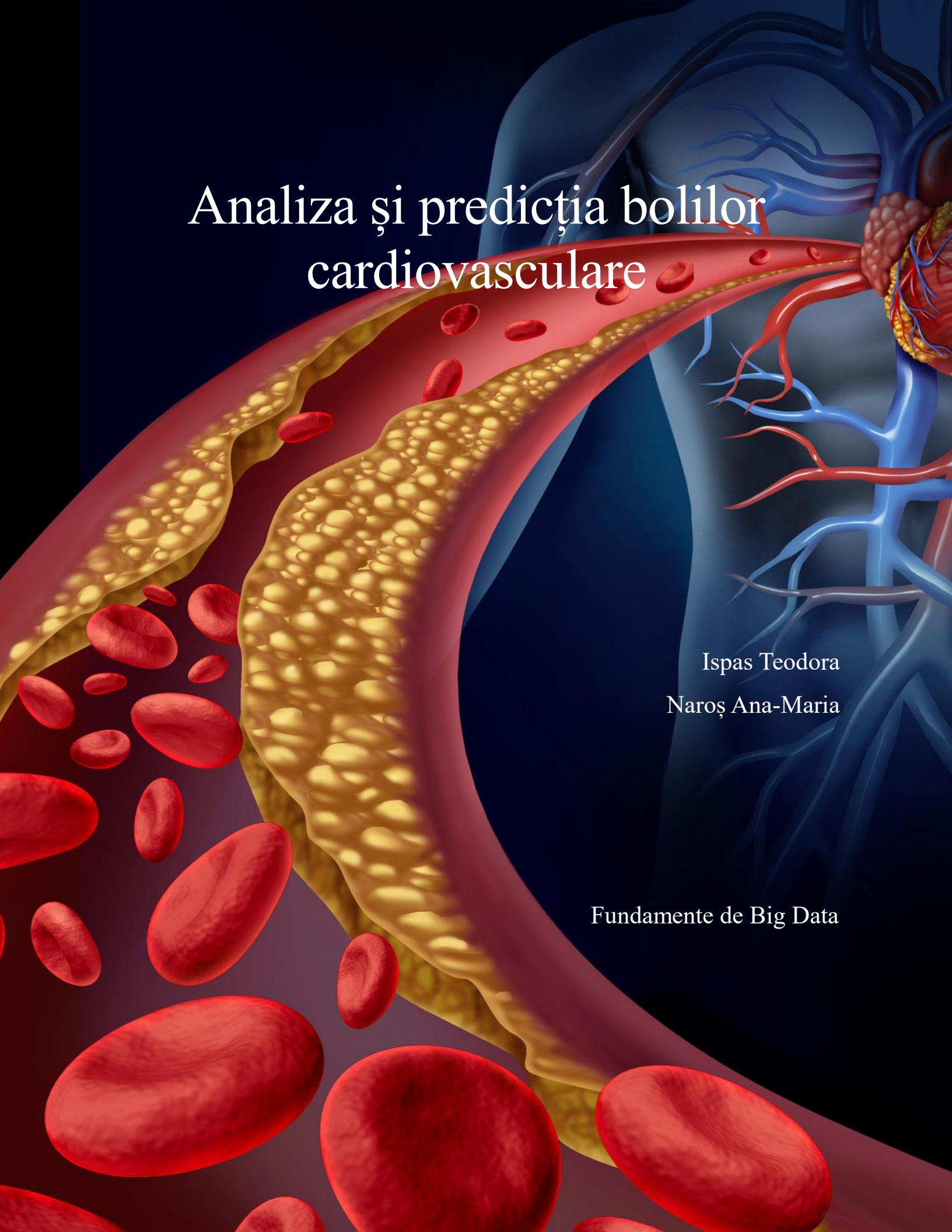
Analiza și predicția bolilor cardiovasculare



Ispas Teodora

Naroș Ana-Maria

Fundamente de Big Data

Cuprins

1. Introducere

Bolile cardiovasculare (CVD) sunt un grup de tulburări ale inimii și ale vaselor de sânge și includ:

- *Boală coronariană* - boală a vaselor de sânge care alimentează mușchiul inimii:

-angina stabilă, angina instabilă, infarctul miocardic și decesul cardiac subit

- *Boală cerebrovasculară* - boală a vaselor de sânge care alimentează creierul

- accidentele vasculare ischemice și atacurile ischemice tranzitorii, hemoragiile intracerebrale, hemoragiile subarhnoidiene

- *Boală arterială periferică*, *Boală cardiacă reumatică*, *Boli cardiace congenitale*, *Tromboză venoasă profundă și embolie pulmonară*

Potrivit Organizației Mondiale a Sănătății, CVD sunt prima cauză de deces la nivel global, adică mai multe persoane mor anual din cauza CVD decât din orice altă cauză. Se estimează că 17,9 milioane de persoane au murit din cauza bolilor cardiovasculare în 2016, reprezentând 31% din totalul deceselor la nivel mondial. Dintre aceste decese, 85% se datorează infarctului și accidentului vascular cerebral.[1]

Prin analiza acestor seturi de date se dorește găsirea principalilor factori care cresc șansele de a dezvolta o astfel de afecțiune. Datele din setul ales au fost colectate la momentul examinării medicale, printre care se numără: vârsta, sexul, nivelul de colesterol, glicemia și altele. Tot în urma analizei ar trebui să reușim să răspundem la întrebări precum: Cât de exact putem prezice dacă o persoană suferă sau nu de o boală cardiovasculară dacă știm tensiunea sistolică? Dar știind mai multe informații despre un pacient? O altă întrebare se referă sexul și vârsta cardiacilor. Influențează sexul și vârsta apariția unei boli?

[1] Potrivit https://www.kaggle.com/rahulgulia/data-science-and-cardiovascular-diseases-cvds

Într-un studiu făcut de către *American Heart Association (AHA)* și *National Institutes of Health* s-au analizat principalii factori de risc printre care se numără: fumatul, lipsa activității fizice, nutriție, obezitate, colesterol ș.a.

Hipertensiunea arteriala este unul dintre cei mai importanți factori de risc în afecțiunile cardiovasculare. Aproximativ 54% dintre infarcturi și 47% dintre atacurile cerebrale sunt atribuite unei tensiuni arteriale crescute. Ea este o boală comună, iar riscul ei crește odată cu vârsta. Estimările spun că afectează 65% dintre persoanele de peste 60 ani.[2]

Un alt studiu realizat de BMJ Group Health arată că deși bolile cardiovasculare sunt o cauză principală a decesurilor la nivel mondial, există o diferență importantă între bărbați și femei. Bărbații dezvoltă această boală mai repede și au un risc ridicat de boli coronariene, ex. infarct miocardic, pe când femeile sunt mai predispuse la atacuri vasculare cerebrale care se declanșează la o vârstă mai înaintată (conform *Netherlands Heart Journal*[3], femeile dezvoltă afecțiuni cu 7 ani mai târziu decât bărbații). Se consideră că această întârziere se datorează perioadei de viață fertile a femeilor, observându-se înainte de menopauză o rată scăzută a bolilor cardiovasculare și majoritatea fiind cauzate de fumat. Mortalitatea cauzată atât de infarctul miocardic cât și de atacuri cerebrale rămâne totuși mai ridicată la bărbați decât la femei până la o vârstă înaintată.

Obiectivul urmărit în cadrul proiectului este acela de a analiza factorii principali care duc la apariția acestor boli, de a prezice dacă o persoană suferă sau nu de o astfel de afecțiune bazându-ne pe câțiva dintre acei factori esențiali și de a vedea în ce măsură există o corelație între datele din setul nostru.

