Klasifikacijska i regresijska stabla odlučivanja

Šabić, Mate

Master's thesis / Diplomski rad

2016

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:546677

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-04-18**

Repository / Repozitorij:

Repository of the Faculty of Science - University of Zagreb

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO – MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Mate Šabić

KLASIFIKACIJSKA I REGRESIJSKA STABLA ODLUČIVANJA

Diplomski rad

Voditelj rada:

prof. dr. sc. Anamarija Jazbec

Zagreb, 2016.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu

1._____, predsjednik

2._____, član

3._____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem se svojoj mentorici Anamariji Jazbec na
pristupačnosti i savjetima koje mi je pružila pri izradi ovog diplomskog rada.
Također, zahvaljujem se svojim prijateljima i kolegama koji su mi olakšali studiranje
svojim savjetima i ugodno provedenom vremenu.
Posebnu zahvalnost iskazujem svojim roditeljima i bratu, koji su uvijek bili uz mene,
bilo da se radi o sretnim ili teškim trenutcima.

Sadržaj

1. Uvod 1

2. Opis podataka 2

2.1. Deskriptivna statistika 6

3. Klasifikacijska i regresijska stabla odlučivanja 9

4. Klasifikacijsko stablo 10

4.1. Izgradnja stabla 10

4.2. Algoritam podjele stabla 12

4.3. Adekvatnost modela 13

4.4. Procjena greške grupiranja 16

4.5. Skraćivanje stabla 18

4.6. Zaustavljanje stabla 22

4.7. Izbor najboljeg skraćenog stabla 22

4.7.1. Nezavisnost testiranih podataka (eng. *independent test set*) 23

4.7.2. Krosvalidacija (eng. *cross-validation*) 24

4.8. Pravilo procijene pogreške 26

5. Regresijska stabla 27

5.1. Vrijednost konačnog čvora 27

5.2. Strategija dijeljenja čvora 28

5.3. Skraćivanje stabla 29

5.4. Izbor najboljeg skraćenog stabla 30

6. Dodatni pristupi 31

6.1. Višestruki pristupi 31

6.2. Stabla doživljenja (eng. survival trees) 31

6.3. Višestruki prilagodljivi regresijski splajnovi – MARS (eng. - multivariate adaptive *regression splines*) 32

6.4. Bagging i Boosting 33

6.5. Nedostajući podaci (eng. missing values) 34

6.6. Softverski paketi 35
7. Primjena CART analize u medicini 36
7.1. R Studio 36
7.2. Rpart (eng. Recursion partitioning and Regression Trees) 37
8. Literatura 51
9. Sažetak 52
10. Summary 53
11. Životopis 54

1. Uvod

Razvitkom raznih statističkih i numeričkih metoda, stvorile su se raznovrsne i detaljnije tehnike za analizu podataka. Unatoč činjenici što su se sve te metode pokazale korisne u praksi i teoriji, najveći problem se javlja kod odlučivanja koju metodu koristiti na problemu s kojim se suočavamo. S takvim su se zaprekama najviše mučili specijalisti u zdravstvenom sektoru čiji su softveri omogućavali implementaciju složenih tehnika ali bez tumačenja rezultat ili smjernica potrebnih za izradu analize.

Cilj medicinskih istraživanja je razvoj pouzdanog pravila odlučivanja koji se koristi za klasifikaciju novih pacijenata u važnim kategorijama. Kao odgovor na takvu složenost klasifikacijsko i regresijsko stablo odlučivanja, CART (eng. *Classification and Regression Tree*), postalo je vrlo popularno i korisno u mnogim područjima.

Počeci CART analize pojavljuju se u knjizi *„Classification and Regression Tree“* (1984) čiji su autori L. Breiman, J.H. Friedman, R.A. Olshen i C.J Stone, koji se ujedno smatraju njenim izumiteljima, a među kojima se najviše ističe profesor Jerome H. Friedman zbog svog truda i zalaganja u toj grani statistike.

Ovaj diplomski rad opisuje metodu CART analize popraćenu praktičnim primjerima uz primjenu statističkog programa R.

2. Opis podataka

Tema studije: Prognoza vrijednosti dobutamina u predviđanju srčanog događaja kod pacijenata koji imaju ili za koje se pretpostavlja da imaju koronarnu arterijsku bolest. Podaci su preuzeti iz odjela kardiologije u UCLA školi medicine koja se nalazi u Los Angeles-u, California. Podaci se mogu pronaći na web stranici www.stat.ucla.edu/projects/datasets/cardiac-explanation.html (2016), a dodatna analiza kao i izrada stabla mogu se pronaći na popisu literature pod brojem [3].

Skup podataka s kojim radimo analizu sastoji se od 558 podataka i 28 varijabli. Varijable koje koristimo u ovom radu su:

- *Bhr* – osnovna brzina otkucaja srca (eng. *bazal heart rate*)
- *Basebp* – osnovni krvni tlak (eng. *basal blood pressure*)
- *Basedp* – osnovni dupli produkt (eng. *basal double product*) $= bhr \times basebp$
- *Pkhr* – maksimalni puls kod snimanja holtera (eng. *peak heart rate*)
- *Sbp* – sistolički krvni tlak (eng. *systolic blood pressure*)
- *Dp* – dupli produkt (eng. *double product*) $= pkhr \times sbp$
- *Dose* – doza primljenog dobutamina (eng. *dose of dobutamine given*)
- *Maxhr* – maksimalna brzina otkucaja srca (eng. *maximum heart rate*)
- *%mphr.b* – postotak predviđene maksimalne brzine otkucaja srca koju postiže pacijent (eng. *percent of maximum predicted heart rate achieved by patient*)
- *Mbp* – maksimalni (sistolički) krvni tlak (eng. *Maximum blood pressure*)
- *Dpmaxdo* – dupli produkt maksimalne doze dobutamina (eng. *double product on maximum dobutamine dose*)
- *Dobdose* – doza dobutamina kod koje je došlo do maksimalnog duplog produkta (eng. *dobutamine dose at which maximum double product occured*)
- *Age* – broj godina pacijenta
- *Gender* – spol pacijenta

