UNIVERSITATEA BABEŞ-BOLYAI
Facultatea de Ştiinţe Economice şi Gestiunea Afacerilor

Proiect
Fundamente de Big Data

Crișan Ioana
Vlad Adina-Miruna

2022

Cuprins

1. Introducere

Un domeniu foarte important al economiei este reprezentat de piața imobiliarelor. Pornind de la această afirmație am decis ca aria de cercetare a acestui proiect să fie reprezentată de piața imobiliarelor în unele dintre cele mai mari orașe din Polonia.

Folosindu-ne de un set structurat de date am încercat să determinăm relații și modele specifice și inteligibile pentru a ne ajuta să ajungem la o concluzie validă în legătură cu modul în care se desfășoară piața imobiliară din țara aleasă.

Factorul principal analizat în cadrul acestui proiect este reprezentat de preț, mai exact nivelul acestuia în orașele Warszawa, Krakow și Poznan și factorii determinanți pe care acesta îi are.

Întrebările cele mai importante cărora dorim să le aflăm răspunsul ca urmare a analizei efectuate sunt:

- Prețul unei locuințe din Polonia este influențat de orașul în care se află, numărul de camere pe care îl are, etajul la care este și suprafața pe care o are?
- Care sunt cei mai importanți factori care determină prețul?
- Metodele cu ajutorul cărora analizăm relațiile dintre factori vor avea același rezultat pentru setul de date folosit?

Am fost interesate să studiem piața imobiliară deoarece studiind primele trei semestre de facultate economia am înțeles cât de important poate să devină venitul pasiv în viața unui om cu o puternică stabilitate financiară. Față de diferite investiții în acțiuni, obligațiuni etc., investițiile în imobiliare oferă investitorilor bunuri tangibile, mai exact case și apartamente ce nu sunt doar surse de venit ci și surse de momente de bucurie și locuri de creștere și dezvoltare pentru persoanele care locuiesc în acestea. Imobiliarele sunt o necesitate primară a fiecărui om în parte și de aceea găsirea unei locuințe potrivite este esențială tuturor.

Ca arie de cercetare am hotărât să ne axăm pe unele dintre cele mai importante orașe din Polonia, mai exact Varșovia, Cracovia și Poznan, deoarece această țară este asemănătoare României atât din punct de vedere economic cât și al modului de gândire și comportament al populației. Totodată, având ca interese și călătoritul am fost curioase să aflăm mai multe despre economia unei alte țări europene.

2. Setul de date

Setul de date folosit este *Houses.csv* , preluat de pe site-ul Kaggle, mai exact https://www.kaggle.com/datasets/dawidcegielski/house-prices-in-poland .

Pentru a nu supraîncărca componentele hardware avute la dispoziție ne-am decis să lucrăm cu un maxim de 4200 de intrări. Totodată, deoarece au existat diferite caractere care îngreunau buna funcționare a R Studio am redenumit numele orașelor cu caractere obișnuite.

Astfel, inițial când am început să lucrăm cu setul de date în R Studio acesta a avut unsprezece coloane: numerotarea, adresa, orașul, etajul, numărul de identificare, latitudinea, longitudine, preț, camere, suprafață. Suprafața fiind exprimată în metri pătrați iar prețul în zloți polonezi.

Într-un nou proiect în RStudio am încărcat tabelul cu intrări ale caselor într-o variabilă denumită generic *"Houses"*. Acest lucru a fost făcut cu ajutorul următoarei linii de cod:

```
Case <- read_csv("Houses.csv")
View(Case)
```

Am reușit vizualizarea tabelului memorat în variabila *Case* cu ajutorul comenzii View, rezultatul fiind:

	...1	address	city	floor	id	latitude	longitude	price	rooms	sq	year
1	0	Podg�rze Zab�ocie Stanis�awa Klimeckiego	Krakow	2	23918	50.04922	19.97038	749000.0	3	74.05	2021
2	1	Praga-Po�udnie Grochowska	Warszawa	3	17828	52.24977	21.10689	240548.0	1	24.38	2021
3	2	Krowodrza Czarnowiejska	Krakow	2	22784	50.06696	19.92002	427000.0	2	37.00	1970
4	3	Grunwald	Poznan	2	4315	52.40421	16.88254	1290000.0	5	166.00	1935
5	4	Ochota Gotowy budynek. Stan deweloperski. Ostatnie ...	Warszawa	1	11770	52.21222	20.97263	996000.0	5	105.00	2020
6	5	Nowa Huta Czy�yny ul. Wo�nic�w	Krakow	2	26071	50.04694	19.99715	414600.0	1	34.55	2022
7	6	Podg�rze P�asz�w Koszykarska	Krakow	0	22569	50.04989	19.99060	750000.0	4	81.40	2021
8	7	Mokot�w Pory	Warszawa	10	13308	52.18406	21.04430	2890000.0	6	280.00	2003
9	8	Ursyn�w Wy�yny	Warszawa	3	11387	52.14028	21.05635	615000.0	4	63.40	1982
10	9	Bemowo	Warszawa	1	10904	52.23897	20.91329	429000.0	1	40.00	1999
11	10	�r�dmie�cie	Warszawa	0	16251	52.23281	21.01907	375000.0	1	29.00	1968
12	11	Praga-Po�udnie Goc�aw	Warszawa	2	13355	52.22863	21.10657	520000.0	3	67.00	1989
13	12	Bia�o��ka	Warszawa	0	15740	52.31966	21.02118	400000.0	3	57.00	2020
14	13	Nowe Miasto Malta Florentyny Luboi�skiej 5	Poznan	0	6081	52.39161	16.99406	421427.0	3	60.29	2019
15	14	Wola	Warszawa	0	12035	52.23624	20.95478	591771.8	2	52.27	2021
16	15	Grunwald �wierzawska	Poznan	8	1085	52.40066	16.91973	547000.0	4	77.39	2020
17	16	Bia�o��ka	Warszawa	2	10118	52.31966	21.02118	489000.0	2	52.00	2005
18	17	Grunwald G�rczyn ul. Ceglana	Poznan	1	3518	52.37099	16.86315	618636.0	3	66.52	2020
19	18	Krowodrza Stanis�awa Konarskiego	Krakow	2	27706	50.08404	19.97816	280000.0	1	16.20	1930

