu 4.70 puncte mai mici decat media.

Coeficienții cu un p-value mic indică o asociere semnificativă între aspirația de carieră respectivă și media notelor. De exemplu, valorile p-value pentru "Doctor", "Construction Engineer", "Scientist", "Lawyer", "Software Engineer" și "Writer" fiind mici, indică că există o corelație puternică între aceste aspirații de carieră și media notelor. Pe de altă parte, coeficienții cu p-value ridicat nu sunt considerați semnificativi statistic, sugerând o asociere mai slabă între aspirația de carieră asociată și performanța academică a elevilor(Government Officer,Policeman,Real Estate Developer, Teacher, Banker). R-squared este 0.3774, rezultând că 37.74% din variabilitatea mediei notelor poate fi explicată de aspirațiile de carieră incluse în model.




```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               79.1302     0.5074 155.946  < 2e-16 ***
career_aspirationArtist                    0.9959     0.8437   1.180 0.238055
career_aspirationBanker                    2.6388     0.6775   3.895 0.000103 ***
career_aspirationBusiness Owner           -4.7030     0.6078  -7.738 1.94e-14 ***
career_aspirationConstruction Engineer     5.0562     0.8725   5.795 8.44e-09 ***
career_aspirationDesigner                  3.0444     0.9493   3.207 0.001372 **
career_aspirationDoctor                   10.7582     0.7238  14.864  < 2e-16 ***
career_aspirationGame Developer            3.0659     0.8924   3.436 0.000609 ***
career_aspirationGovernment Officer        1.9048     0.8789   2.167 0.030385 *
career_aspirationLawyer                    4.7908     0.7099   6.748 2.19e-11 ***
career_aspirationPoliceman                 0.8785     0.6308   1.393 0.163947
career_aspirationReal Estate Developer     0.4827     0.7945   0.608 0.543548
career_aspirationScientist                 7.2746     1.0148   7.168 1.23e-12 ***
career_aspirationSoftware Engineer         3.2260     0.6052   5.330 1.14e-07 ***
career_aspirationStock Investor            2.0008     0.8604   2.325 0.020190 *
career_aspirationTeacher                   1.0493     0.9228   1.137 0.255714
career_aspirationWriter                    4.7015     1.0417   4.513 6.92e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.814 on 1382 degrees of freedom
Multiple R-squared:  0.3774,    Adjusted R-squared:  0.3702
F-statistic: 52.37 on 16 and 1382 DF,  p-value: < 2.2e-16
```

Metoda	Coeficient	Std. Error	t value	p-value	RSE	R^2
Regresie simpla weekly_self_study_hours	0.25635	0.01132	22.64	0	5.19	0.2685
Regresie simpla absence_days	-0.52967	0.06101	-8.681	0	5.911	0.05119
Regresie simpla gender	0.5104	0.3242	1.575	0.116	6.063	0.001771
Regresie simpla part_time_job	-3.3517	0.4357	-7.693	0	5.944	0.04064
Regresie simpla activitai_extra	-0.1832	0.4038	-0.454	0.65	6.068	0.0001473
Regresie simpla career_Aspiration					4.814	0.3774

Regresia cu numărul de ore de studiu pe săptămână ca predictor are un R-squared mai mare decât regresia cu zilele de absență, indicând că numărul de ore de studiu pe săptămână este un predictor mai puternic pentru nota medie decât zilele de absență. Cea mai mare influență asupra notei medii a unui student este posibila carieră pe care elevul și-o dorește. Modelele care au cel mai mare R^2 sunt :Regresia weekly_self_study_hours ,career_Aspiration, absence_days si part_time_job. În urma datelor de mai sus, modelele gender și activitati_extra nu au un efect semnificativ asupra notei medii, astfel că aceste variabile nu vor mai fi luate în considerare în calculele viitoare, importanța lor fiind scăzută.