2. Setul de date

Setul de date ales conține înregistrări furnizate în urma anumitor examinări medicale, cât și informații furnizate de către pacienți, conținând astfel 70000 de instanțe și 12 atribute, printre care și clasa țintă, *cardio*. Atributele setului de date pot fi regăsite în cadrul figurii de mai jos.

```
> names(cardio_train)
 [1] "id"          "age"         "gender"      "height"      "weight"      "ap_hi"       "ap_lo"
 [8] "cholesterol" "gluc"        "smoke"       "alco"        "active"      "cardio"
>
```

[2] Conform https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059018/
[3] Netherlands Heart Journal https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3018605/

Fiecare coloană reprezintă un factor important în analiza bolilor cardiovasculare în rândul pacienților. În acest sens, ***age*** reprezintă vârsta pacienților care au participat la examinarea medicală, reprezentată în zile, ***gender*** reprezintă sexul participanților, care este o variabilă nominală, unde 1-Feminin, 2-Masculin, ***height*** furnizează înălțimea participanților, exprimată în cm, în timp ce ***weight*** greutatea acestora, în kg. Trecând mai departe, ***ap_hi*** reprezintă datele corespunzătoare nivelului de tensiune sistolică, sau tensiunea mare, acea presiune exercitată de sânge împotriva pereților arterelor atunci când inima bate , iar în ceea ce privește ***ap_lo***, această coloană cuprinde valorile înregistrate pentru tensiunea diastolică, sau tensiunea mică, adică presiunea exercitată de sânge pe pereții arterelor în timp ce inima se odihnește între bătăi, ambele cuprinzând valori numerice.

Cholesterol și ***gluc*** furnizează informații despre nivelul colesterolului, repectiv al glucozei printre pacienții analizați, fiind împărțite în trei categorii: *Normal*, *Above normal*, respectiv *Well above normal*. Variabila nominală ***smoke*** reprezintă date despre stagiul de fumător/nefumător al unei persoane, ***alco*** prezintă consumul de alcool la nivelul fiecărui individ, iar ***active*** este nivelul de intensitate fizică. Nu în cele din urmă, clasa țintă este reprezentată de coloana ***cardio*** care conține date nominale despre stadiul bolilor pacienților, astfel încât 0 seminifică acele persoane care nu au avut probleme cu boli cardiovasculare până în momentul examinării medicale, iar 1 seminifică acele persoane care suferă de afecțiuni cardiovasculare.

2.1 Curățarea datelor

Datorită unui număr mare observat de *outliers* în setul de date au avut loc o serie de pași de curățare pentru a aduce setul de date într-o formă cât mai adecvată pentru a învăța modelele noastre. Astfel, am considerat coloana *id* irelevantă, fapt pentru care aceasta a fost eliminată complet din setul de date cu ajutorul comenzii

```
> cardio_train = cardio_train %>% select(-id) %>% select(age, everything())
> names(cardio_train)
 [1] "age"         "gender"      "height"      "weight"      "ap_hi"       "ap_lo"       "cholesterol"
 [8] "gluc"        "smoke"       "alco"        "active"      "cardio"
>
```

O altă variabilă importantă, reprezentată de coloana *age*, conține valori ale vârstei pacienților măsurate în zile, însă acest lucru nu este foarte reprezentativ, iar pentru a schimba modul de afișare al valorilor am transformat valorile în ani cu ajutorul diviziunii simple la numărul de zile de pe parcursul unui an:

```
> head(cardio_train$age, n=10)
 [1] 18393 20228 18857 17623 17474 21914 22113 22584 17668 19834
> cardio_train$age <- as.numeric(cardio_train$age) %/% 365.25
> head(cardio_train$age, n=10)
 [1] 50 55 51 48 47 59 60 61 48 54
> 
```

Conform tabelului corespunzător distribuției normale a tensiunii sistolice și diastolice prezent în imaginea de mai jos, pentru coloanele reprezentate de valorile acestora s-au aplicat o serie de filtre astfel încât să păstrăm doar înregistrările relevante, într-un interval relevant. Astfel, în acest sens s-au creat subseturi de date care au fost alocate setului de date inițial pentru tensiunea sistolică cu valori normale între 90-240 mmHg, pentru tensiunea diastolică cu valori între 50-190 mmHg.

Observație: s-au constat multe înregistrări cu valori ale tensiunii diastolice de 1000 și 1100 care s-au transformat în 100, respectiv 110, considerând că a fost vorba despre o eroare.

Pentru înălțime considerăm valori cuprinse între 120-230cm și pentru greutate un minim de 40kg și un maxim de 200kg.

```
> #transform 1000 in 100 ap_lo
> #transform 1100 in 110 ap_lo
> cardio_train$ap_lo[cardio_train$ap_lo == 1000] <- 100
> cardio_train$ap_lo[cardio_train$ap_lo == 1100] <- 110
> #range
> cardio_train <-  subset(cardio_train, ap_hi
+                         %in% (90:240) & ap_lo
+                         %in% (50:190) & height
+                         %in% (120:230) & weight
+                         %in% (40:200))
> 
```

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

Figura 1. *Distribuția valorilor normale ale tensiunii*

Mai departe, datele factoriale au fost interpretate corespunzător, coloana înregistrărilor genului participanților a primit valoarea *Woman* pentru acele înregistrări reprezentate de valoarea 1, respectiv *Man* pentru celelalte.

```
> #M/F
> cardio_train <- cardio_train %>%
+   mutate(gender = ifelse(gender == 1, "Woman", "Man"))
> |
```

Conform setului de date, coloanele colesterol și glucoză cuprind înregistrări factoriale pentru 1-*Normal*, 2-*Above normal*, respectiv 3-*Well above normal*, iar acest mod de vizualizare a fost adoptat și în setul nostru de date cu ajutorul funcției *mutate():*

```
> #cholesterol + glu
> cardio_train <- cardio_train %>%
+   mutate(cholesterol = ifelse(cholesterol >=2, "Above normal", "Normal")) %>%
+   mutate(gluc = ifelse(gluc >=2, "Above normal","Normal"))
```

Tot prin această funcției au fost grupate două clase: *Above normal* și *Well above normal* sub denumirea de *Above normal* care reprezintă valorile peste limita normală atât a colesterolului cât și a glucozei.

Nu în cele din urmă, toate înregistrările reprezentate de factorii 0-1, și anume dacă pacienții sunt fumători sau nu, dacă aceștia depun efort fizic sau nu, dacă sunt consumători de alcool sau dacă suferă de boli cardiovasculare, au primit valorile predefinite *Yes* și *No*.

```
>
> #smoke, alco, active, cardio "Yes"/"No"
> cardio_train <- cardio_train %>%
+   mutate(smoke = ifelse(smoke == 0, "No", "Yes")) %>%
+   mutate(active = ifelse(active == 0, "No", "Yes")) %>%
+   mutate(alco = ifelse(alco == 0, "No", "Yes")) %>%
+   mutate(cardio = ifelse(cardio == 0, "No", "Yes"))
> |
```

Pentru a dispune de o serie de factori mai reprezentativi, au fost calculate două coloane extra pe baza datelor deținute: indicele corporal ***BMI***, cu ajutorul formulei:

```
> #bmi
> cardio_train$bmi <- with(cardio_train, weight/((height/100)^2))
> head(cardio_train$bmi, n=10)
 [1] 21.96712 34.92768 23.50781 28.71048 23.01118 29.38468 37.72973 29.98359 28.44095 25.28257
> |
```

și pulsul pacienților:

```
> #pulse
> cardio_train$pulse <- with(cardio_train,  ap_hi-ap_lo)
> head(cardio_train$pulse, n=10)
 [1] 30 50 60 50 40 40 50 40 40 50
> |
```

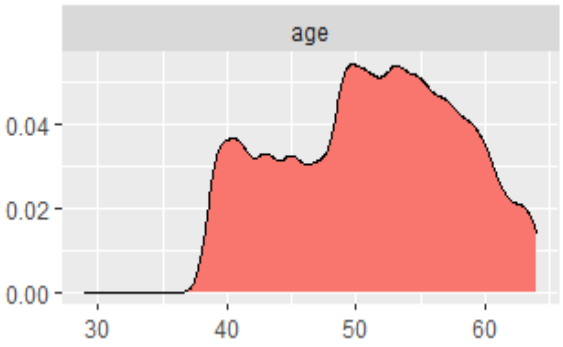
De asemenea, au fost eliminate valorile înregistrate pentru tensiunea diastolică ce depășeau valorile tensiunii sistolice deoarece acest lucru este imposibil d.p.d.v medical. Acest lucru s-a realizat prin selectarea doar a acelor valori care respectă condiția ca tensiunea diastolică să fie cel mult egală cu tensiunea sistolică.

```
> #delete data where ap_lo is higher than ap_hi
> cardio_train<-cardio_train[!(cardio_train$ap_lo>=cardio_train$ap_hi),]
> |
```