- BaseEF – osnovna srčana ejekcijska frakcija, tj. mjera učinkovitosti pumpanja srca (eng. *baseline cardiac ejection fraction - a measure of the heart's pumping efficiency*)
- *DobEF* – ejekcijska frakcija dobutamina (eng. *Dobutamin ejection fraction*)
- *Chestpain* – bol u prsima
- *PosECG* – znakovi srčanog udara na ECG-u (eng. *signs of heart attack on ECG*)
- *Equivecg* – ECG je dvosmislen (eng. *ECG is equivocal*)
- *RestWMA* – kardiolog vidi nenormalno gibanje oko srca na ehokardiogramu (eng. *Rest wall motion abnormalities*)
- *PosSE* – indikator za pozitivnost ehokardiograma (eng. *Positive stress echocardiogram*)
- *MI* – imao infarkt miokarda ili srčani udar (eng. *recent myocardial infarction or heart attack*) - da = 0
- *PTCA* – imao revaskularizaciju po perkutnoj koronarnoj angioplastici (eng. *recent angioplasty*) – da = 0
- *CABG* – imao operaciju presađivanja koronarne arterije (eng. *recent bypass surgery*) – da = 0
- *Death* – pacijent umro – da = 0
- *HxofHT* – indikator dali je pacijent imao hipertenziju, tj. povišeni krvni tlak (eng. *History of hypertension*)
- *Hxofdm* - indikator dali je pacijent imao dijabetes (eng. *patient has history of diabetes*)
- *Hxofcig* - indikator dali je pacijent pušio (eng. *patient has history of smoking*)
- *HxofMI* – pacijent ima povijest srčanog udara (eng. *patient has history of heart attack*)
- *HxofPTCA* – pacijent ima povijest angioplastike (eng. *patient has history of angioplasty*)
- *HxofCABG* – pacijent imao operaciju (eng. *patient has history of bypass surgery*)

- *Any* – zavisna varijabla – ako je pacijent imao barem jedan od idućih događaja: MI, PTCA, CABG ili umro (eng. *it is defined as "death or MI or PTCA orCABG". if any of these variables is positive then "any" is also postive*) (any poprima vrijednosti: 0 što označava potvrdu nekog događaja i 1 da se nijedan nije dogodio)

Stresna ehokardiografija dobutamina (*DSE*) se često i uspješno koristi za određivanje da li pacijenti bez ili sa koronarnom arterijskom bolesti ima ishemiju, tj. nedovoljan priljev krvi. Nadalje, *DSE* se uglavnom koristi kad pacijent nije u stanju fizički vježbati do stupnja koji bi trebao pružiti korisne kliničke informacije. Prednosti su da ne zahtijeva suradnju pacijenta i da ehokardiogram, tj. ultrazvuk srca ima dovoljno informacija da se dobije odgovarajuća obrada slike na svim razinama stresa.

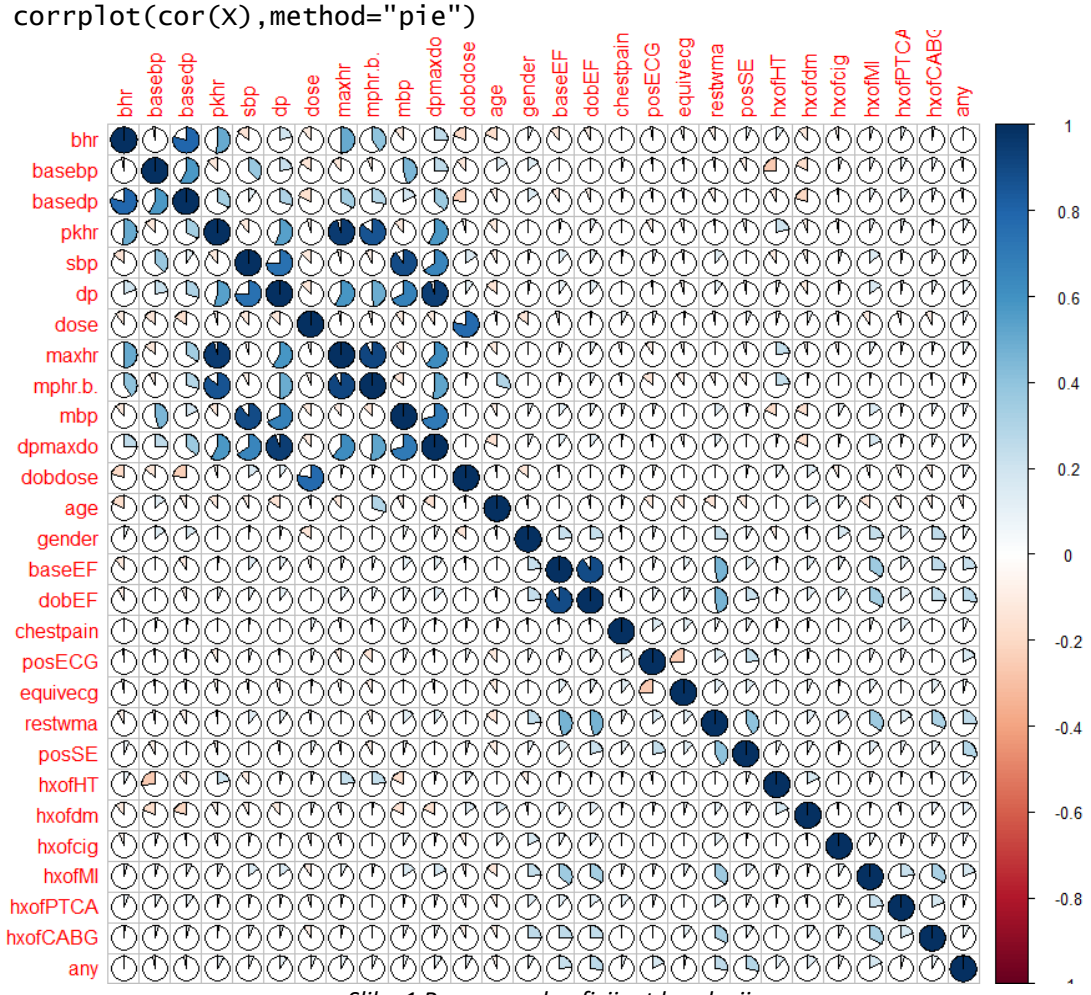
Cilj ove studije jest praćenje pacijenata koji su se podvrgli *DSE*-u tijekom pet godina i određivanje koje su studije vezane za najbolje testiranja stresa za predviđenu izlaznu informaciju, tijekom 12 mjeseci.