Regresia liniară multiplă (weekly_self_study_hours+ career_aspiration+ part_time_job+absence_days)

Coeficientul de determinare (R^2) este 0.4098, ceea ce înseamnă că aproximativ 40.98% din variația în notele medii este explicată de variabilele din modelul nostru. Conform acestui model, pentru fiecare oră suplimentară petrecută învățând pe săptămână, se așteaptă o creștere a notei medii cu aproximativ 0.1665 puncte. În legătură cu coeficienții asociați cu diferitele aspirații de carieră , un student care aspiră să devină doctor are, în medie, o notă medie cu aproximativ 8.62 puncte mai mare decât un student care nu aspiră la această carieră. Coeficientul part_time este negativ (-0.23038), dar p-value-ul mare (0.54956) sugerează că nu există suficiente dovezi pentru a susține o corelație semnificativă între existența unui job și nota medie. Coeficientul absence_days este pozitiv (0.05366), dar p-value-ul mare (0.34901) sugerează că nu există suficiente dovezi pentru a susține o corelație semnificativă. Variabilele cele mai semnificative din acest model par să fie aspirațiile de carieră și orele de studiu individuale.

Interpretare valorilor (RMSE) este esențială pentru a înțelege cât de bine performează modelul și cum se generalizează acesta pe date noi.Aceasta valoare indică faptul că, în medie, predicțiile modelului sunt la aproximativ 4.66 puncte distanță de valorile reale pentru datele din setul de antrenament. Valorile RMSE pentru setul de antrenament și setul de test sunt foarte apropiate (4.66 pentru antrenament și 4.62 pentru test). Aceasta indică faptul că modelul nu suferă de overfitting (unde modelul ar performa foarte bine pe setul de antrenament, dar mult mai prost pe setul de test) sau underfitting (unde modelul ar performa prost pe ambele seturi de date).
Un RMSE mic și apropiat între cele două seturi de date indică faptul că modelul are o performanță bună și este capabil să facă predicții precise atât pe datele de antrenament, cât și pe datele de test.

Limitările rezultatelor

Totuși, aproape 59%(la regresie), din variația notei nu poate fi explicată, acesta fiind un procent destul de mare. Aceste limitari pot fi datorate existenței altor variabile importante care ar influența nota medie a studenților dar care nu au fost incluse în analiza noastră. Spre exemplu, nu avem date despre caracteristicile personale ale studenților (cum ar fi vârsta, mediu socio-economic) sau factori legați de mediul academic (calitatea profesorilor, resursele școlare etc.). Pentru a obține rezultate mai satisfăcătoare, am decis sa exploram și alte modele cum ar fi arborii de decizie ca să îmbunătățim performanța modelului.

Arbori de decizie

Ușurința de interpretare a arborilor de decizie face ca aceștia să fie preferați în mediul de cercetare, fiind folosiți în mai mult de 50% dintre modele. Performanța acestora este mai scăzută însă, cu ajutorul ansamblurilor de arbori, vom reuși să creștem acest factor. De aceea, pentru analiza noastră, am ales ca în continuare să folosim arbori de decizie folosind metodele bagging și random Forest.

Am început analiza prin stabilirea unei valori seed pentru a fi consistenți în ceea ce privește reproductibilitatea, deoarece atunci când se realizează împărțirea datelor se aleg date aleatoare. Apoi am împărțit setul de date în setul de antrenament și setul de test cu proporția de 70%-30%. Astfel, am obținut setul de antrenament cu 1400 instanțe și setul de test cu 600 de instanțe.

Pentru început am format un arbore în care am folosit ca formula, variabila dependentă împreună cu toate celelalte variabile disponibile și am observat faptul că variabilele “gender”, “absence_days”, “part_time_job” și “extraculicural_activities” nu influențează ramificarea arborelui nostru așa că am decis să le omitem în continuarea cercetării noastre. Astfel, variabila “average_score” este dependentă de variabilele independente ”self_study_hw” și ”career_aspirations” cu ajutorul cărora am creat arborele de decizie.

Mai departe am aplicat funcția ”rpart()” care a creat un arbore de decizie și l-a atribuit variabilei ”m1”. La crearea acestuia, am avut în vedere faptul că atributul țintă este ”average_score” iar în formula de modelare am adăugat variabilele de care media studenților este dependentă. Am

indicat metoda folosită ca fiind ”anova”. După rulare, afișăm în mod text arborele pentru a putea observa conținutul fiecărui nod.