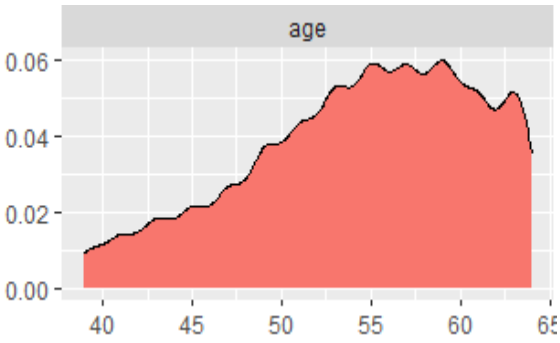
2.2 Analiza datelor

Un prim pas în analiza datelor este dat de reprezentarea grafică a variabilelor numerice în funcție de cele două clase Yes/No. Astfel s-a realizat distribuția acestor variabile în urma cărora s-a constatat:

-în distribuția clasei *age* în cazul clasei majoritare(persoane sănătoase) se observă un număr mai mare de înregistrări în intervalul 40-50 ani. De aici putem concluziona că odată cu înaintarea în vârstă (peste 50 de ani) crește riscul de apariție a unei boli cardiovasculare.

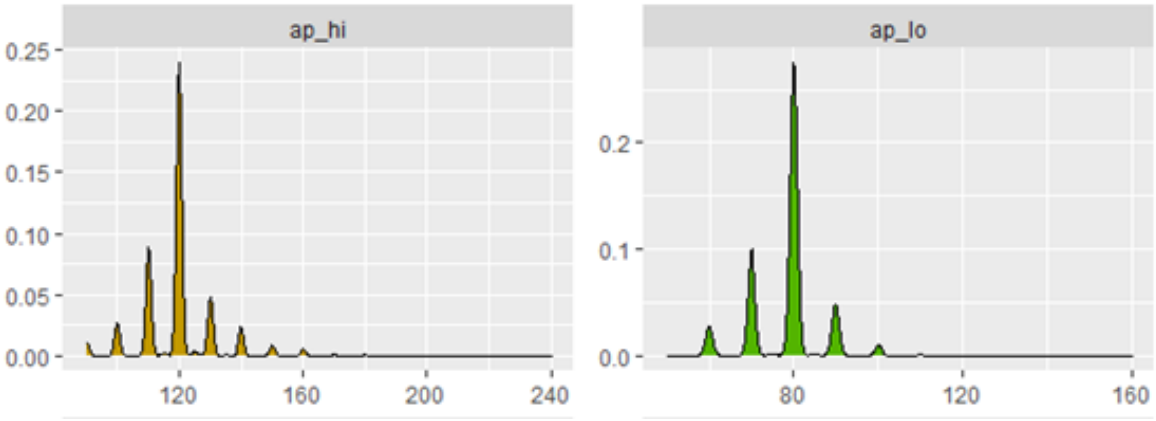


Distribuția vârstei persoanelor sănătoase

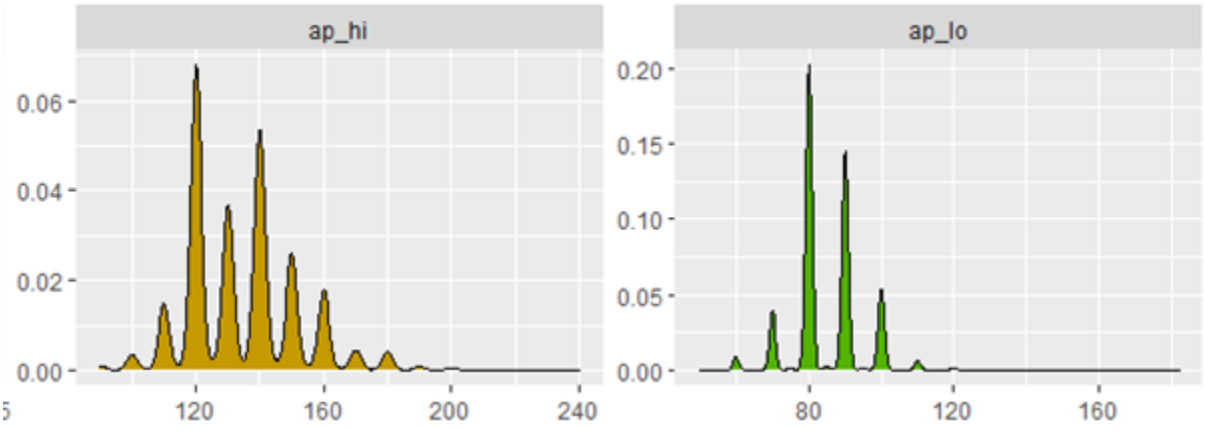


Distribuția vârstei persoanelor cardiace

-in cazul *ap_hi* si *ap_lo*, care corespund tensiunii sistolice și diastolice observăm o diferență majoră între distribuțiile fiecărei clase, a persoanelor sănătoase, respectiv a cardiacilor:



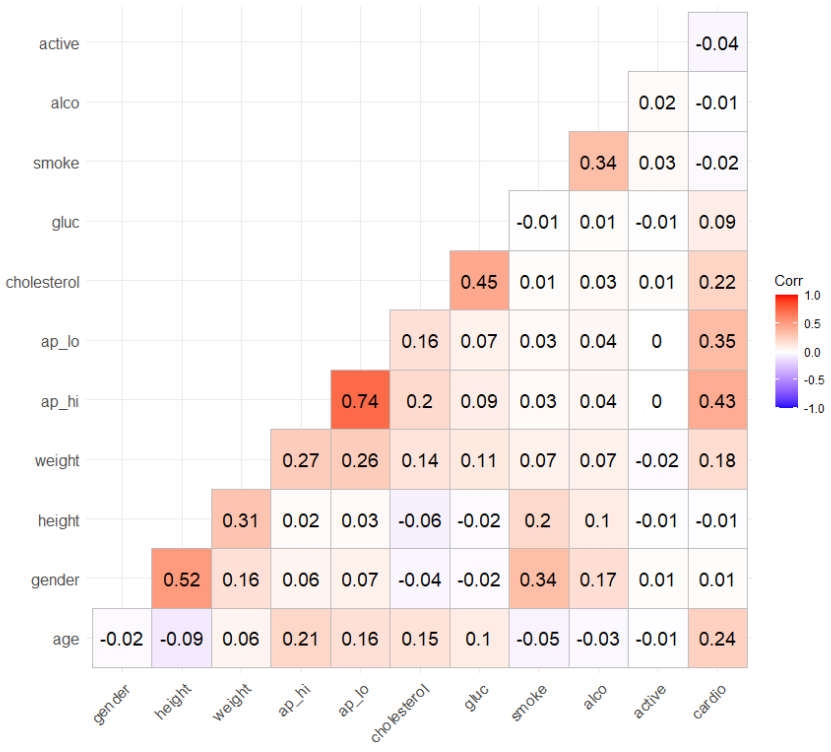
Distribuția tensiunii persoanelor sănătoase