Showing 1 to 20 of 4,199 entries, 11 total columns

Pentru o analiză cât mai concretă și mai relevantă ne-am decis să nu luăm în calcul factori precum adresa, id-ul, latitudinea, longitudinea și anul construcției. Astfel, printr-un select am păstrat doar acei factori pe care am vrut să-i analizăm într-o nouă variabilă numită *PreţuriCase*.

```
PreturiCase <- select(Case, city, floor , price, rooms, sq)
View(PreturiCase)
```

Pentru o legibilitate sporită am redenumit coloanele în limba română cu ajutorul funcției rename.

```
PreturiCase <- rename(PreturiCase, c("Oras" = "city"))
PreturiCase <- rename(PreturiCase, c("Etaj" = "floor"))
PreturiCase <- rename(PreturiCase, c("Pret" = "price"))
PreturiCase <- rename(PreturiCase, c("NrCamere" = "rooms"))
PreturiCase <- rename(PreturiCase, c("Suprafata" = "sq"))
View(PreturiCase)
```

Astfel coloana city s-a redenumit Oras, coloana floor a devenit Etaj, coloana price s-a transformat în preț, coloana rooms a ajuns NrCamere iar sq s-a schimbat în Suprafață.

	Oras	Etaj	Pret	NrCamere	Suprafata
1	Krakow	2	749000.0	3	74.05
2	Warszawa	3	240548.0	1	24.38
3	Krakow	2	427000.0	2	37.00
4	Poznan	2	1290000.0	5	166.00
5	Warszawa	1	996000.0	5	105.00
6	Krakow	2	414600.0	1	34.55
7	Krakow	0	750000.0	4	81.40
8	Warszawa	10	2890000.0	6	280.00
9	Warszawa	3	615000.0	4	63.40
10	Warszawa	1	429000.0	1	40.00
11	Warszawa	0	375000.0	1	29.00
12	Warszawa	2	520000.0	3	67.00
13	Warszawa	0	400000.0	3	57.00
14	Poznan	0	421427.0	3	60.29
15	Warszawa	0	591771.8	2	52.27
16	Poznan	8	547000.0	4	77.39
17	Warszawa	2	489000.0	2	52.00
18	Poznan	1	618636.0	3	66.52
19	Krakow	2	280000.0	1	16.20

Showing 1 to 20 of 4,199 entries, 5 total columns

Am vrut apoi să aflăm mai exact ce tipuri de date are fiecare coloană în parte și astfel am rulat comanda str() pe variabila noastră.

```
> str(PreturiCase)
tibble [4,199 x 5] (S3: tbl_df/tbl/data.frame)
 $ Oras     : chr [1:4199] "Krakow" "Warszawa" "Krakow" "Poznan" ...
 $ Etaj     : num [1:4199] 2 3 2 2 1 2 0 10 3 1 ...
 $ Pret     : num [1:4199] 749000 240548 427000 1290000 996000 ...
 $ NrCamere : num [1:4199] 3 1 2 5 5 1 4 6 4 1 ...
 $ Suprafata: num [1:4199] 74 24.4 37 166 105 ...
```

Astfel am aflat că toate coloanele sunt numerice în afară de cea Oraș care este string.

Unele date precum Oraș, Etaj și Nr Camere nu sunt utile sub formă numerică iar de aceea acestea au fost transformate în factori, acest lucru a fost făcut cu o structuri de cod ca cea de mai jos, acestea fiind aplicate în parte fiecărei coloane ce a avut parte de modificări în acest pas.

```
PreturiCase <- PreturiCase %>%
  mutate(
    Oras=factor(Oras)
  )
```

Am vizualizat apoi tipul de date pentru a ne convinge că modificarea a fost făcută:

```
> str(PreturiCase$Oras)
 Factor w/ 3 levels "Krakow","Poznan",..: 1 3 1 2 3 1 1 3 3 3 ...
```

Prin urmare s-au format 3 nivele pentru fiecare oraș analizat:

```
> levels(PreturiCase$Oras)  #vizualizare nivele factor
[1] "Krakow"   "Poznan"   "Warszawa"
```

Fiecărui oraș îi corespunde un anumit număr de intrări, mai exact:

```
> freq_oras<-table(PreturiCase$Oras)
> freq_oras

  Krakow   Poznan Warszawa
    1704      731     1764
```

Numărul maxim de camere fiind 8, am obținut următorul tabel de frecvență:

```
> freq_NrCamere<-table(PreturiCase$NrCamere)
> freq_NrCamere

   1    2    3    4    5    6    7    8
 473 1568 1484  527  115   28    3    1
```