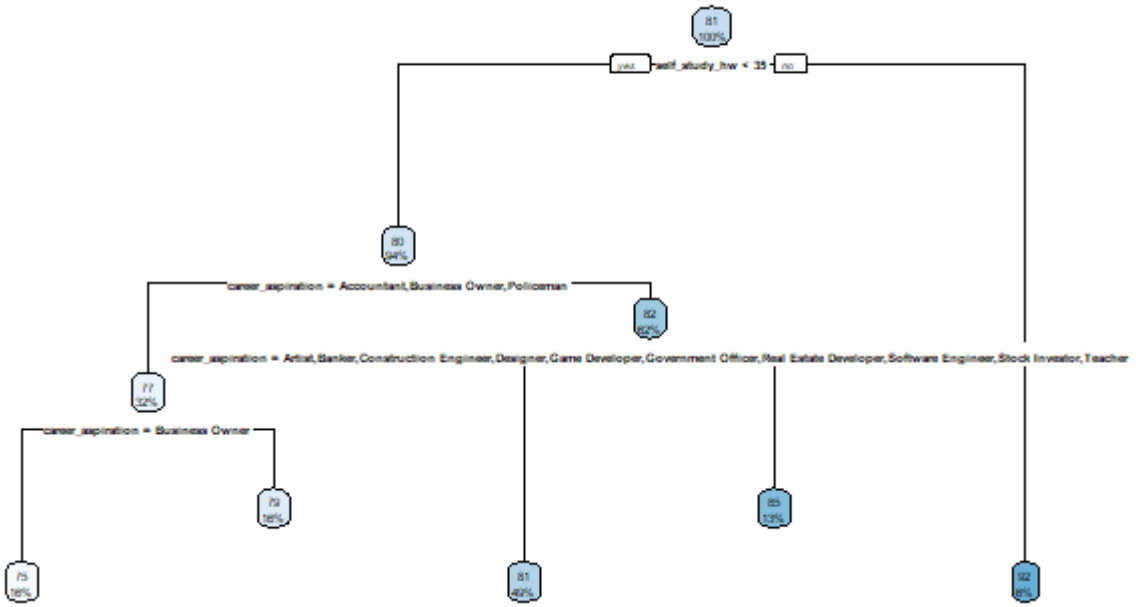
```
1) root 1400 50409.3500 80.91796
```

Mai sus este rezultatul nodului rădăcină ce oferă informatii referitoare la numărul de observații care sunt cuprinse în nod înainte de a fi divizat (1400), devierea totală (50409.3500) și deviația medie sau SSE-ul (Sumar squerd error) (80.91796).

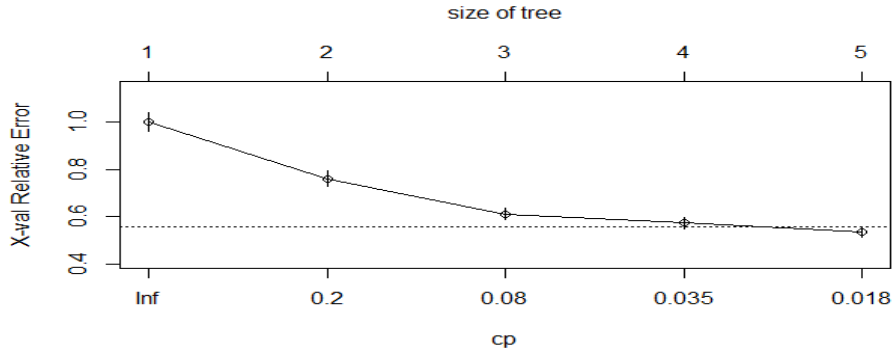
Afișarea arborelui de decizie în mod text ne ajută să observăm datele din fiecare nod și astfel remarcăm cum, odată cu scăderea numărului de observații din noduri valoarea SSE-ului scade.

```
      * denotes terminal node

 1) root 1400 50409.3500 80.91796
   2) self_study_hw< 34.5 1313 37975.7300 80.16005
     4) career_aspiration=Accountant,Business Owner,Policeman 443 13574.6700 76.61271
       8) career_aspiration=Business Owner 219  7017.3680 74.51272 *
       9) career_aspiration=Accountant,Policeman 224  4647.3100 78.66582 *
     5) career_aspiration=Artist,Banker,Construction Engineer,Designer,Doctor,Game De
veloper,Government Officer,Lawyer,Real Estate Developer,Scientist,Software Engineer,S
tock Investor,Teacher,Writer 870 15987.9700 81.96634
      10) career_aspiration=Artist,Banker,Construction Engineer,Designer,Game Develop
er,Government Officer,Real Estate Developer,Software Engineer,Stock Investor,Teacher
690 11989.8000 81.27039 *
      11) career_aspiration=Doctor,Lawyer,Scientist,Writer 180  2382.9050 84.63413 *
   3) self_study_hw>=34.5 87   296.6479 92.35632 *
```