Distribuția tensiunii persoanelor cardiace

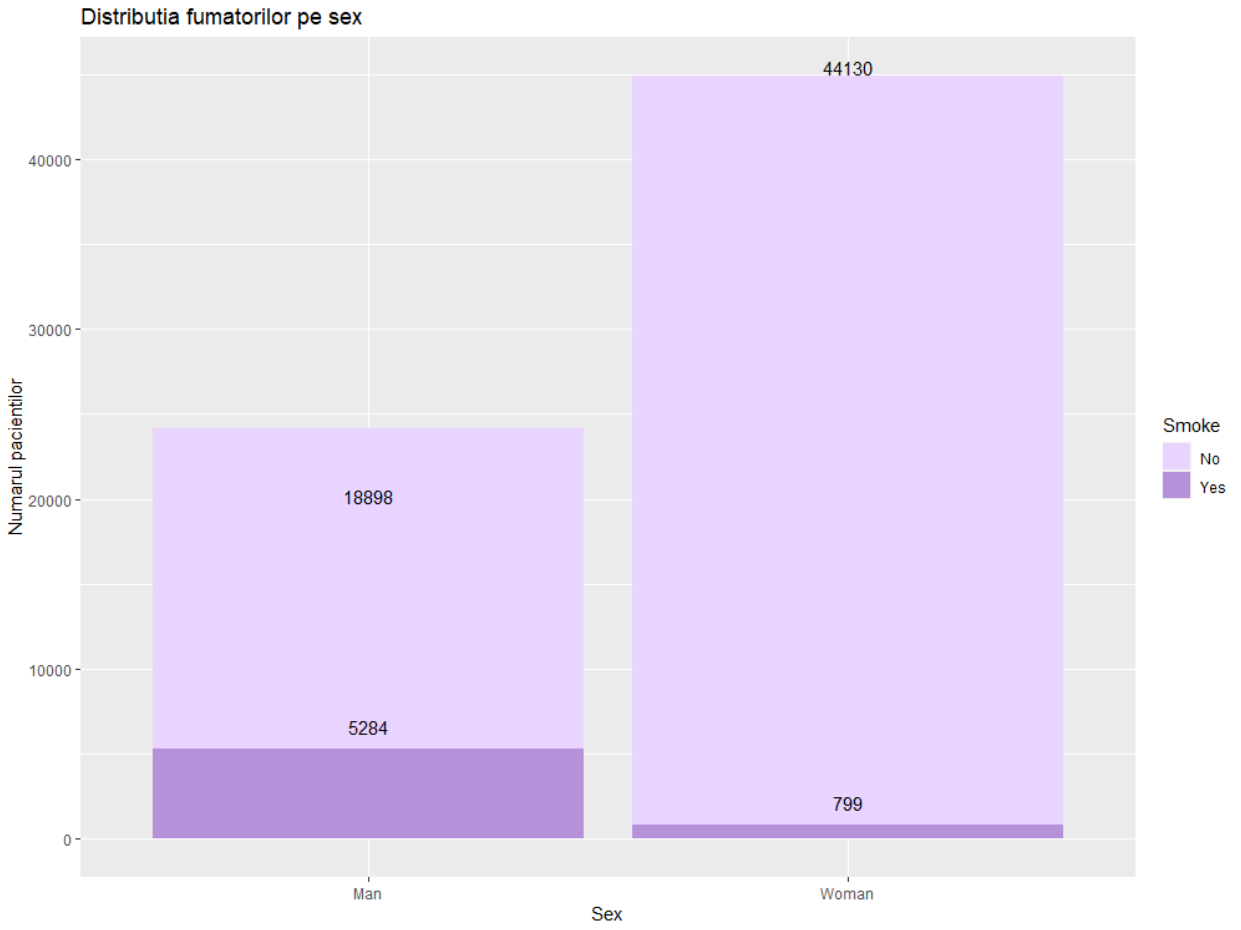
Astfel, se poate observa un nivel crescut atât al tensiunii sistolice în rândul cardiacilor, aceasta având mult mai multe valori înregistrate în intervalul 140-160mmHg decât în rândul persoanelor sănătoase unde majoritatea se află situați la valoarea normală de 120mmHg, cât și în rândul tensiunii diastolice, unde aproximatix 20% din persoanele cardiace au valori egale sau chiar mai mari decât limita maximă de 90mmHg.

Cu ajutorul matricei de corelație am reușit să vizualizăm coeficientul de asociere a atributelor din setul nostru de date, după cum se poate observa și în cadrul imaginii din dreapta. Se constată o asociere strânsă între vârsta pacienților și starea lor de sănătate, iar acest lucru confirmă concluzia de mai sus, pe măsură ce persoanele înaintează în vârstă, crește șansa instaurării bolilor cardiovasculare. De asemenea, un alt atribut cu un grad mare de asociere este nivelul colesterolului, urmat imediat de greutatea pacienților și glucoză, la un nivel mai scăzut, însă reprezentativ.

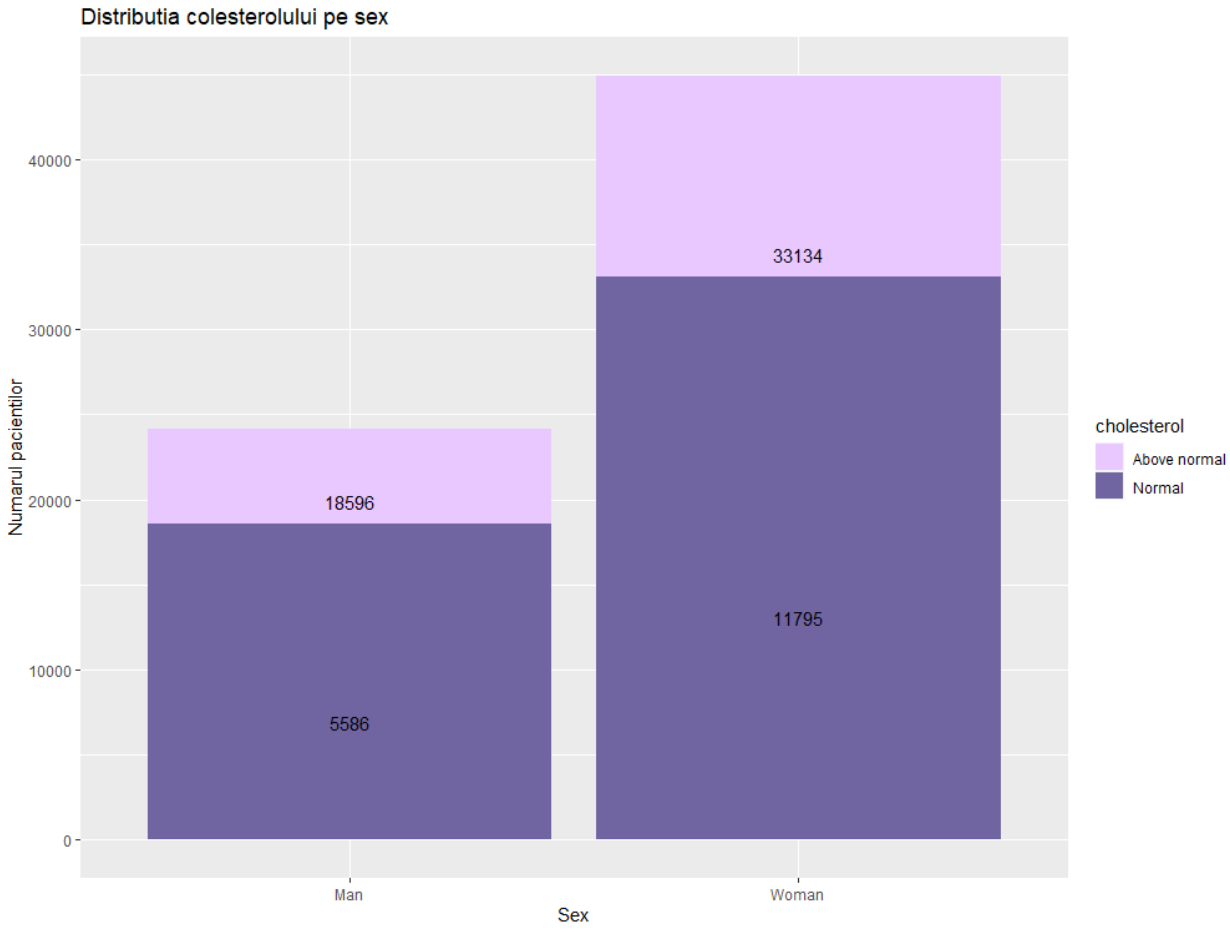


Matricea de corelație

Fumatul este unul dintre factorii care influențează apariția unei boli cardiovasculare și conform studiilor menționate în partea introductivă, declanșarea afecțiunii diferă în funcție de sex. În graficul de mai jos putem observa numărul fumătorilor pentru fiecare sex. Numărul bărbaților care fumează este de peste 6 ori mai mare decât cel al femeilor.



De cealaltă parte avem colesterolul care poate avea valori normale sau peste limita normală. Se face o altă distribuție pe sexe pentru colesterol. Chiar dacă în aparență numărul femeilor cu colesterol peste limită este mai mare, în realitate setul de date conține mai multe înregistrări pentru sexul feminin. Astfel în procente vedem că de fapt proporția este aproximativ egală între femei și bărbați: 73,74% dintre femei au colesterol peste medie, iar la bărbați avem un procent de 76,9%.