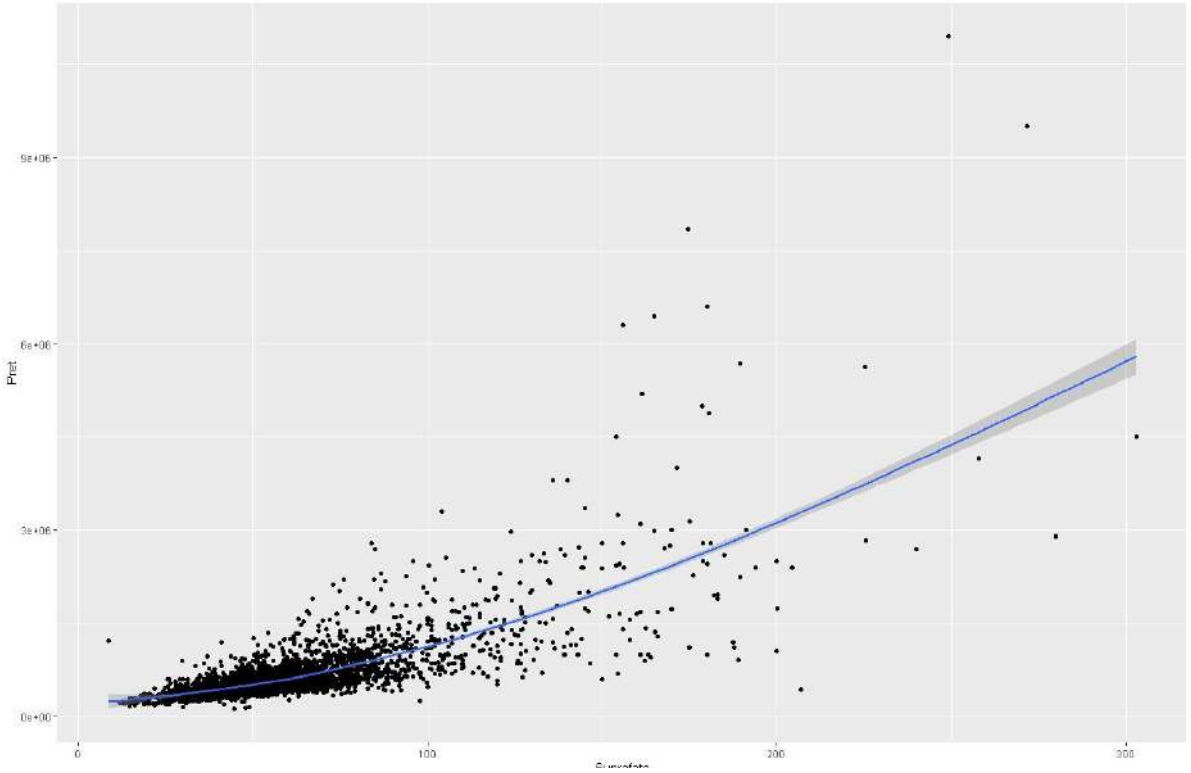
Iar numărul maxim de etaje este de 10, acestora corespunzându-le următoarele numere de frecvență:

```
> freq_Etaj<-table(PreturiCase$Etaj)
> freq_Etaj

  0   1   2   3   4   5   6   7   8   9  10
685 892 727 641 502 248 141 109  92  37 125
```

După ce am făcut aceste schimbări am verificat încă o dată tipurile de date pe care le avem:

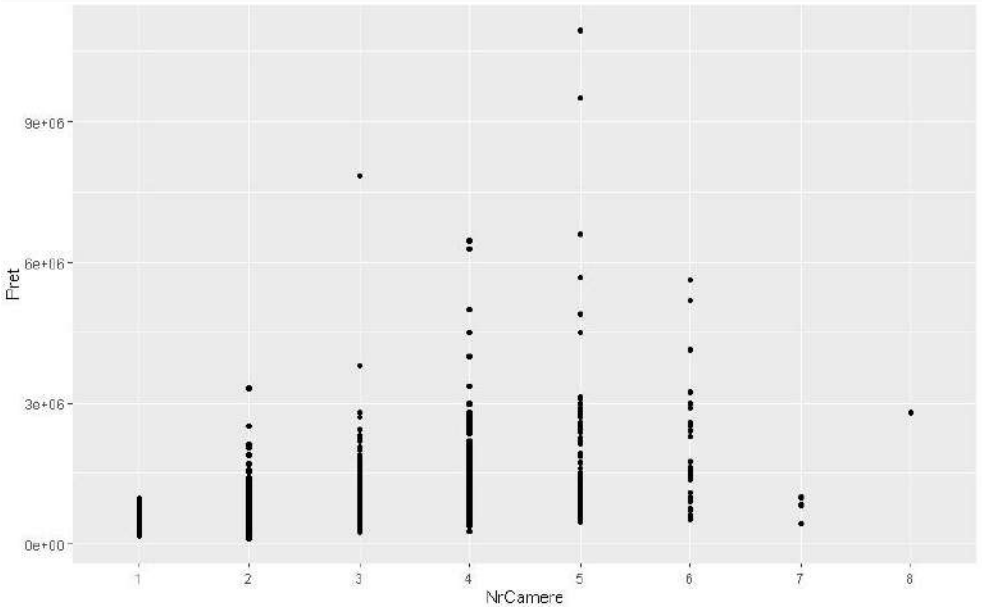
```
> str(PreturiCase)
tibble [4,199 x 5] (S3: tbl_df/tbl/data.frame)
 $ Oras     : Factor w/ 3 levels "Krakow","Poznan",..: 1 3 1 2 3 1 1 3 3 3 ...
 $ Etaj     : Factor w/ 11 levels "0","1","2","3",..: 3 4 3 3 2 3 1 11 4 2 ...
 $ Pret     : num [1:4199] 749000 240548 427000 1290000 996000 ...
 $ NrCamere : Factor w/ 8 levels "1","2","3","4",..: 3 1 2 5 5 1 4 6 4 1 ...
 $ Suprafata: num [1:4199] 74 24.4 37 166 105 ...
```

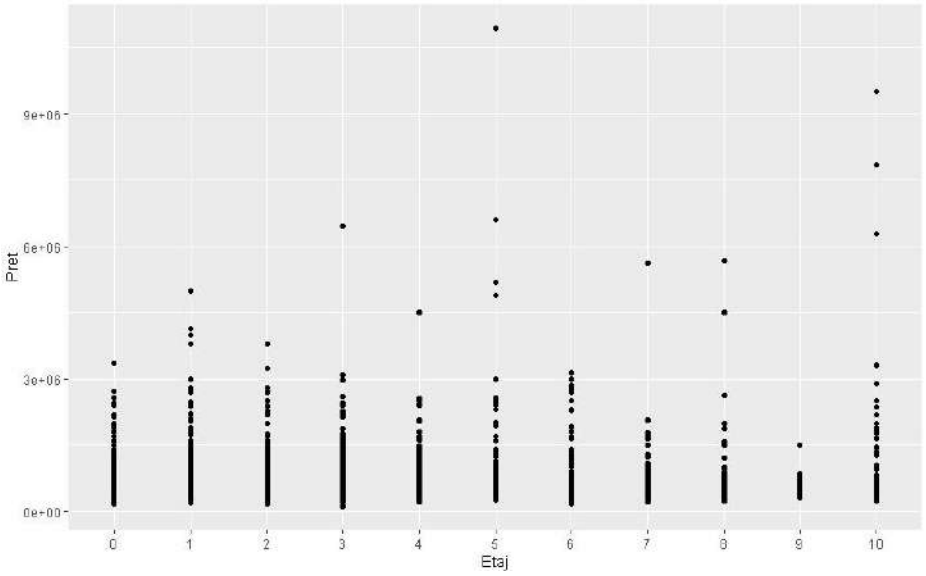
Fiindcă ne-am propus să analizăm fluctuația prețului în funcție de restul factorilor am continuat prin vizualizarea acestuia cu fiecare factor în parte.