Afișăm apoi arborele de decizie în mod grafic utilizând funcția: ”rpart.plot(m1)”. Astfel, în cadrul arborelui se afișează condiția, numărul de instanțe și ponderea lor din total, condiția și valoarea prezisă.



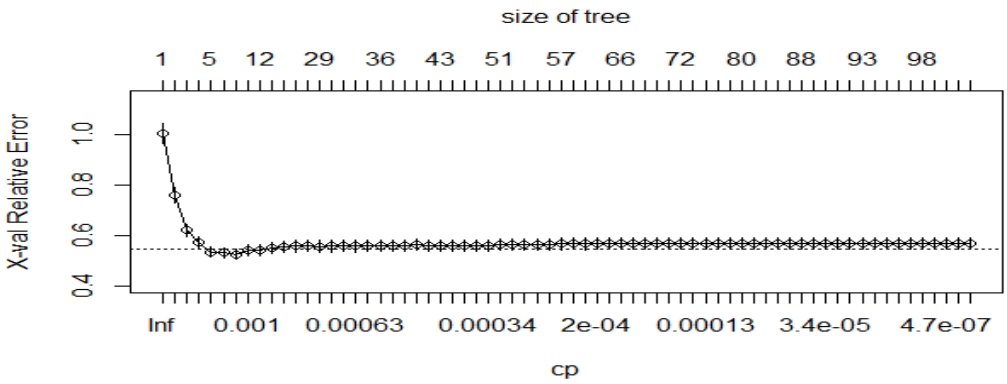
Atunci când SSE-ul nu se va mai îmbunătăți și deci $minimize\{SSE + \alpha|T|\}$ nu va mai fi cel mai mic, va fi selectat de către algoritm, parametrul α(cost complexity). Afișăm graficul costului complexității prin ” plotcp(m1)”.



Observăm că α începe de la ponderi foarte mari iar eroarea relativă este 1.0. Când α va avea o pondere de 0.2, eroarea relativă scade mai jos de 0.8 adică cu mai mult de 0.2. Continuând, α va ajunge la valoarea minimă 0.018 oprindu-se în acest punct, mărimea arborelui fiind 5.

```
          CP nsplit rel error    xerror       xstd
1 0.24076822      0 1.0000000 1.0013259 0.03992917
2 0.16689536      1 0.7592318 0.7602012 0.03249654
3 0.03788971      2 0.5923364 0.6129454 0.02405881
4 0.03204309      3 0.5544467 0.5768689 0.02264590
5 0.01000000      4 0.5224036 0.5298172 0.02108684
```

Setarea unui cp de 0 va rezulta într-un arbore imens, cu foarte multe noduri făcându-se overplotting, însă eroarea nu va mai scădea față de cea inițială.