3. Rezultate si discuții

3.1 Naive Bayes

Metodele ***Naive Bayes*** sunt un set de metode de clasificare care se bazează pe aplicarea teoremei *Bayes* cu „naiva" presupunere că toate perechile de clase sunt independente unele de altele și că toți factorii au aceeași importanță în generarea modelului. În situații reale aceste ipoteze nu sunt în general valabile. Cu ajutorul teoremei putem afla probabilitatea ca un eveniment să aibă loc știind probabilitatea unui alt eveniment care deja s-a întâmplat.

Pentru a construi modelul vom aplica un algoritm prin care inițial împărțim setul de date cardio_train cu 69.058 înregistrări în două subseturi: setul de antrenament (c_train) care va conține 70% din date și setul de test (c_test) cu restul de 30%.

Data	
c_split	Large mc_split (4 elements, 4.9 MB)
c_test	20717 obs. of 12 variables
c_train	48341 obs. of 12 variables
cardio_train	69058 obs. of 12 variables

Apoi vom antrena modelul și vom putea analiza acuratețea. Vom crea alt model cu cei mai buni parametrii, pe baza căruia vom face predicțiile pentru setul de test și vom analiza matricea de confuzie cu specificitatea și senzitivitatea rezultată.

```
> confusionMatrix(pred, as.factor (c_test$cardio))
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  8383 3983
       Yes 1994 6373

               Accuracy : 0.7117
                 95% CI : (0.7055, 0.7179)
    No Information Rate : 0.5005
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4233

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8078
            Specificity : 0.6154
         Pos Pred Value : 0.6779
         Neg Pred Value : 0.7617
             Prevalence : 0.5005
         Detection Rate : 0.4043
   Detection Prevalence : 0.5964
      Balanced Accuracy : 0.7116

       'Positive' Class : No
```

Cu ajutorul probabilităților vom putea prezice apartenența unei înregistrări din setul de test la clasa No/Yes a atributului cardio.

```
> head(pred, 20)
 [1] Yes Yes Yes No  No  No  Yes No  No  Yes No  No  Yes Yes Yes No  No  No  Yes Yes
Levels: No Yes
> head(predProb, 20)
           No        Yes
1  0.428705233 0.57129477
2  0.127333615 0.87266639
3  0.154539088 0.84546091
4  0.923486117 0.07651388
5  0.801067358 0.19893264
6  0.851338880 0.14866112
7  0.438026369 0.56197363
8  0.870966010 0.12903399
9  0.797835959 0.20216404
10 0.399390407 0.60060959
11 0.859575373 0.14042463
12 0.525645748 0.47435425
13 0.034923992 0.96507601
14 0.216127221 0.78387278
15 0.005783552 0.99421645
16 0.940891356 0.05910864
17 0.927236848 0.07276315
18 0.897012391 0.10298761
19 0.437605143 0.56239486
20 0.016293798 0.98370620
> |
```

În funcție de probabilitățile rezultate vom realiza curba ROC. Pe axa x vom avea specificitatea, iar pe axa y senzitivitatea.

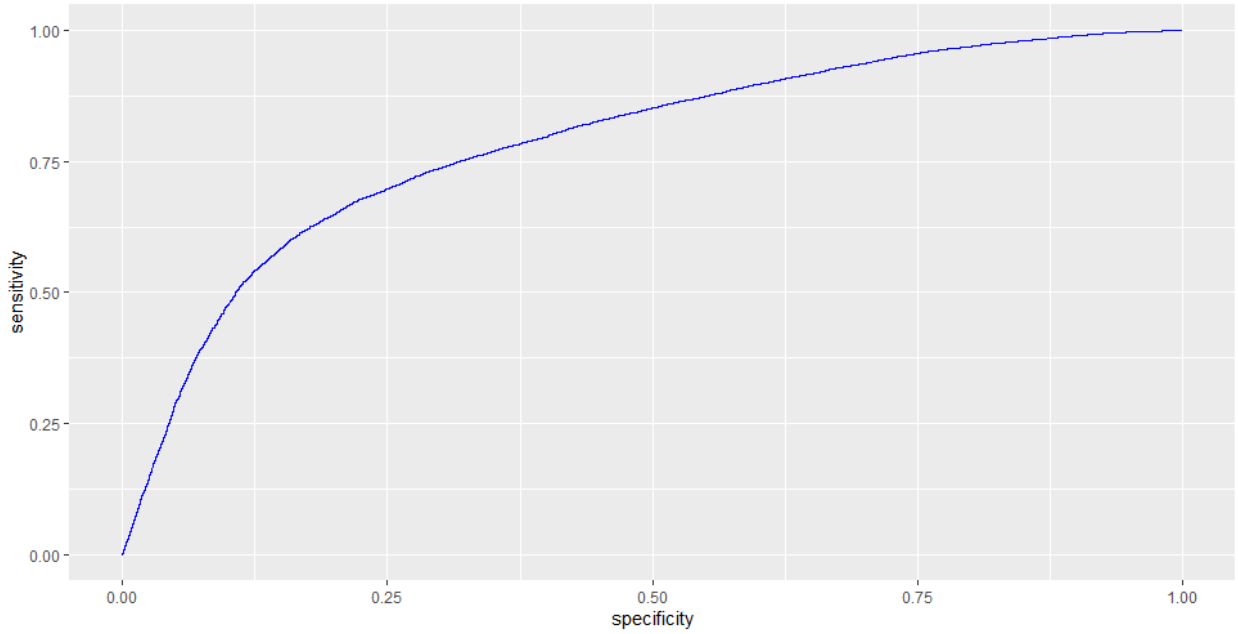
Curba ROC

```
> roc.val

Call:
roc.formula(formula = actual.class ~ probability, data = dataset)

Data: probability in 10371 controls (actual.class No) > 10346 cases (actual.class Yes).
Area under the curve: 0.7863
>
```

Aria de sub curba(AUC) este de 78.63%, destul de mică am putea spune, deci modelul va prezice spre exemplu 75% oameni bolnavi, cu costul de a găsi alți 30% care nu sunt bolnavi.

Senzitivitate = True positive/(True positive + False negative)

Specificitate = True negative/(True negative + False positive)



Antrenăm modelul cel mai bun pe tot setul de antrenament (usekernel TRUE, fL=0.5, adjust=4), deci fara a mai folosi cross-validation sau alte metode de validare.

```
289 searchOne <- expand.grid(
290   usekernel = TRUE,
291   fL = 0.5,
292   adjust = 4
293 )
294 fitControlNone <- trainControl(
295   method = "none"
296 )
```

Aplicând pe setul de test obținem o acuratețe de 72.42%, și o specificitate de 63.46%; concluzia: nu prea reușește să descopere multe persoane dintre cele bolnave, puțin mai bine de jumătate.