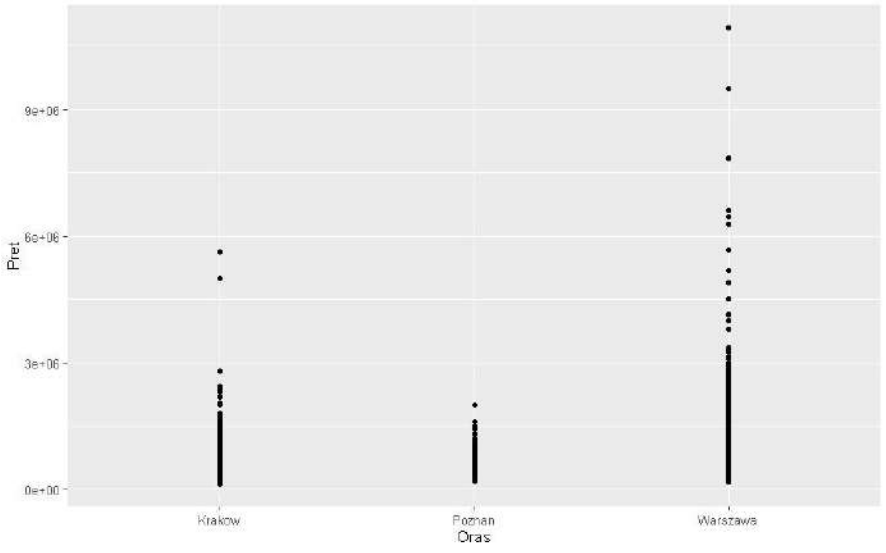


Din acest prim tabel putem observa că suprafața influențează într-o anumită măsură prețul unei case sau a unui apartament prin faptul că prețul este scăzut pentru o suprafață mică și mai ridicat pentru o suprafață mai mare.

Următoarele trei tabele nu sunt concludente deoarece nu putem observa dacă un preț scăzut are un număr mai mic de camere, este situat la un anumit etaj sau într-un anumit oraș.


Pret
9e+06
6e+06
3e+06
0e+00
1
2
3
4
5
6
7
8
NrCamere


Pret
9e+06
6e+06
3e+06
0e+00
0
1
2
3
4
5
6
7
8
9
10
Etaj


Pret
9e+06
6e+06
3e+06
0e+00
Krakow
Poznan
Warszawa
Oras

3. Rezultate și discuții

Arbori de decizie

Am ales să lucrăm cu metoda arborilor de decizie deoarece sunt mai ușor de înțeles de către persoanele din afara sferei informatice care prezintă adesea dificultăți în priceperea mecasismului de Machine Learning.Astfel, cei interesați de conținutul și concluziile acestei analize vor putea consulta notițele fără a fi intimidați foarte mult de metodele folosite, prin urmare fiindu-le facil să „digereze” informațiile aflate.

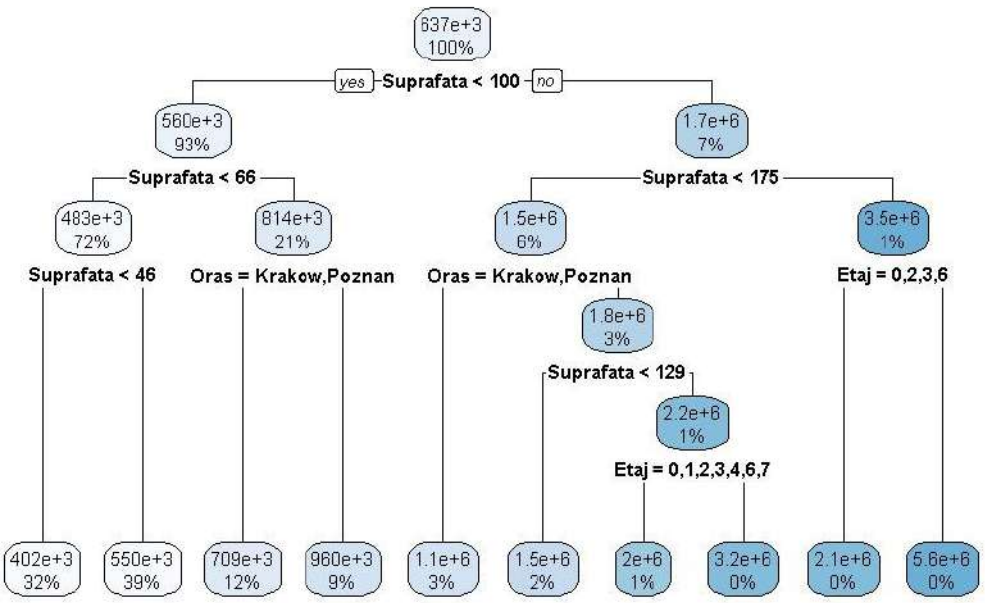
Componentele arborilor sunt: rădăcina, nodurile intermediare și frunzele. Nodurile intermediare împart datele după un anumit factor.

Pentru că există riscul de overfitting atunci când nu este utilizat corespunzător trebuie să înțelegem unele dintre cele mai importante proprietăți ale arborelui, mai exact adâncimea și minimul de instanțe de unde nu mai este posibilă diviziunea. Aceste proprietăți sunt esențiale în efectuarea unei predicții cu eroare minimală.