Sveukupno je 1 183 pacijenata izloženo DSE-u između 1991. i 1996. godine na UCLA laboratoriju, koji je dao pristanak da se preispitaju njihovi medicinski podaci. Ako je pacijent imao više od jedne DSE tijekom tog perioda tada gledamo samo prvi test. Međutim, nisu svi pacijenti ušli u analizu studije. 208 pacijenata koji su uzeti u obzir za transplantaciju jetre su isključeni, kao i 12 onih koji su umrli od ne srčanog uzroka tijekom hospitalizacije. Zatim, 2 pacijenta koji su imali transplantaciju srca i 13 njih koji su imali PTCA unutar šest mjeseci prije testiranja. Ostalih 376 nisu doživjeli kraj testiranja, a 14 ih je odbačenu jer im je nedostajao jedan ili više podataka. Končano, studija se provodi na 558 pacijenata koji nisu imali nijedan od prethodno navedenih događaja ili su imali neki srčani događaj: smrt, MI, PTCA ili CABG unutar 12 mjeseci. Od 558 pacijenata njih 90 je imalo jedan ili više od četiri srčana događaja, dok je preostalih 468 prošlo bez srčanih posljedica. Dobna granica pacijenata je između 26 i 93 godine, s prosječnom vrijednosti od 67 godina.

Dobutamin je davan intravenozno, tj. direktno u venu, pomoću standardnog uređaja za odmjerenu količinu počevši sa 5 µg/kg/min. Nova primljena količina se svaki put povećavala za 5 µg/kg/min svako 3 minute do maksimalne doza od 40 µg/kg/min. U iduće 2 godine doze od 25 i 35 µg/kg/min su izostavljanje, a antropin je davan u venu kada je broj otkucaja srca bio ne adekvatan.

Nadalje, test je simptomatski ograničen i zaustavljen ako je pacijent imao više od 2 mm ST depresije, tj. električne aktivnosti na srcu tijekom nekog perioda pomoću elektroda postavljenih po tijelu osobe. Zatim zaustavljen je zbog nenormalnih gibanja u području srčanog mišića, ventrikularne tahikardije, sistoličkog krvnog tlaka većeg od 220 mm Hg, mučnine, umjerene do jake boli u prsima ili zadaha.

```
> corrplot(cor(X),method="pie")
```



Slika 1 Pearsonov koeficijent korelacije

Pearsonovim koeficijentom korelacije uočavamo međusobne odnose varijabli. Vrijednost Pearsonovog koeficijenta kreće se od -1 (negativne korelacije) do +1 (pozitivne korelacije). Predznak koeficijenta nas upućuje na smjer korelacije.

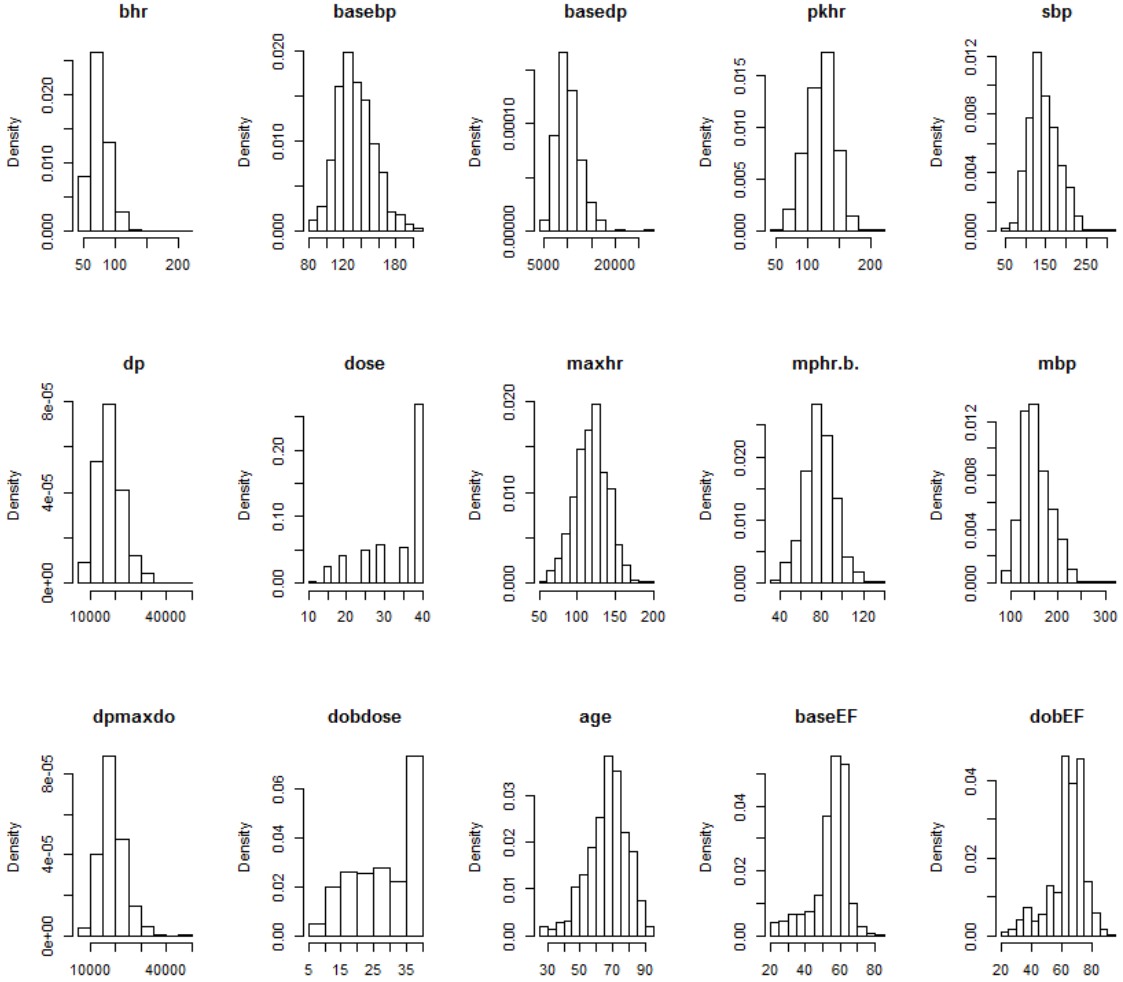
Iz prikupljenih podataka znamo da su nam neke varijable linearno zavisne, međutim većina ih je nezavisna što se može učiti iz slike 1. Iz slike ujedno vidimo da nam je manje dio podataka u nekoj značajnoj korelaciji.