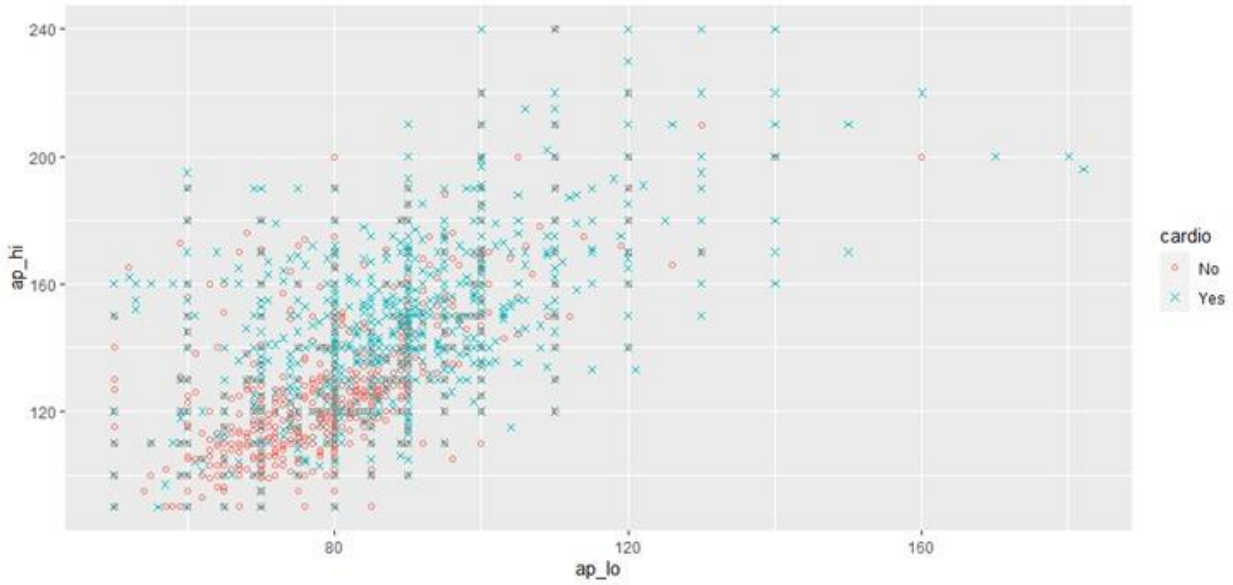
```
> confusionMatrix(predNone, test$cardio)
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  8437 3780
       Yes 1934 6566

               Accuracy : 0.7242
                 95% CI : (0.718, 0.7303)
    No Information Rate : 0.5006
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4483

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8135
            Specificity : 0.6346
         Pos Pred Value : 0.6906
         Neg Pred Value : 0.7725
             Prevalence : 0.5006
         Detection Rate : 0.4073
   Detection Prevalence : 0.5897
      Balanced Accuracy : 0.7241

       'Positive' Class : No

> |
```

3.2 Regresia logistică

Regresia logistică modelează probabilitatea unei instanțe de a aparține unei categorii particulare a clasei țintă, mai concret măsoară relația dintre variabila dependentă categorică (caracteristică) și una sau mai multe variabile independente prin estimarea „probabilităților”, folosind o funcție logistică. În acest sens, se dorește găsirea acelor valori ale coeficienților funcției logistice care vor da valori apropiate de 1 pentru acei pacienți care aparțin clasei respective și valori apropiate de zero pentru persoanele care nu fac parte din acea clasă.

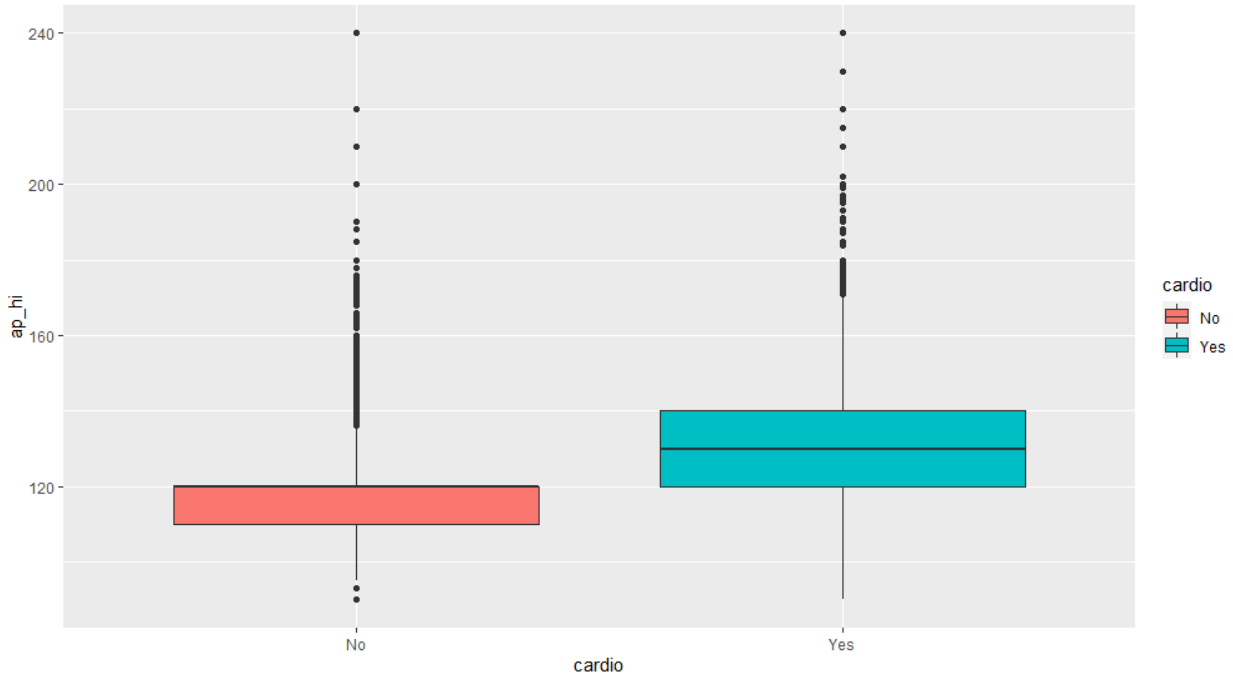
Cu ajutorul acestui model am reușit să răspundem la cea de-a doua întrebare adresată în acest studiu și anume măsura în care putem spune dacă o persoană suferă de boli cardiovasculare doar pe baza tensiunii acesteia.

În graficul de mai jos se poate observă distribuția dispersată a variabilelor corespunzătoare valorilor tensiunii din setul nostru de date, însă cu toate acestea putem trage concluzia că la un nivel crescut al tensiunii există totuși o șansă mai mare de a fi diagnosticat cu boli cardiovasculare.

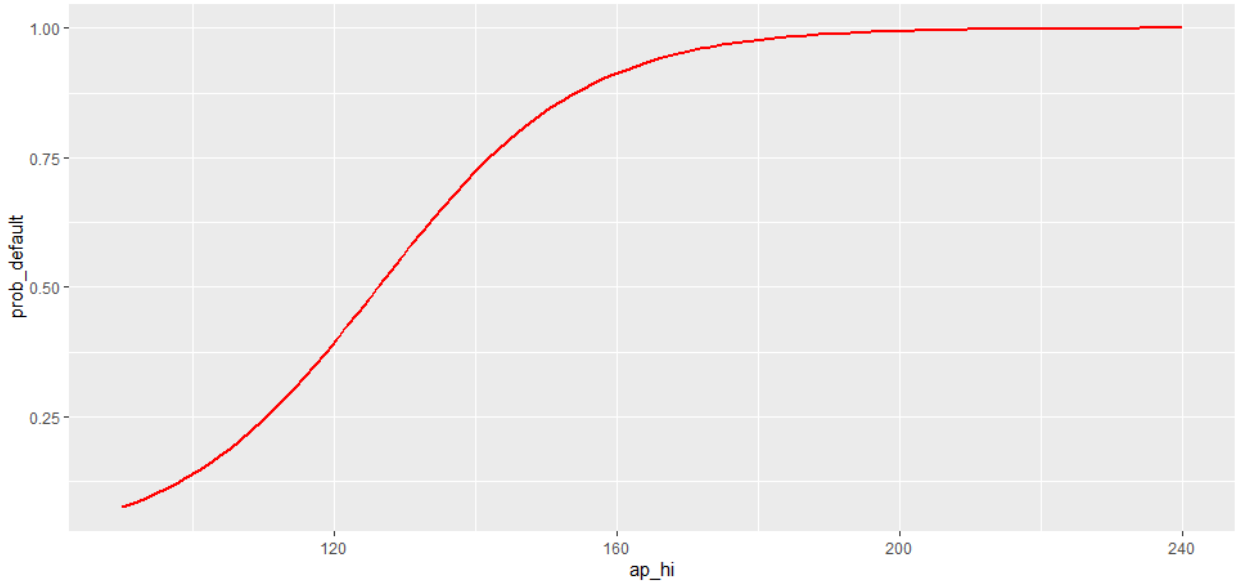


Distribuția tensiunii pentru cardio

În următorul grafic observăm că cei care au boli cardiovasculare au valori ale tensiunii mult mai ridicate și majoritatea se situează între 120-140 mmHg, în timp ce cei care nu au boli au valori ale tensiunii mai mici, între 110-120 mmHg.