În primă fază am împărțit setul de date în două: un set de date pentru antrenamentul modelului, format din 70% din totalul datelor, respectiv un set de date pentru testarea acestuia, format din restul de 30%.

Deoarece valorile de antrenament și de testare din setul inițial sunt alese aleatoriu, pentru a nu fi o diferență mare între execuții am setat un seed de 1. Astfel rularea codului de mai multe ori va asigura o oarecare similaritate a rezultatelor.

Am continuat cu învățarea modelului printr-un minim de diviziuni și maxim de lungime cu setările prestabilite și am obținut următorul arbore:



În formă textuală, arborele de mai sus este reprezentat astfel:

```
n= 2939

node), split, n, deviance, yval
      * denotes terminal node

 1) root 2939 7.327647e+14  636822.1
   2) Suprafata< 99.9 2737 1.697426e+14  559655.9
     4) Suprafata< 65.525 2106 5.395589e+13  483363.8
       8) Suprafata< 45.985 949 1.391183e+13  401836.1 *
       9) Suprafata>=45.985 1157 2.856250e+13  550234.8 *
     5) Suprafata>=65.525 631 6.261721e+13  814285.2
      10) Oras=Krakow,Poznan 366 2.162267e+13  708869.4 *
      11) Oras=Warszawa 265 3.131007e+13  959878.4 *
   3) Suprafata>=99.9 202 3.258971e+14 1682386.0
     6) Suprafata< 175.05 182 1.139940e+14 1486216.0
      12) Oras=Krakow,Poznan 84 1.278196e+13 1102524.0 *
      13) Oras=Warszawa 98 7.824573e+13 1815096.0
        26) Suprafata< 129.15 60 1.517078e+13 1544337.0 *
        27) Suprafata>=129.15 38 5.173113e+13 2242610.0
          54) Etaj=0,1,2,3,4,6,7 31 2.430480e+13 2024812.0 *
          55) Etaj=5,8,10 7 1.944354e+13 3207143.0 *
     7) Suprafata>=175.05 20 1.411645e+14 3467530.0
      14) Etaj=0,2,3,6 12 7.784078e+12 2063000.0 *
      15) Etaj=1,5,10 8 7.419927e+13 5574325.0 *
```

Acest arbore subliniază faptul că suprafața este un factor important în determinarea prețului, astfel diviziunile se realizează preponderent după acesta după care alți factori relevanți sunt orașul și etajul.

Eroarea pe care am obținut-o în urma acestor calcule a fost de 387.418,2, o valoare destul de ridicată pe care am încercat să o corectăm prin următoarele calcule.

```
> pred <- predict(m1, newdata = test_case)
> RMSE(pred = pred, obs = test_case$Pret)
[1] 387418.2
```

Am încercat să găsim cea mai bună diviziune minimă și cea mai bună lungime maximă cu ajutorul următorului segment de cod:

```
hyper_grid <- expand.grid(
  minsplit = seq(5, 20, 1),
  maxdepth = seq(8, 15, 1)
)
head(hyper_grid)
models <- list()
for (i in 1:nrow(hyper_grid)) {
  minsplit <- hyper_grid$minsplit[i]
  maxdepth <- hyper_grid$maxdepth[i]
  models[[i]] <- rpart(
    formula = Pret ~. ,
    data = antrenament_case,
    method = "anova",
    control = list(minsplit = minsplit, maxdepth = maxdepth)
  )
}
```

Astfel am creat o hiper grilă formată din totalitatea combinațiilor dintre minimul de diviziuni și maximul de lungime din [5-100] respectiv [8-100]. În *for* s-a generat câte un arbore de decizie pentru fiecare valoare posibilă a minimului de diviziuni și a maximului de

lungime din grilă. Apoi ne-am folosit de o listă pentru a aduna în aceasta totalitatea modelelor generate în *for*.

Dorind să aflăm cp-ul și eroarea minimă a fiecărui arbore ne-am folosit de două funcții care furnizează cele mai mici valori pentru cp și eroare din arbori.

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

Grilei i-am lipit două coloane cu cp-ul cel mai mic găsit cât și cea mai mică eroare pentru minimul de diviziuni și maximul de lungime.

```
> mutated_grid %>%      #top 5 arbori cu eroarea cea mai mica
+   arrange(error) %>%
+   top_n(-5, wt=error)
  minsplit maxdepth   cp     error
1        9       13 0.01 0.4087261
2        5        8 0.01 0.4116743
3       19       10 0.01 0.4161760
4       12       10 0.01 0.4186702
5        6       10 0.01 0.4247033
```

Am creat apoi un model căruia i-am adăugat combinația cea mai bună:

```
optimal_tree <- rpart(
  formula = Pret ~ .,
  data = antrenament_case,
  method = "anova",
  control = list(minsplit = 9, maxdepth = 13, cp = 0.01)
)
```

După care am calculat predicția acestui model:

```
> pred <- predict(optimal_tree, newdata = test_case)
> RMSE(pred = pred, obs = test_case$Pret)
[1] 387418.2
```

Se pare că setările prestabilite au generat o predicție bună până la urmă, deoarece avem o valoare egală a arborelui anterior și arborele optim.

Regresia liniară

Pentru a concluziona, am realizat o serie de regresii - liniare și multiple. Se constată că există o relație liniară între suprafață și preț și între numărul de camere și preț.

În ceea ce privește relația suprafață-preț, am creat o primă regresie cu următoarele rezultate:

```
> Pret_Suprafata <- lm(data = PreturiCase , Pret ~ Suprafata)
> summary(Pret_Suprafata)

Call:
lm(formula = Pret ~ Suprafata, data = PreturiCase)

Residuals:
     Min       1Q   Median       3Q      Max
-2237603  -136665   -16037   102056  7708593

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -166355.1    12608.8  -13.19   <2e-16 ***
Suprafata     13685.8      192.4   71.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 347500 on 4197 degrees of freedom
Multiple R-squared:  0.5465,    Adjusted R-squared:  0.5464
F-statistic:  5058 on 1 and 4197 DF,  p-value: < 2.2e-16
```

Din aceast rezultat putem observa, conform estimării, că dacă suprafața ar crește cu un metru pătrat, prețul unui apartament ar putea crește cu 13685.8 euro.