2.1. Deskriptivna statistika

```
> summary(X)
      bhr             basebp          basedp           pkhr      
 Min.   : 42.00   Min.   : 85.0   Min.   : 5000   Min.   : 52.0  
 1st Qu.: 64.00   1st Qu.:120.0   1st Qu.: 8400   1st Qu.:106.2  
 Median : 74.00   Median :133.0   Median : 9792   Median :122.0  
 Mean   : 75.29   Mean   :135.3   Mean   :10181   Mean   :120.6  
 Stdev  : 15.42   Stdev  : 20.77  Stdev  :2579.75 Stdev   :22.57 
 3rd Qu.: 84.00   3rd Qu.:150.0   3rd Qu.:11663   3rd Qu.:135.0  
 Max.   :210.00   Max.   :203.0   Max.   :27300   Max.   :210.0  

     sbp              dp             dose           maxhr      
 Min.   : 40.0   Min.   : 5100   Min.   :10.00   Min.   : 58.0  
 1st Qu.:120.0   1st Qu.:14033   1st Qu.:30.00   1st Qu.:104.2  
 Median :141.0   Median :17060   Median :40.00   Median :120.0  
 Mean   :146.9   Mean   :17634   Mean   :33.75   Mean   :119.4  
 Stdev  : 36.53  Stdev  :5220.53 Stdev  : 8.134   Stdev  : 21.91 
 3rd Qu.:170.0   3rd Qu.:20645   3rd Qu.:40.00   3rd Qu.:133.0  
 Max.   :309.0   Max.   :45114   Max.   :40.00   Max.   :200.0  

   mphr.b.           mbp           dpmaxdo         dobdose      
 Min.   : 38.00   Min.   : 84.0   Min.   : 7130   Min.   : 5.00  
 1st Qu.: 69.00   1st Qu.:133.2   1st Qu.:15260   1st Qu.:20.00  
 Median : 78.00   Median :150.0   Median :18118   Median :30.00  
 Mean   : 78.57   Mean   :156.0   Mean   :18550   Mean   :30.24  
 Stdev  : 15.12   Stdev  : 31.71  Stdev  :4901.43 Stdev  : 9.53  
 3rd Qu.: 88.00   3rd Qu.:175.8   3rd Qu.:21239   3rd Qu.:40.00  
 Max.   :133.00   Max.   :309.0   Max.   :45114   Max.   :40.00  

     age            baseEF          dobEF      
 Min.   :26.00   Min.   :20.0   Min.   :23.00  
 1st Qu.:60.00   1st Qu.:52.0   1st Qu.:62.00  
 Median :69.00   Median :57.0   Median :67.00  
 Mean   :67.34   Mean   :55.6   Mean   :65.24  
 Stdev  :12.05   Stdev  :10.32  Stdev  :11.76  
 3rd Qu.:75.00   3rd Qu.:62.0   3rd Qu.:73.00  
 Max.   :93.00   Max.   :83.0   Max.   :94.00  
```

Histogram

```
> for (i in 1:15) hist(X[,i] ,main=toString(names(X[i])))
```



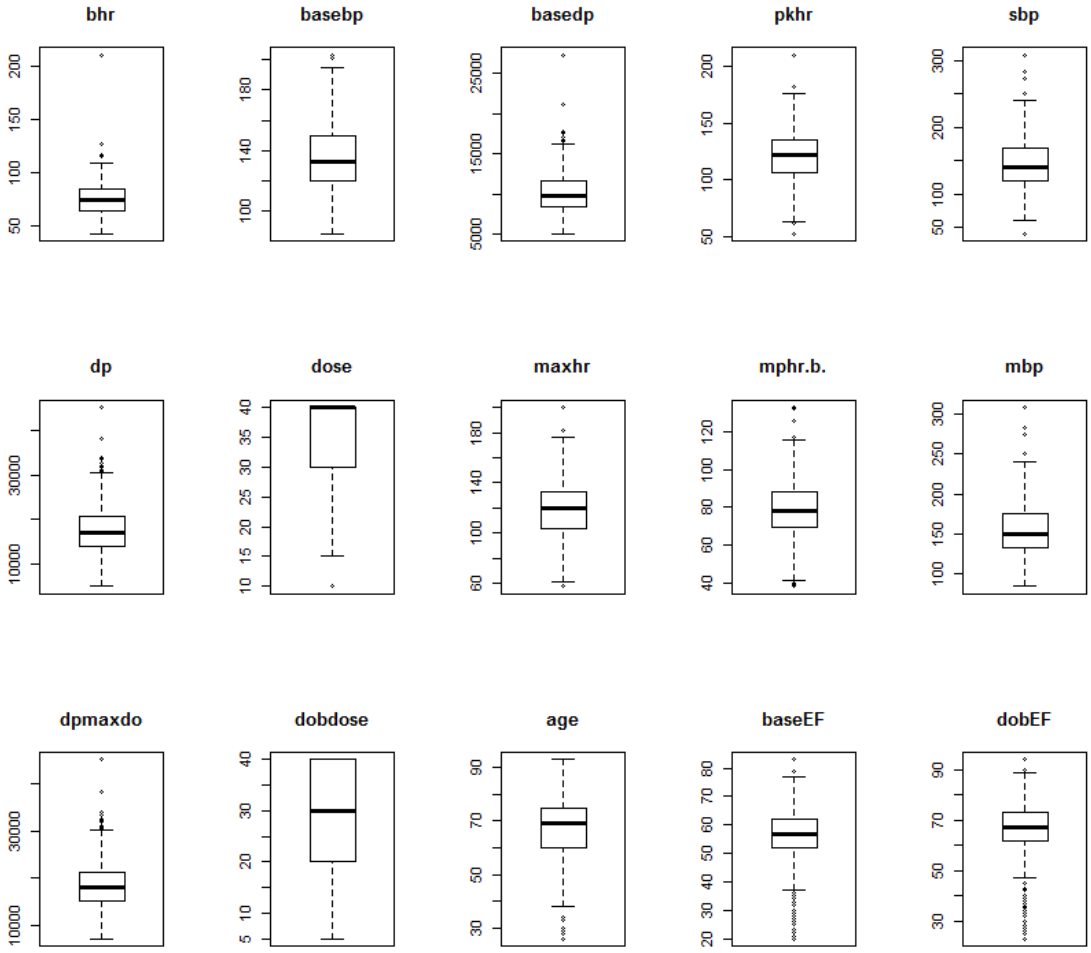
Slika 2 Histogram podataka

	Ukupno	(%)	Godine	SD	Imao srčani		Nije imao srčani	
					Ukupno	(%)	Ukupno	(%)
Pacijenti	558		67	12	90	16	468	84
Muški	220	39	68	12	43	20	177	80
Ženski	338	61	67	12	47	14	291	86