Cu ajutorul funcției logistice vom putea observa corelația dintre clasa cardio și valoarea tensiunii, iar mai apoi vom construi curba S pe baza funcției, care ajută la o mai bună vizualizare a corelației dintre cele două atribute. Exemplu: pentru o valoare de 120 mmHg a tensiunii sistolice, probabilitatea de apartenență la clasa Yes este de 39%. În cazul unei persoane cu tensiunea de 160 mmHg, probabilitatea crește la 91%.

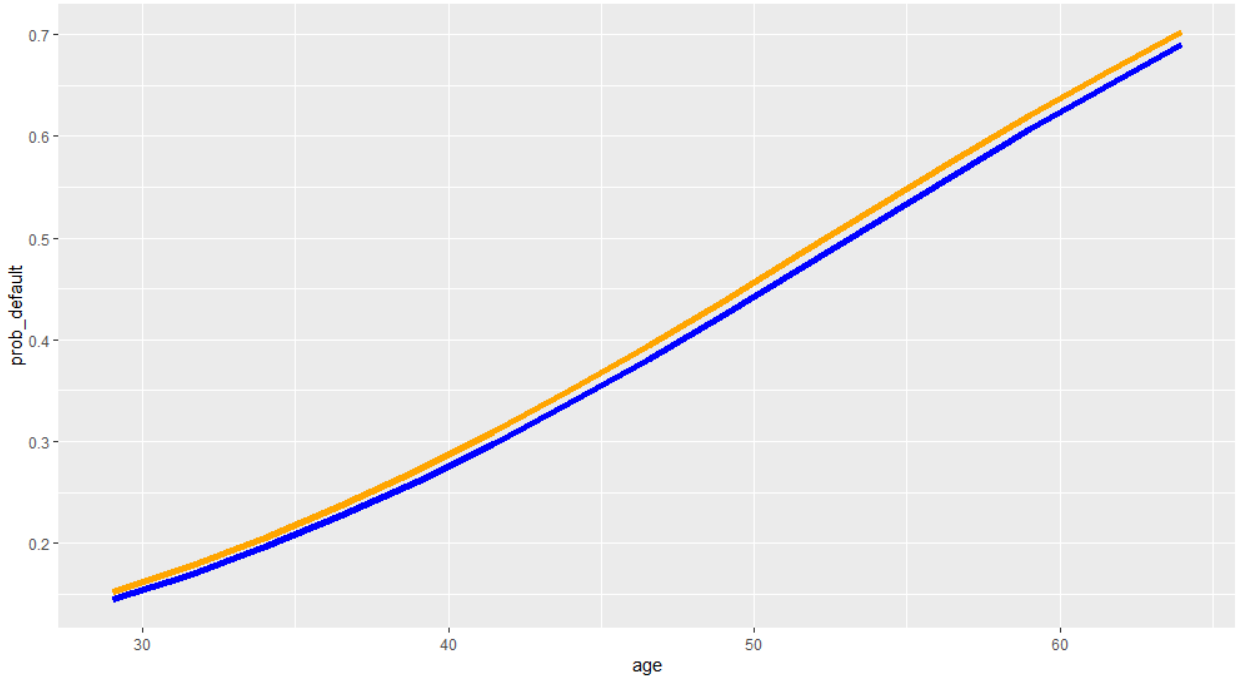


Apoi putem analiza corelația între colesterol și atributul cardio: cu cât valoarea colesterolului se apropie de o valoare normală, cu atât cresc șansele ca o persoană să fie sănătoasă.

```
> summary(mod_cholesterol)

Call:
glm(formula = as.factor(cardio) ~ cholesterol, family = binomial,
    data = cardio_train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.503  -1.077  -1.077   1.281   1.281

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.73819    0.01622   45.52   <2e-16 ***
cholesterolNormal -0.97945    0.01848  -53.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95808  on 69110  degrees of freedom
Residual deviance: 92845  on 69109  degrees of freedom
AIC: 92849

Number of Fisher Scoring iterations: 4
```

Se analizează corelația dintre atributele cardio, sex și vârstă pentru a confirma sau infirma datele studiilor. Conform studiilor menționate rata de îmbolnăvire crește odată cu vârsta și diferă în

funcție de sex: bărbații se îmbolnăvesc mai repede și femeile mai târziu.



Graficul confirmă faptul că bărbații se îmbolnăvesc într-un număr puțin mai mare, diferența fiind mai vizibilă de la 40 ani, însă datorită setului de date echilibrat, această diferență nu este vizibil mai mare.

Aria de sub curbă (AUC) înregistrată în cadrul metodei regresiei logistice este de 79.35%, față de 78.63% în cadrul Naive Bayes, fapt pentru care, comparând cele două metode putem spune, pe baza acestu indicator că regresia logistică este mult mai reprezentativă pe setul de date ales.

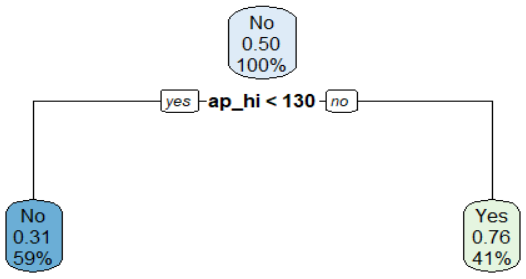
3.3 Arbori de decizie + bagging

Arborii de decizie sunt modele de predicție folosite atât pentru predicții numerice cât și pentru clasificare. Arborii nu au robustețe: dacă pentru aceeași problemă se schimbă puțin setul de date se va produce o schimbare foarte mare în arborele care se obține. Construiția unui astfel de arbore se realizează cu ajutorul recursivității. La fiecare nod al arborelui, toate atributele din setul de date sunt evaluate în vederea găsirii acelui atribut care împarte cel mai bine observațiile din setul de date. În general, arborii de decizie sunt intuitivi și ușor de interpretat, însă sunt destul de slabi, un singur arbore de decizie nu este suficient pentru a face predicții, ci în schimb e nevoie de combinarea mai multor arbori în vederea obținerii unui model mai bun.[4]

După construirea arborelui de decizie se vizualizează importanța fiecărui atribut. Ordinea importanței acestora este*: ap_hi*, *ap_lo*, *cholesterol* etc.

```
> summary(m1)
Call:
rpart(formula = cardio ~ ., data = c_train, method = "class")
  n= 48378

         CP nsplit rel error    xerror        xstd
1 0.4352756      0 1.0000000 1.0166363 0.004550639
2 0.0100000      1 0.5647244 0.5647244 0.004096140

Variable importance
      ap_hi       ap_lo cholesterol      weight         age        gluc
         51          32           7           6           3           2
```

Arborele rezultat este următorul:

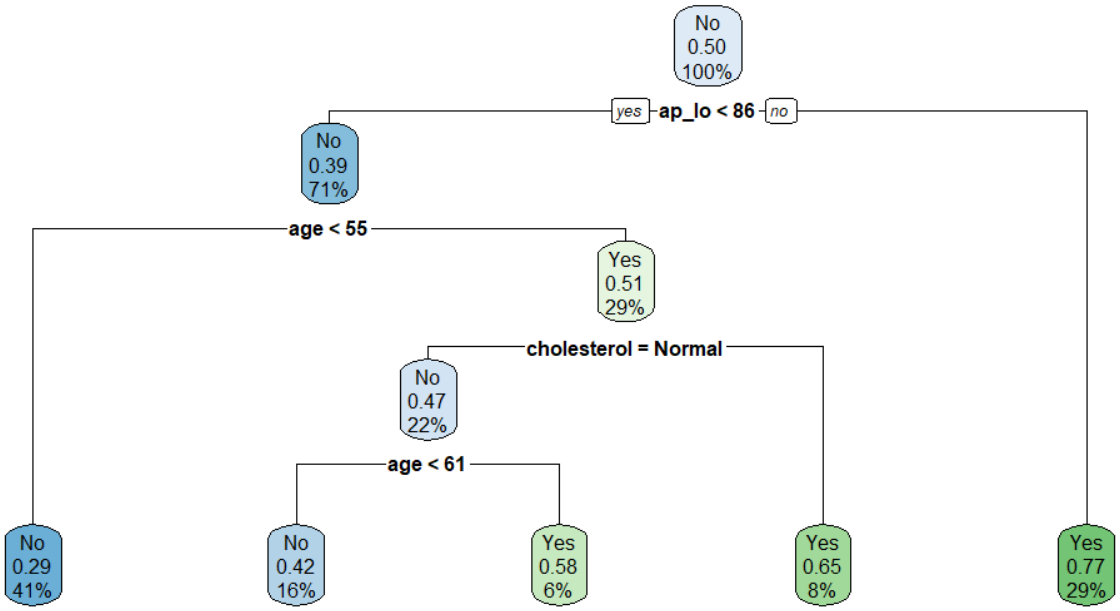


Arbore de clasificare după ap_hi (tensiunea sistolică)