Standard error (SE) pentru suprafață este 192.4. Putem spune că estimarea acestui parametru diferă în medie cu 192.4 față de valoarea reală. Pentru calcularea intervalelor de încredere se foloseste această măsură.

T-statistic în acest caz este de 71.12. Acesta este o valoare absolută care ne arată numărul de deviații standard față de zero.

P-value este o probabilitate. Aceasta poate să spună dacă asocierea dintre variabila dependentă și parametru este datorată șansei sau nu. În acest caz asocierea dintre preț și suprafață există , deoarece valoarea lui P este foarte mică.

Residual Standard Error (RSE), arată distanța de la punctele din norul de puncte la dreapta de liniaritate. Acest RSE are valoarea 347500, o valoare destul de mare. Cu cât acest parametru este mai mic cu atat putem observa lipsa de potrivire a modelului.

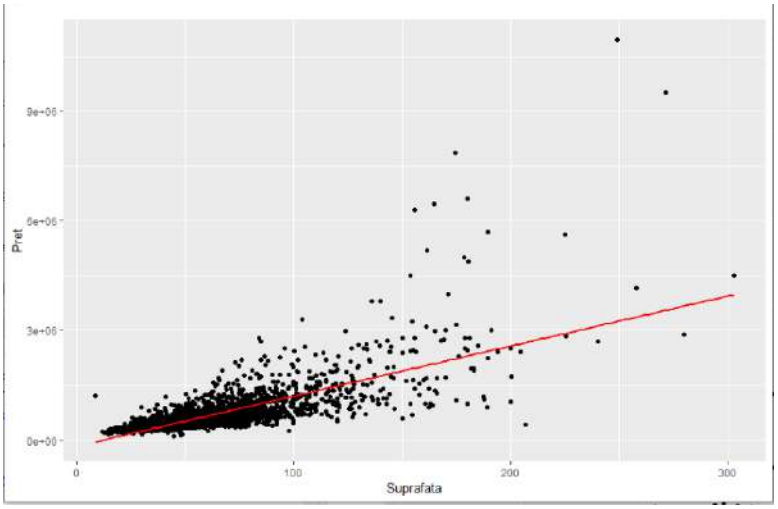
Multiple R-squared (R^2) indică proporția din variabila dependentă (Y) explicată de variabila independentă(X), adică gradul de corelație. Acest parametru ar fii indicat să fie cât mai mare posibil. În acest caz este 0.546, adică 54% din prețul unei case este susținut de suprafața acesteia.

F-statistic este similar cu parametrul T-statistic, dar răspunde pentru modelul întreg.

Acesta este intervalul de încredere. Putem observa că intervalul pentru suprafață este destul de mare.

```
> confint(Pret_Suprafata) #ne da intervalele de incredere
                  2.5 %     97.5 %
(Intercept) -191075.1 -141635.12
Suprafata     13308.5   14063.08
>
```

Am realizat următoarele predicții și următorul grafic:



Următoarea regresie este pentru Preț si numărul de camere ale imobilului.

```
Call:
lm(formula = Pret ~ NrCamere, data = PreturiCase)

Residuals:
     Min       1Q   Median       3Q      Max
-1423164  -172676   -66087    74869  9337100

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   370211      20487  18.071  < 2e-16 ***
NrCamere2     124840      23374   5.341 9.73e-08 ***
NrCamere3     306861      23527  13.043  < 2e-16 ***
NrCamere4     593876      28221  21.044  < 2e-16 ***
NrCamere5    1242689      46325  26.825  < 2e-16 ***
NrCamere6    1572953      86660  18.151  < 2e-16 ***
NrCamere7     382456     258060   1.482    0.138
NrCamere8    2419789     446034   5.425 6.12e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 445600 on 4191 degrees of freedom
Multiple R-squared:  0.2556,    Adjusted R-squared:  0.2544
F-statistic: 205.6 on 7 and 4191 DF,  p-value: < 2.2e-16
```

```
> confint(Pret_NrCamere) #ne da intervalele de incredere
                2.5 %    97.5 %
(Intercept)  330045.8  410376.6
NrCamere2     79014.9  170664.4
NrCamere3    260736.5  352985.2
NrCamere4    538547.7  649204.1
NrCamere5   1151867.0 1333511.2
NrCamere6   1403053.4 1742852.0
NrCamere7   -123479.4  888390.4
NrCamere8   1545326.8 3294250.8
>
```

Din aceste rezultate putem să observăm că aproape toate valorile sunt destul de mari față de modelul suprafeței. Indicatorul RSE este cam cel mai important punct de comparație , acesta fiind mai mare (445600>347500). P-value are și aici o valoare mică, acest lucru înseamnă că relația dintre aceste 2 proprietăți nu este datorată șansei. Până și intervalele de încredere sunt mai mari.

Regresia Preț și Etaj este un model mai puțin reușit decât cel cu prețul și numărul camerelor.