Tablica 1 Deskriptivna statistika

Box-plot

```
> for (i in 1:15)  boxplot(X[,i],main=toString(names(X[i])))
```



Slika 3 Box-plot analiziranih podataka

3. Klasifikacijska i regresijska stabla odlučivanja

CART analiza, je analitički alat koji, koristeći povijesne podatke, pomaže u određivanju važnih i značajnih varijabli u skupu podataka za izradu stabla odluke, koji služi za daljnju izradu modela. To je ne-parametrijska statistička metoda koja se razlikuje ovisno o definiciji izlazne varijable.

Ako je zavisna varijabla kategorijska, najčešće binarna, klasifikacijsko stablo organizira podatke u grupe bazirane na homogenosti i sličnosti. Međutim, ako je kontinuirana podaci se koriste da se predvidi vrijednost varijable izlaza te se primjenjuje regresijski model na svaku od nezavisnih varijabli.

Ideja koja se koristi u CART analizi jest rekurzivno dijeljenje podataka u manje podgrupe zbog poboljšanja bolje izrade modela. Naime, metoda rekurzije jest da se podaci koje koristimo optimalno dijele na svim varijablama u svim mogućim čvorovima na dva dijela, te se svaki novi dio opet dijeli na analogan način.

Prema Gordon-u glavne komponente CART metode su: pravilo odabira varijable na određenom čvoru te pravilo zaustavljanja stabla. Pravilo odabira određuju kakve podjele izvesti u svakom koraku, a pravilo zaustavljanja određuje konačne podgrupe koje su stvorene tijekom rekurzije.

Funkcija nečistoće (eng. *impurity*) svake podgrupe je funkcija koja mjeri stupanj čistoće za područje koje sadrži podatke različitih klasa.

Postoje tri slična načina mjerenja nečistoće čvora kod klasifikacijskih stabala. To su pogreška grupiranja (eng. *missclassification error*), Gini indeks (eng. *Gini index*) i funkcija entropije (eng. *cross-entropy ili deviance*). Zadnje dvije metode su diferencijabilne te ih je lakše numerički optimizirati i ujedno su dosta osjetljive na promjene vjerojatnosti u čvorovima. Kod regresijskih stabala za mjerenje nečistoće čvora koristi se metoda najmanjih kvadrata (eng. *least square*).(vidi [4])

4. Klasifikacijsko stablo

Klasifikacijsko stablo je rezultat niza događaja koji se trebaju dogoditi, a rezultat svakog događaja ovisi o rezultatima iz prethodnog niza događaja. U primjeni ga najčešće koristimo kada za svaki podatak znamo kojoj klasi pripada.

Klasifikacijski problem sastoji se od četiri glavne komponente. Prva komponenta je kategorijska zavisna varijabla koju pokušavamo predvidjeti a bazirana je na nezavisnim varijablama što je ujedno i druga komponenta. Treća komponenta je skup podataka kojeg opažamo koji sadrži izlaznu i nezavisne varijable, dok je četvrta komponenta skup podataka kojeg želimo procijeniti.

4.1. Izgradnja stabla

Shema za izradu stabla sastoji se od:

- odabira Bolean-ovog uvjeta za dijeljenje svakog čvora i kriterija za podjelu prolaznog čvora
- odabira konačnog čvora, odnosno lista i dodjele klase konačnom čvoru

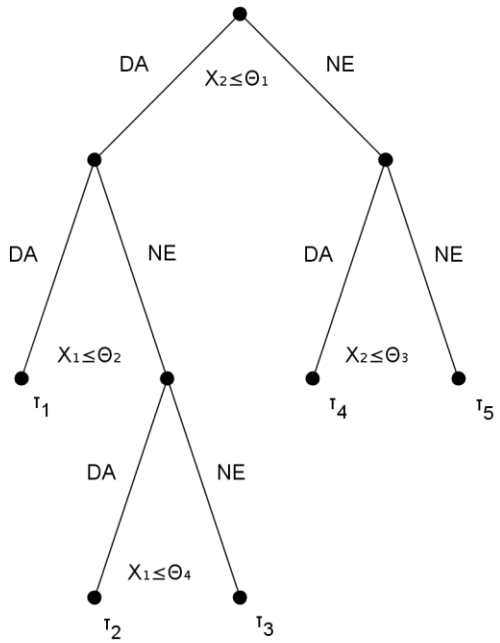
Kod svakog čvora, algoritam za izgradnju čvora treba odlučiti na kojoj varijabli je najbolje dijeliti. Treba gledati svako moguće dijeljenje u odnosu na sve varijable u tom čvoru, zatim numerirati sva moguća dijeljenja, procijeniti ih te odrediti koje je najbolje.

U izgradnji stabla polazimo od jedinstvene početne točke koja se zove glavni ili korijenski čvor (eng. *root node*) te sadrži cijeli skup podataka na vrhu stabla. Čvor je podskup od skupa varijabli koji može biti konačan (eng. *terminal*) ili prolazni (eng. *non-terminal*). Roditeljski čvor se razdvaja, na dva nova čvora (kćeri), binarno sa maksimalnom homogenosti. Takva podjela je određena Boolean-ovim uvjetom na vrijednost varijable koja se promatra, gdje je uvjet zadovoljen („da“) ili nije zadovoljen („ne“). Čvor koji se ne razdvaja, tj. onaj koji dođe do kraja ogranka stabla zove se konačni čvor ili list i poprima oznaku klase. Kada stablo u jednoj grani dođe do kraja, tj.

završi u konačnom čvoru, ta grana poprima klasu koja odgovara oznaci klase vezane za taj konačni čvor. Može se dogoditi da istu oznaku ima više čvorova.