Următorul pas în analiza noastră este găsirea unui arbore optim. Pentru acest lucru am setat numărul minim de exemple pentru ca divizarea să se facă la 10 și adâncimea maximă la 5. Construirea unui hyper-grid este necesară pentru a găsi toate combinațiile dintre minsplit și maxdepth. Am creat o secvență de valori pentru minsplit de la 5 la 20 având pasul 1 și o secvență de valori pentru maxdepth de la 5 la 9 având de asemenea pasul 1. Vor rezulta 80 de astfel de combinații, pentru fiecare dintre acestea construindu-se un arbore de decizie. Pentru toate aceste combinații vom crea câte un model iar toate modelele le vom adăuga într-o listă. Creăm apoi două funcții care extrag: parametrul de complexitate pentru valoarea minimă a erorii

```
get_cp <- function(x) {
  min <- which.min(x$cptable[,"xerror"])
  cp <- x$cptable[min, "CP"]
}
```

și eroarea minimă

```
get_min_error <- function(x) {
  min <- which.min(x$cptable[, "xerror"])
  xerror <- x$cptable[min, "xerror"]
}
```

	minsplit	maxdepth	cp	error
1	5	5	0.01	0.5353842
2	6	5	0.01	0.5307628
3	7	5	0.01	0.5295313
4	8	5	0.01	0.5314467
5	9	5	0.01	0.5323947
6	10	5	0.01	0.5358911
7	11	5	0.01	0.5306542
8	12	5	0.01	0.5344952
9	13	5	0.01	0.5322116
10	14	5	0.01	0.5344128
11	15	5	0.01	0.5301506
12	16	5	0.01	0.5313126
13	17	5	0.01	0.5295810
14	18	5	0.01	0.5352437

Adăugăm apoi, cu ajutorul funcției ”mutate()” două noi coloane hyper-grid-ului: coloana cp care conține parametrul de complexitate pentru fiecare model și coloana error care conține eroarea minimă a fiecărui model. Aceste informații sunt obținute cu ajutorul funcțiilor definite mai sus.

Vom extrage și vom afișa apoi primele 5 combinații din hyper-grid pentru a o găsi pe cea optimă.

```
  minsplit maxdepth   cp     error
1       14         6 0.01 0.5276678
2        6         7 0.01 0.5287086
3       15         9 0.01 0.5290950
4       20         7 0.01 0.5292953
5        7         8 0.01 0.5293316
```

După cum se observă și mai sus, combinația optimă este: minsplit 14, maxdepth 6 și cp 0.01.

Pasul următor va fi să aplicăm pe arborele optim setul de test și apoi vom calcula eroarea pătratică medie între predicții și valorile reale din setul de test. Pentru datele noastre, RMSE va avea valoarea egală cu 4.694673.

La generarea în format text a arborelui optim se observă că împărțirea se face în primul rând după orele individuale de studiu per săptămână iar apoi după cariera la care aspiră studentul, astfel arătându-se că variabila ”self_study_hw” afectează mai mult variabila dependentă average score decât variabila ”career_aspirations”.

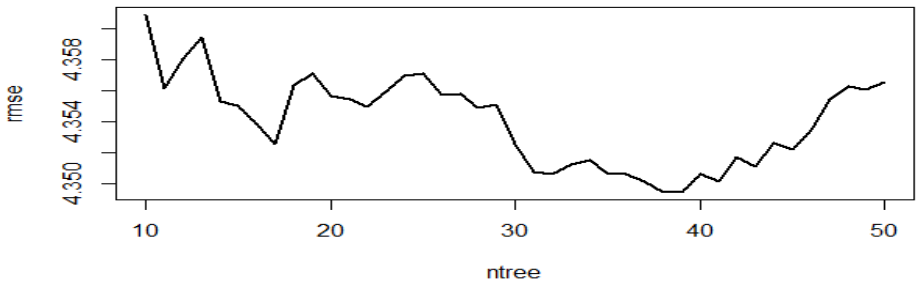
```
 1) root 1400 50409.3500 80.91796
   2) self_study_hw< 34.5 1313 37975.7300 80.16005
     4) career_aspiration=Accountant,Business Owner,Policeman 443 13574.6700 76.61271
       8) career_aspiration=Business Owner 219  7017.3680 74.51272 *
       9) career_aspiration=Accountant,Policeman 224  4647.3100 78.66582 *
     5) career_aspiration=Artist,Banker,Construction Engineer,Designer,Doctor,Game De
veloper,Government Officer,Lawyer,Real Estate Developer,Scientist,Software Engineer,S
tock Investor,Teacher,Writer 870 15987.9700 81.96634
       10) career_aspiration=Artist,Banker,Construction Engineer,Designer,Game Develop
er,Government Officer,Real Estate Developer,Software Engineer,Stock Investor,Teacher
690 11989.8000 81.27039 *
       11) career_aspiration=Doctor,Lawyer,Scientist,Writer 180  2382.9050 84.63413 *
   3) self_study_hw>=34.5 87   296.6479 92.35632 *
```