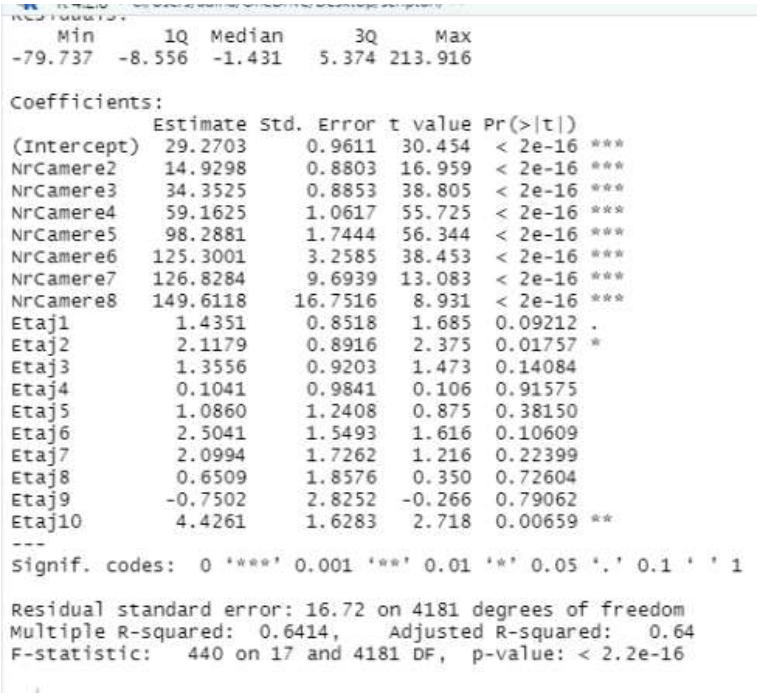
```
Call:
lm(formula = Pret ~ Etaj, data = PreturiCase)

Residuals:
     Min       1Q   Median       3Q      Max
 -669832  -231305  -114496    66730 10182130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   565697      19539  28.953  < 2e-16 ***
Etaj1          66573      25979   2.563 0.010426 *
Etaj2          52223      27230   1.918 0.055196 .
Etaj3          68793      28102   2.448 0.014407 *
Etaj4          63799      30045   2.123 0.033774 *
Etaj5         202173      37897   5.335 1.01e-07 ***
Etaj6         225396      47291   4.766 1.94e-06 ***
Etaj7         177090      52734   3.358 0.000792 ***
Etaj8         165896      56782   2.922 0.003501 **
Etaj9          21698      86310   0.251 0.801525
Etaj10        344135      49737   6.919 5.24e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 511400 on 4188 degrees of freedom
Multiple R-squared:  0.02018,   Adjusted R-squared:  0.01784
F-statistic: 8.624 on 10 and 4188 DF,  p-value: 4.263e-14

> |
```

Regresia Preț și oraș:

```
Call:
lm(formula = Pret ~ Oras, data = PreturiCase)

Residuals:
     Min       1Q   Median       3Q      Max
 -613118  -223118  -103417    68178 10167882

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    582046      12138  47.955  < 2e-16 ***
OrasPoznan    -120107      22152  -5.422 6.23e-08 ***
OrasWarszawa   200072      17018  11.756  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 501000 on 4196 degrees of freedom
Multiple R-squared:  0.05763,   Adjusted R-squared:  0.05718
F-statistic: 128.3 on 2 and 4196 DF,  p-value: < 2.2e-16

> |
```

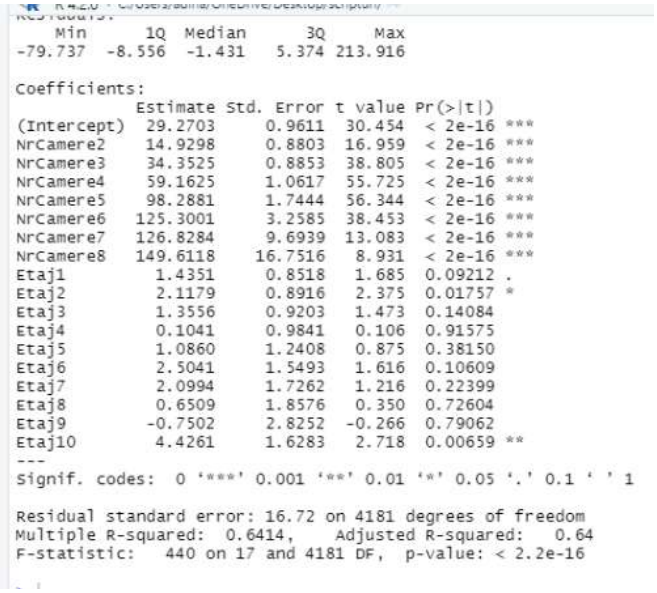
Mai departe urmează regresiile liniare care au mai mulți predictori care ar putea să ne dea o concluzie asupra relației între o variabilă dependentă și un predictor. Spre deosebire de regresia liniara simplă care ne poate da o concluzie în privința relației între o variabilă dependentă și un predictor, în condițiile în care ignorăm ceilalți factori.

Regresia multiplă între suprafață și numărul de camere dintr-o casă:



```
    Min      1Q  Median      3Q     Max
-79.737  -8.556  -1.431   5.374 213.916

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.2703     0.9611  30.454  < 2e-16 ***
NrCamere2    14.9298     0.8803  16.959  < 2e-16 ***
NrCamere3    34.3525     0.8853  38.805  < 2e-16 ***
NrCamere4    59.1625     1.0617  55.725  < 2e-16 ***
NrCamere5    98.2881     1.7444  56.344  < 2e-16 ***
NrCamere6   125.3001     3.2585  38.453  < 2e-16 ***
NrCamere7   126.8284     9.6939  13.083  < 2e-16 ***
NrCamere8   149.6118    16.7516   8.931  < 2e-16 ***
Etaj1         1.4351     0.8518   1.685  0.09212 .
Etaj2         2.1179     0.8916   2.375  0.01757 *
Etaj3         1.3556     0.9203   1.473  0.14084
Etaj4         0.1041     0.9841   0.106  0.91575
Etaj5         1.0860     1.2408   0.875  0.38150
Etaj6         2.5041     1.5493   1.616  0.10609
Etaj7         2.0994     1.7262   1.216  0.22399
Etaj8         0.6509     1.8576   0.350  0.72604
Etaj9        -0.7502     2.8252  -0.266  0.79062
Etaj10        4.4261     1.6283   2.718  0.00659 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.72 on 4181 degrees of freedom
Multiple R-squared:  0.6414,    Adjusted R-squared:   0.64
F-statistic:   440 on 17 and 4181 DF,  p-value: < 2.2e-16
```