Stablo koje se podijelilo samo jednom, tj. ono koje ima samo dva konačna čvora zove se panj (eng. *stump*), a skup svih konačnih čvorova zove se particija (eng. *partition*). Svaki konačan čvor i pripadno područje, koje odgovara uniji uvjeta na čvorovima od glavnog do konačnog čvora, poprima oznaku klase. (vidi [1], str. 282)

Primjer 1: Slika 4 prikazuje rekurzivnu podjelu sa dvije varijable ulaza koja ima stazu: ako je $X_2 \leq \theta_1$ grananje se dijeli na lijevu stranu, a u suprotnom na desnu. Ako je upit $X_2 \leq \theta_1$ potvrdan, dalje gledamo kakav je $X_1 \leq \theta_2$ upit, ako je potvrdan tada je τ_1 konačan čvor sa područjem $R_1 = \{X_1 \leq \theta_2, X_2 \leq \theta_1\}$. Ako je negativan tada gledamo upit $X_1 \leq \theta_4$. Za potvrdan odgovor τ_2 jest konačan čvor sa područjem $R_2 = \{X_1 \leq \theta_4, X_1 > \theta_2, X_2 \leq \theta_1\}$, a negativan konačan čvor je τ_3 sa $R_3 = \{X_1 > \theta_4, X_1 > \theta_2, X_2 \leq \theta_1\}$. Analogno se dobije za konačne čvorove τ_4 i τ_5 sa njihovim pripadnim područjima.



Slika 4 Rekurzivna podjela stabla

4.2. Algoritam podjele stabla

U knjizi „*Modern Multivariate Statistical Techniques: Regression, Classification, and Monifold Learning*“, Izenman je objasnio algoritam za definiranje i podjelu stabla koji je opisan u daljnjem dijelu rada.

Sa $\tau_R, \tau_L \, i \, \tau_D$ označimo redom čvorove roditelja, lijeve i desne kćeri. Sa matricom X označimo primjer promatranih podataka koji se sastoji od N podataka i M varijabli, a sa Y zavisni N dimenzionalni vektor s najviše K klasa.

Neka je x_j varijabla j=1,…,M, a x_j^R najbolja vrijednost varijable x_j koja razdvaja čvor. Roditeljska nečistoća je konstantna za svaku varijablu za koju vrijedi $x_j \leq x_j^R$, dok je maksimalna homogenost kćeriju ekvivalentna maksimalnoj vrijednosti funkcije nečistoće:

$$\Delta i(\tau) = \; i(\tau_R) - E[i(\tau_k)] \tag{1.1}$$

gdje je τ_K lijeva i desna kćer roditeljskog čvora τ_R.

Pod pretpostavkom da su $p_L \, i \, p_D$ vjerojatnosti lijevog i desnog čvora, prikladnost modela (eng. *goodness of fit*) u čvoru τ je dana dijeljenjem čvora na dvije kćeri oblikom

$$\Delta i(\tau) = \; i(\tau_R) - p_L i(\tau_L) - p_D i(\tau_D) \tag{1.2}$$

gdje se problem maksimalne homogenosti funkcije rješava kao:

$$\underset{x_j \leq x_j^R, j=1,\dots,M}{\arg\max} \; [i(\tau_R) - p_L i(\tau_L) - p_D i(\tau_D)] \tag{1.3}$$

Nadalje, neka su $\Pi_1, \dots, \Pi_K : K \geq 2$ klase. Za čvor τ funkcija $i(\tau)$ se definira kao

$$i(\tau) = \Phi\big(p(1|\tau), \dots, p(K|\tau)\big) \tag{1.4}$$

gdje je $p(K|\tau)$ procjena od $P(\boldsymbol{X} \in \Pi_K|\tau)$, tj. vjerojatnost da opažanje **X** unutar klase Π_K upada u čvor τ, a Φ simetrična funkcija definirana na skupu svih K-torki vjerojatnosti $(\mathrm{p}_1, \dots, \mathrm{p}_K)$ sa jediničnom sumom.

Primjeri takvih funkcija Φ su:

- Funkcija entropije

$$i(\tau) = -\sum_{k=1}^{K} p(k|\tau) \log p(k|\tau) \tag{1.5}$$

- Gini pravilo ili gini indeks

$$i(\tau) = \sum_{k=1}^{K} p(k|\tau)(1 - p(k|\tau)) = \sum_{k \neq k'} p(k|\tau)p(k'|\tau) = 1 - \sum_{k} \{p(k|\tau)\}^2 \tag{1.6}$$

Ako postoje samo dvije klase, tada se prethodne dvije funkcije redom reduciraju na oblike:

- $i(\tau) = -p \log p - (1-p) \log(1-p)$ (1.7)
- $i(\tau) = 2p(1-p)$ (1.8)

4.3. Adekvatnost modela

Koristeći funkciju entropije pravilo razdvajanja dobijemo na sljedeći način.

Za početak uvedemo nove oznake $n_{ab}, n_{a+}, n_{+a}, n_{++}: a, b \in \{1,2\}$, za koje vrijedi: $n_{a+} = n_{a1} + n_{a2}, n_{+a} = n_{1a} + n_{2a}, n_{++} = n_{1+} + n_{2+} = n_{+1} + n_{+2}$. Ostale veličine procjenjujemo iz tablice 2 na način da gledamo broj podataka iz skupa koji zadovoljavaju uvjete u križanju određenog retka i stupca za promatranu varijablu.

Nadalje procijenimo p_L sa n_{+1}/n_{++} i p_D sa n_{+2}/n_{++} i funkciju nečistoće za roditeljski čvor

$$i(\tau_R) = -\left(\frac{n_{+1}}{n_{++}}\right) \log\left(\frac{n_{+1}}{n_{++}}\right) - \left(\frac{n_{+2}}{n_{++}}\right) \log\left(\frac{n_{+2}}{n_{++}}\right) \tag{1.9}$$

Analogno procijenimo vjerojatnosti za čvorove kćeriju τ_L i τ_D:
za $x_j \leq x_j^R$ procijenimo p_L sa n_{11}/n_{1+} i p_D sa n_{12}/n_{1+},

za $x_j > x_j^R$ procijenimo p_L sa n_{21}/n_{2+} i p_D sa n_{22}/n_{2+},

i funkciju za obje

$$i(\tau_L) = -\left(\frac{n_{11}}{n_{1+}}\right)\log\left(\frac{n_{11}}{n_{1+}}\right) - \left(\frac{n_{12}}{n_{1+}}\right)\log\left(\frac{n_{12}}{n_{1+}}\right) \tag{1.10}$$

$$i(\tau_D) = -\left(\frac{n_{21}}{n_{2+}}\right)\log\left(\frac{n_{21}}{n_{2+}}\right) - \left(\frac{n_{22}}{n_{2+}}\right)\log\left(\frac{n_{22}}{n_{2+}}\right) \tag{1.11}$$