Pentru a îmbunătății performanța analizei noastre, vom calcula RMSE -ul și prin procedura bagging. Am creat mai întâi un model de bagging în care am precizat formula arborelui de decizie, setul de date care va fi utilizat (studenti_train) și estimarea erorii out-of-bag.

```
Bagging regression trees with 25 bootstrap replications

Call: bagging.data.frame(formula = average_score ~ self_study_hw +
    career_aspiration, data = studenti_train, coob = TRUE)

Out-of-bag estimate of root mean squared error:  4.3655
```

Observăm că avem 25 bootstrap replications iar valoarea RMSE va fi îmbunătățită, ajungând la 4.3655, valoarea precedentă fiind mai mare, 4.694673.

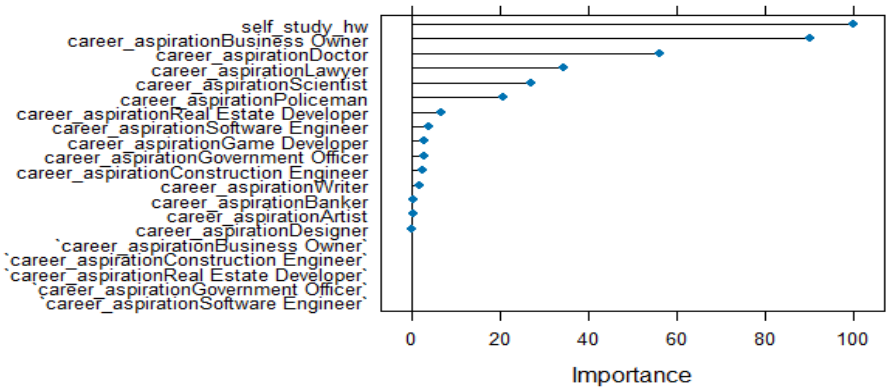
Definim numărul de arbori pentru bagging utilizând o secvență de numere de la 10 la 50 (ntree) și un vector în care vor fi stocate rezultatele RMSE corespunzătoare. Creăm apoi un model de bagging pentru fiecare valoare din tree, iar rezultatul fiecărui model îl vom stoca în vectorul declarat anterior.



După cum se observă din graficul prezentat mai sus, odată cu creșterea numărului de bag-uri, eroarea va scădea, ceea ce demonstrează că metoda aceasta este eficientă însă, după un anumit punct optim (în jur de 40), adăugarea altor arbori nu aduce îmbunătățiri ci poate afecta în mod negativ performanța.

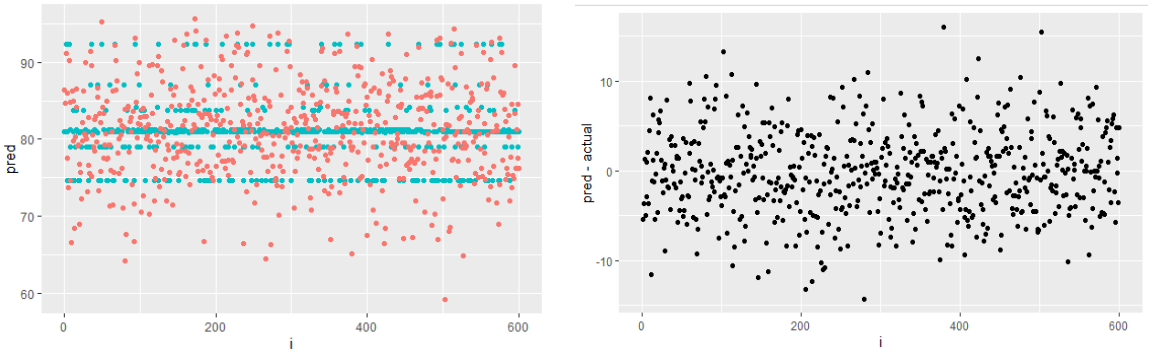
Vom utiliza și metoda bagging cu caret. Creăm un obiect pentru a defini modul de control al antrenamentului apoi antrenăm un model de bagging pe setul de date de antrenament. Următorul grafic afișează variabilele și importanța lor pentru modelul de bagging. Așa cum am

specificat și în cadrul arborilor de decizie, variabila self_study_hw influențează cel mai mult variabila average_score.