Regresia multiplă cu suprafața , etajul și numărul de camere



```
    Min      1Q  Median      3Q     Max
-79.737  -8.556  -1.431   5.374 213.916

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.2703     0.9611  30.454  < 2e-16 ***
NrCamere2    14.9298     0.8803  16.959  < 2e-16 ***
NrCamere3    34.3525     0.8853  38.805  < 2e-16 ***
NrCamere4    59.1625     1.0617  55.725  < 2e-16 ***
NrCamere5    98.2881     1.7444  56.344  < 2e-16 ***
NrCamere6   125.3001     3.2585  38.453  < 2e-16 ***
NrCamere7   126.8284     9.6939  13.083  < 2e-16 ***
NrCamere8   149.6118    16.7516   8.931  < 2e-16 ***
Etaj1         1.4351     0.8518   1.685  0.09212 .
Etaj2         2.1179     0.8916   2.375  0.01757 *
Etaj3         1.3556     0.9203   1.473  0.14084
Etaj4         0.1041     0.9841   0.106  0.91575
Etaj5         1.0860     1.2408   0.875  0.38150
Etaj6         2.5041     1.5493   1.616  0.10609
Etaj7         2.0994     1.7262   1.216  0.22399
Etaj8         0.6509     1.8576   0.350  0.72604
Etaj9        -0.7502     2.8252  -0.266  0.79062
Etaj10        4.4261     1.6283   2.718  0.00659 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.72 on 4181 degrees of freedom
Multiple R-squared:  0.6414,    Adjusted R-squared:   0.64
F-statistic:   440 on 17 and 4181 DF,  p-value: < 2.2e-16
```

Regresia multiplă cu suprafața, numărul de camere, etajul și orașul

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.4593     1.0041  27.347  < 2e-16 ***
NrCamere2     14.9431     0.8764  17.050  < 2e-16 ***
NrCamere3     34.3991     0.8810  39.045  < 2e-16 ***
NrCamere4     59.1155     1.0570  55.929  < 2e-16 ***
NrCamere5     98.0825     1.7363  56.489  < 2e-16 ***
NrCamere6    124.3819     3.2454  38.326  < 2e-16 ***
NrCamere7    127.4696     9.6466  13.214  < 2e-16 ***
NrCamere8    151.3541    16.6703   9.079  < 2e-16 ***
Etaj1          1.5028     0.8477   1.773   0.0763 .
Etaj2          2.1866     0.8876   2.464   0.0138 *
Etaj3          1.4821     0.9162   1.618   0.1058
Etaj4          0.1670     0.9793   0.171   0.8646
Etaj5          0.7354     1.2358   0.595   0.5518
Etaj6          1.8079     1.5454   1.170   0.2421
Etaj7          1.6960     1.7212   0.985   0.3245
Etaj8          0.0903     1.8522   0.049   0.9611
Etaj9         -1.5723     2.8143  -0.559   0.5764
Etaj10         3.4131     1.6281   2.096   0.0361 *
OrasPoznan     1.5208     0.7382   2.060   0.0395 *
OrasWarszawa   3.7913     0.5729   6.618 4.11e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.64 on 4179 degrees of freedom
Multiple R-squared:  0.6452,    Adjusted R-squared:  0.6436
F-statistic:   400 on 19 and 4179 DF,  p-value: < 2.2e-16

> |
```

Ultima regresie este cea în care am folosit toți factorii de interes:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -210495.2    20788.0 -10.126  < 2e-16 ***
Suprafata      17242.8      294.9  58.461  < 2e-16 ***
Etaj1            564.2    16168.1   0.035  0.97217
Etaj2           2524.1    16935.6   0.149  0.88153
Etaj3           9779.5    17474.2   0.560  0.57574
Etaj4          39797.0    18671.4   2.131  0.03311 *
Etaj5         125653.2    23564.3   5.332 1.02e-07 ***
Etaj6          90863.2    29470.7   3.083  0.00206 **
Etaj7          90004.8    32822.5   2.742  0.00613 **
Etaj8          79348.0    35315.7   2.247  0.02470 *
Etaj9           3920.0    53662.2   0.073  0.94177
Etaj10        202160.6    31059.6   6.509 8.47e-11 ***
OrasPoznan   -147661.2    14082.7 -10.485  < 2e-16 ***
OrasWarszawa  102509.5    10980.4   9.336  < 2e-16 ***
NrCamere2    -119326.3    17281.8  -6.905 5.79e-12 ***
NrCamere3    -273159.7    19624.4 -13.919  < 2e-16 ***
NrCamere4    -414574.6    26649.0 -15.557  < 2e-16 ***
NrCamere5    -473892.9    43964.6 -10.779  < 2e-16 ***
NrCamere6    -616243.3    71936.8  -8.566  < 2e-16 ***
NrCamere7   -1823729.4   187734.3  -9.714  < 2e-16 ***
NrCamere8    -122978.2   320970.9  -0.383  0.70163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 317300 on 4178 degrees of freedom
Multiple R-squared:  0.6237,    Adjusted R-squared:  0.6219
F-statistic: 346.3 on 20 and 4178 DF,  p-value: < 2.2e-16

> |
```

Rezultate

Dacă luăm în considerare toți parametrii (R^2 , P-value, F-statistic și RSE) și comparăm toate modelele de regresie realizate, putem spune că cel mai concret model este cel în care examinăm variabilele independente , iar suprafața și numărul de camere sunt cele mai concludente dintre aceste variabile.

Dacă comparăm rezultatele arborilor cu cele ale regresiilor, putem spune că metoda optimă de predicție pentru setul de date pe care l-am ales este regresia liniara multipla deoarece eroarea medie standard generată este mult mai mică decât eroarea pe care am reușit să o prezicem în cadrul arborilor. (317300>387418)

Concluzii

- Prețul unei locuințe din Polonia este influențat de orașul în care se află, de numărul de camere pe care îl are, etajul la care este și suprafața pe care o are;
- Analizând regresiile liniare individuale, constatăm că cele doua variabile independente (suprafața și numărul de camere) sunt cele mai importante. În urma arborilor de decizie, se poate observa că în funcție de atributul suprafață se face prima diferențiere, de unde rezultă că este cel mai semnificativ iar numărul de camere nu este la fel de semnificativ ca în cazul regresiei;

- Am avut rezultate diferite folosindu-ne de arborii de decizie și regresia liniară chiar dacă setul de date folosit a fost identic.