Na kraju računamo vrijednost funkcije (1.2) za promatranu varijablu, što provjerimo za svaku vrijednost iz skupa te varijable formulom

$$\arg\max_{x_j \le x_j^R, j=1,\dots,M} [i(\tau_R) - p_L i(\tau_L) - p_D i(\tau_D)] \tag{1.12}$$

te odaberemo maksimalnu vrijednost funkcije $i(\tau)$. (vidi [1])

	1	0	ZBROJ
$x_j \le x_j^R$	n_{11}	n_{12}	n_{1+}
$x_j > x_j^R$	n_{21}	n_{22}	n_{2+}
ZBROJ	n_{+1}	n_{+2}	n_{++}

Tablica 2 Dijeljenje varijable x_j sa varijablom odaziva koja poprima vrijednosti 0 ili 1

Takav proces traženja maksimuma funkcije radimo za svaki skup varijabli posebno u skupu kojeg promatramo. Zatim iz tog novo dobivenog skupa maksimalnih vrijednosti tražimo maksimalnu vrijednost koja će odrediti varijablu $x_j^R : j = 1, \dots, M$, kao varijablu s najboljim uvjetom za dijeljenje čvora.

Primjer 2: Tražimo vrijednost funkcije za dijeljenje početnog čvora varijablom $x = dobEF$ iz podataka koje koristimo u ovom radu opisanima u poglavlju „Opis podataka“.

Traženjem najveće vrijednosti funkcije (1.3) za varijablu x prethodnim algoritmom dobijemo najveću vrijednost kada je vrijednost varijable $x_j^R = 52$. Skup podataka u početnom čvoru dijelimo na način da lijeva grana odgovara nejednakosti $dobEF \le 52$, a desna $dobEF > 52.5$.

	1	0	ZBROJ
$x_j \leq 52$	44	33	77
$x_j > 52$	424	57	481
ZBROJ	468	90	558

Tablica 3 Broj podataka kod dijeljenja varijable dobEF

Koristeći funkciju entropije kao funkciju nečistoće iz (1.10) i (1.11) za čvorove kćeri dobijemo

$$i(\tau_L) = -\left(\frac{44}{77}\right)\log\left(\frac{44}{77}\right) - \left(\frac{33}{77}\right)\log\left(\frac{33}{77}\right) = 0.6829081$$

$$i(\tau_D) = -\left(\frac{424}{481}\right)\log\left(\frac{424}{481}\right) - \left(\frac{57}{481}\right)\log\left(\frac{57}{481}\right) = 0.3639319$$

Nadalje iz (1.9) za roditeljski čvor

$$i(\tau_R) = -\left(\frac{468}{558}\right)\log\left(\frac{468}{558}\right) - \left(\frac{90}{558}\right)\log\left(\frac{90}{558}\right) = 0.4418033$$

Zatim procijenimo $p_L = 0.1379928$ i $p_D = 0.8620072$.

Koristeći funkciju (1.2) dobijemo da nam je vrijednost funkcije za dijeljenje početnog čvora sa varijablom $dobEF$ jednaka

$$i(\tau) = 0.4418033 - 0.1379928 * 0.6829081 - 0.8620072 * 0.3639319$$
$$= 0.03385501$$

Analogno dobijemo najbolje vrijednosti funkcije kod početnog čvora za ostale varijable:

Restwma	PosSE	dobEF	BaseEF	HxofMI	posECG	Dp	HxofHT
0.03650	0.03569	0.03386	0.02378	0.01437	0.01209	0.00957	0.00696

Tablica 4 Maksimalne vrijednosti funkcije nečistoće za najbolje varijable u početnom čvoru sortirane od veće prema manjoj

Iz tablice 4 možemo uočiti da je $restwma$ najbolja varijabla koja dijeli početni čvor. Međutim, kako je to kategorijska varijabla, s vrijednostima jedan i nula, uvjet na

dijeljenje čvora glasi: ako je vrijednost podatka varijable koju gledamo jednaka nula taj podatak ide na lijevu granu stabla, a za jedan ide na desnu.

Kod korijenskog čvora radimo s cijelim skupom podataka, dok kod ostalih čvorova, na koje se taj početni čvor dijeli, vrijednost funkcije računamo na analogan način samo što ne radimo sa cijelim skupom podataka, već sa onim podskupom koji je zadovoljio ili nije uvjet za određenu varijablu u prethodnom čvoru.

4.4. Procjena greške grupiranja

Procjena greške grupiranja $r(\tau)$ promatranih varijabli dana je formulom

$$r(\tau) = 1 - \max_k p(k|\tau) \tag{1.13}$$

koja se kod dvije klase reducira na $r(\tau) = 1 - \max(p, 1-p) = \min(p, 1-p)$, gdje je k oznaka klase.

Neka je T oznaka stabla, a $\tilde{T} = \{\tau_1, \tau_2, \dots, \tau_L\}$ skup svih konačnih čvorova od T. Tada možemo procijeniti pravu grešku grupiranja za T sa

$$R(T) = \sum_{\tau \in \tilde{T}} R(\tau)P(\tau) = \sum_{l=1}^{L} R(\tau_l)P(\tau_l) \tag{1.14}$$

gdje je $P(\tau)$ vjerojatnost da promatrana vrijednost upadne u čvor τ.