După realizarea predicțiilor pe setul de test observăm că valoarea RMSE este 4.692271, mai mare decât valoarea obținută anterior cu modelul de bagging.

Generăm apoi două grafice cu ajutorul unui dataframe. Graficul din dreapta reprezintă cu culoarea roșie valorile prezise prin acest model la setul de date studenti_test și cu culoarea albastră valorile actuale. Graficul din stânga reprezintă diferența dintre predicție și valoarea actuală pentru fiecare observație din setul de test.



Vom calcula valoarea RMSE prin metoda random forest. Pentru a face acest lucru am avut nevoie de libraria ”randomForest()”. Am setat din nou un seed pentru reproductibilitate, formula arborelui va avea în vedere variabila țintă ”average_score” și predictorii ”career_aspirations” și ”self_study_hw” folosind setul de antrenament.

```
Call:
 randomForest(formula = average_score ~ self_study_hw + career_aspiration,      data
= studenti_train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 19.90332
                    % Var explained: 44.72
```

Tipul arborilor va fi de regresie, modelul conține 500 de arbori, media pătratică a reziduurilor (MSE) este de 19.90332 iar acest model explică doar 44.72% din varianța din variabila

dependentă. Observăm că valoarea RMSE este de 4.456118, care este mai mică decât valoarea obținută prin arbori de decizie dar mai mare decât cea obținută prin metoda bagging.

Comparație între modele

În urma cercetării noastre am observat faptul că modelul care are valoarea RMSE cea mai mică pentru setul de date de test este modelul arborilor de regresie formați cu metoda bagging. Pentru acesta am obținut valoarea de 4.3655 comparativ mai mică față de valorile obținute prin metodele utilizate în cadrul cercetării cu ajutorul celorlalte metode și modele. Prin metodele folosite am obținut următoarele rezultate ale RMSE-ului, în ordine descrescătoare a valorii: metoda arborelui optim 4.694673, metoda bagging cu caret 4.692271, metoda regresiei liniare multiple 4.62, metoda random Forest 4.456118, metoda bagging 4.7823. Astfel, prin compararea rezultatelor obținute pentru media diferenței valorilor prezise față de cele actuale, modelul de regresie bazat pe modelul arborilor bagging este cel mai potrivit pentru studiul nostru.

Concluzii

În urma cercetării noastre, am identificat următorii factori principali care influențează performanța academică a elevilor: orele de studiu săptămânal, aspirațiile de carieră, numărul de absențe și locul de muncă part-time. Există o corelație pozitivă semnificativă între numărul de ore de studiu săptămânal și media notelor, fiecare oră suplimentară de studiu fiind asociată cu o creștere a notei medii. De asemenea, elevii cu aspirații de carieră clare obțin performanțe academice mai bune, aceste aspirații motivându-i să obțină note mai mari. În contrast, numărul de zile de absență are o corelație negativă cu media notelor, fiecare zi de absență suplimentară ducând la o scădere a notei medii. Elevii cu locuri de muncă part-time tind să aibă note mai mici comparativ cu cei care nu lucrează. În ceea ce privește activitățile extracurriculare, analiza noastră nu a găsit o corelație semnificativă între acestea și performanța academică, sugerând că participarea la astfel de activități nu influențează în mod semnificativ notele elevilor din setul de date analizat. În plus, media notelor aferentă elevilor de gen masculin este puțin mai mare decât cea a celor de sex feminin, dar diferența nu este semnificativă statistic, pentru că avem un p-value mare. Prin urmare, nu putem concluziona cu încredere că bărbații au note mai bune decât femeile pe baza acestui model.