[4] https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6

Primul atributul considerat este cel care minimizează rata de eroare de clasificare, în acest caz *ap_hi*. Din cauza diferențelor mari între valorile importanței variabilelor este și singurul luat în considerare. Astfel, în cadrul arborelui reprezentat în imaginea de mai sus, putem observa că primul atribut care împarte setul de date cel mai bine este *ap_hi* care impune un *treshold* de clasificare la o valoare a tensiunii de 130mmHg. În acest sens, pe următorul nivel al arborelui, 31% dintre persoanele din setul nostru de date sunt clasificate în cadrul clasei No deoarece înregistrează valori ale tensiunii mai mici de 130mmHg, în timp ce 76% sunt clasificate în clasa Yes, datorită valorilor crescute ale tensiunii.

Pentru o mai bună analiză se va exclude atributul *ap_hi* și se generează un alt arbore în scopul comparării rezultatelor.



Arbori de decizie cu ap_lo & age & cholesterol

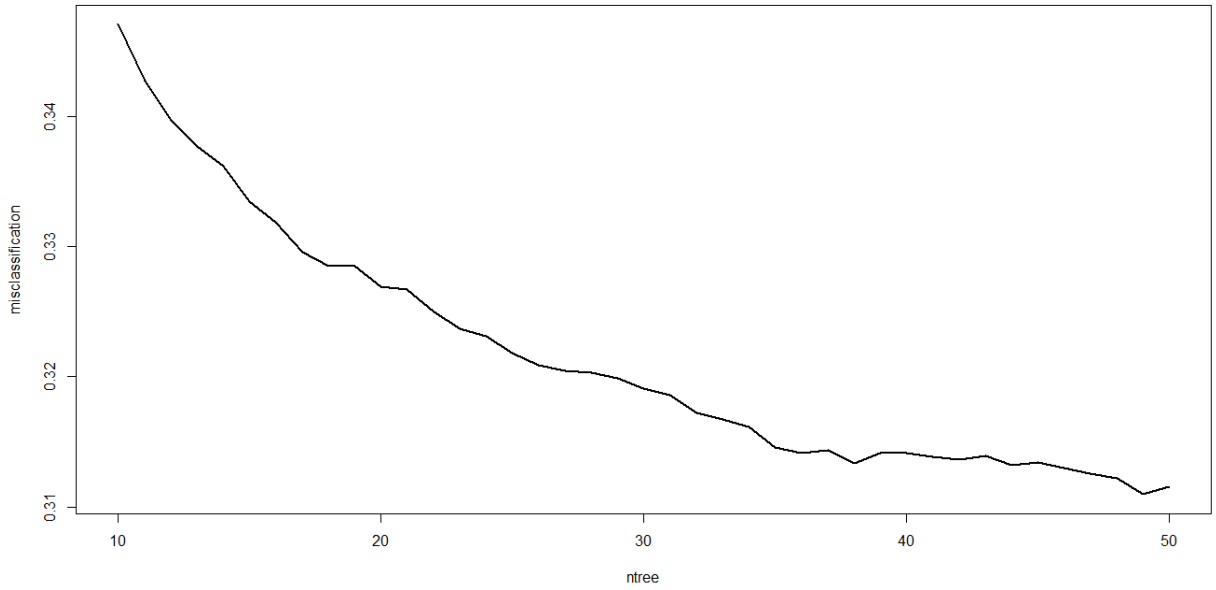
Noul arbore clasifică mai detaliat instanțele din setul de date pe baza a trei atribute: *ap_lo*, *age* și *cholesterol*, fapt pentru care are și o adâncime mai mare. Exemplu: persoanele care au *ap_lo* (tensiunea diastolică) mai mare de 86mmHg sunt automat clasificate ca făcând parte din clasa Yes. Pentru celelalte, vom face o diferențiere atât după vârstă și colesterol.

Acuratețea pe setul de test:

```
> confusionMatrix(factor(pred_m1$class), factor(c_test$cardio))
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  8269 3980
       Yes 2108 6376

               Accuracy : 0.7064
                 95% CI : (0.7001, 0.7126)
    No Information Rate : 0.5005
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4126

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7969
            Specificity : 0.6157
         Pos Pred Value : 0.6751
         Neg Pred Value : 0.7515
             Prevalence : 0.5005
         Detection Rate : 0.3988
   Detection Prevalence : 0.5908
      Balanced Accuracy : 0.7063

       'Positive' Class : No

> |
```

Bagging este un algoritm creat pentru a îmbunătății stabilitatea și acuratețea unor algoritmi de clasificare sau regresie.

Acest algoritm este folosită atunci când dorim să reducem varianța unui arbore de decizie. Obiectivul este să creem mai multe subseturi de date din setul de training prin eșantionare cu *replacement*. Fiecare subset este folosit pentru a antrena arborii de decizie. Ca rezultat obținem un ansamblu de modele diferite. Vom folosi media predicțiilor arborilor care este mai robustă decât un clasificator al unui singur arbore.



După aplicarea algoritmului pe setul nostru de antrenament am constat că sunt necesare 48 de *baggs* pentru a stabiliza rata de eroare.

Acurateţe pe setul de test:

```
> confusionMatrix(pred_bagged_m1_48, factor(c_test$cardio))
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  7264 3111
       Yes 3107 7235

               Accuracy : 0.6999
                 95% CI : (0.6936, 0.7061)
    No Information Rate : 0.5006
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.3997

 Mcnemar's Test P-Value : 0.9697

            Sensitivity : 0.7004
            Specificity : 0.6993
         Pos Pred Value : 0.7001
         Neg Pred Value : 0.6996
             Prevalence : 0.5006
         Detection Rate : 0.3506
   Detection Prevalence : 0.5008
      Balanced Accuracy : 0.6999

       'Positive' Class : No

> |
```

4. Concluzia

În urma studiului și analizei celor 3 modele alese am reușit să răspundem la cele 3 întrebări. La întrebarea care sunt factorii principali care duc la apariția bolilor cardiovasculare s-a răspuns cu ajutorul arborilor de decizie, tensiunea fiind factorul principal, urmată de colesterol și vârstă.

La cea de-a doua întrebare, cât de exact prezicem apartenența la o clasă bazându-ne pe valoarea tensiunii, am răspuns folosind regresia logistică(vezi S curve). Cunoscând toate celelalte atribute despre pacient putem prezice apartenența la o clasă cu o acuratețe de 73% bazându-ne pe modelul Naive Bayes.

Răspunsul celei de-a treia întrebări este: Da, sexul și vârsta influențează apariția unei boli cardiovasculare. Plecând de la concluzia studiului realizat BMJ Group Heath care susține legătura dintre boală sex și vârstă, putem confirma că există într-adevăr o legătură însă pe setul nostru de date ea este mai puțin vizibilă.

	Model	**Acuratețe set test**	**AUC**
1	Naive Bayes	72.23%	78.63%
2	Regresie Logistică	71.74%	79.35%
3	Arbori de decizie	69.99%	70.08%

Bibliografie

Heart Disease Prediction From Patient Data https://www.r-bloggers.com/2019/09/heart-disease-prediction-from-patient-data-in-r/

Heart Disease Prediction https://rstudio-pubs-static.s3.amazonaws.com/396380_639e2f68b09e41a0b05f97b5dc8eb3f2.html

Correlation Plots https://rkabacoff.github.io/datavis/Models.html

Decision Tree in R https://www.guru99.com/r-decision-trees.html

High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5059018/

FE, tuning and comparison of the 20 popular models for Cardiovascular Disease prediction https://www.kaggle.com/vbmokin/20-models-for-cardiovascular-disease-prediction

Naive Bayes Classifiers https://www.geeksforgeeks.org/naive-bayes-classifiers/