Nadalje, ako $P(\tau_l)$ procijenimo po omjeru, $p(\tau_l)$, svih promatranih vrijednosti koje padnu u taj čvor τ_l, tada je izmjenjena procjena od $R(T)$ dana sa

$$R^{re}(T) = \sum_{l=1}^{L} r(\tau_l)p(\tau_l) = \sum_{l=1}^{L} R^{re}(\tau_l) \tag{1.15}$$

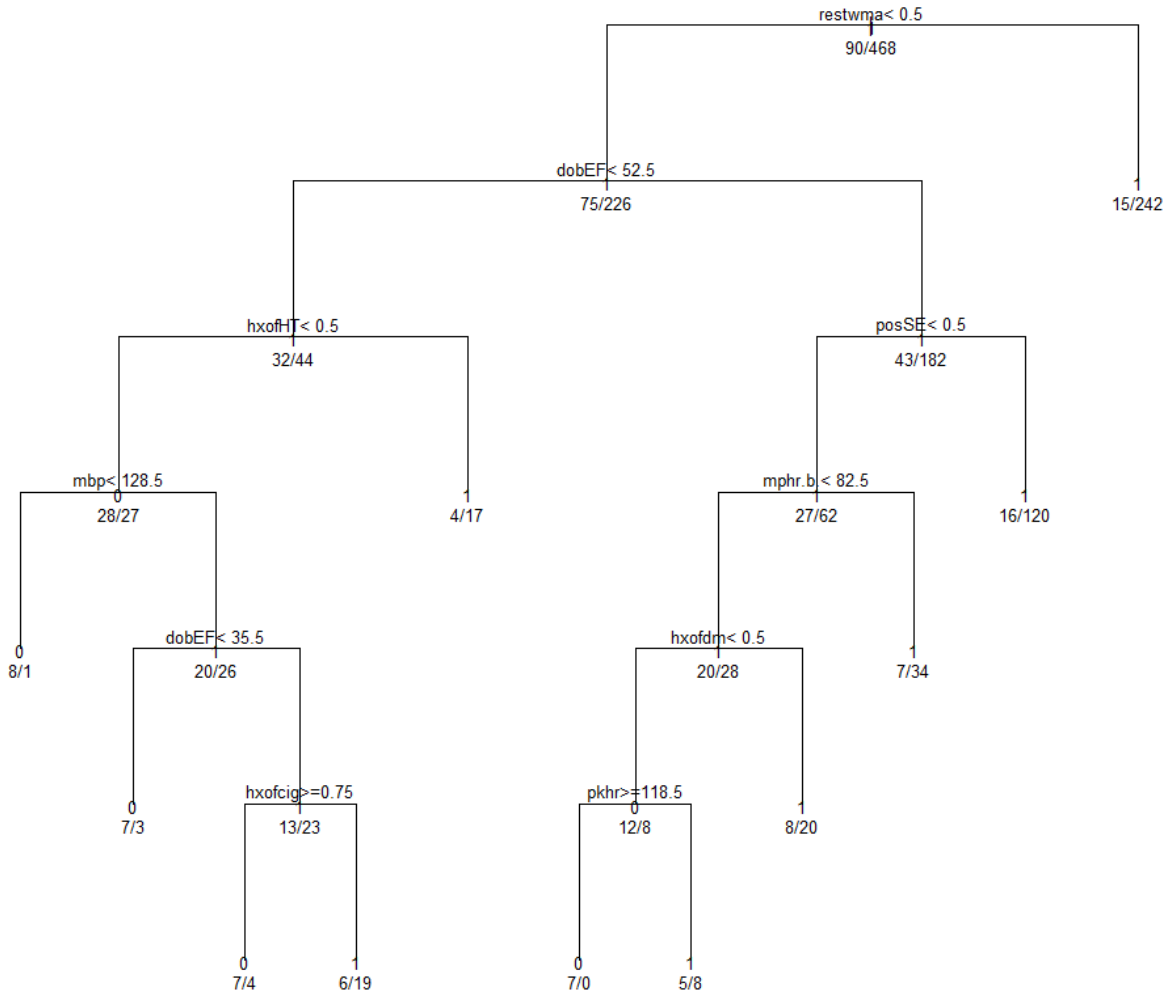
gdje je $R^{re}(\tau_l) = r(\tau_l)p(\tau_l), l = 1, 2, \dots, L.$

Drugim riječima, procjenom greške grupiranja zapravo vidimo koliko nam se razlikuju procijenjene vrijednosti od stvarnih.

Jedan od načina na koji se izmijenjena procjena $R^{re}(T)$, može prikazati kao procjena od $R(T)$ je da veća stabla imaju manje vrijednosti, tj. $R^{re}(T') \leq R^{re}(T)$, gdje je T' formirana dijeljenjem konačnih čvorova. Na primjer, ako se stablo može širiti na način

da svaki konačni čvor ima samo jednu promatranu vrijednost, tada je taj čvor klasificiran sa tom vrijednosti i $R^{re}(T)$=0 (vidi [1])

Primjer 3: Procjenu greške grupiranja na podacima koje koristimo u ovom radu dobijemo tako da sumiramo one podatke iz konačnih čvorova koji su krivo grupirani i podijelimo ih sa sveukupnim brojem podataka, tj. 558. To znači da gledamo one koji se nalaze u čvoru gdje dominira suprotna klasa.



Slika 5 Stablo sa prikazanom podjelom podataka po čvorovima

Iz slike 5 možemo uočiti da je stablo krivo grupiralo 8 jedinica od sveukupno 468 kao nule, a 61 nulu od njih 90 kao da su jedinice. Varijabla izlaza koju procjenjujemo ima vrijednost jedan ili nula, što je dodatno objašnjeno u poglavlju „Opis podataka“. Sumiranjem krivo grupiranih podataka i dijeljenjem sa ukupnom količinom dobijemo da je pogreška grupiranja stabla jednaka

$$R^{re}(T) = \frac{(8+61)}{558} = 0.12366 = 12.37\%$$

4.5. Skraćivanje stabla

Pravilo izgradnje konačnog stabla sastoji se od toga da prvo maksimalno proširimo stablo te mu onda skratimo grane dok ne dobijemo pravu veličinu.

Prema Izenman-u skraćeno stablo je zapravo pod stablo od originalnog stabla, a za njegov nastanak postoji više načina skraćivanja. U nastavku ćemo prikazati najbolje“ pod stablo koristeći procjenu od $R(T)$.

Algoritam koji koristimo je idući:

1. Napravimo veliko stablo, T_{max}, s maksimalnim skupom čvorova tako da dijelimo čvorove sve dok svaki sadrži manje od n_{min} promatranih vrijednosti
2. Izračunamo procjenu od $R(\tau)$ u svakom čvoru $\tau \in T_{max}$
3. Skraćujemo stablo T_{max} prema korijenskom čvoru tako da je u svakom koraku skraćivanja procjena od $R(T)$ minimalna.

Neka je $\alpha \geq 0$ parametar složenosti. Za bilo koji čvor $\tau \in T$, stavimo da je

$$R_\alpha(\tau) = R^{re}(\tau) + \alpha \tag{1.16}$$

iz čega definiramo mjeru posljedične složenosti skraćivanja stabla (eng. *cost-complexity pruning measure*) kao:

$$R_\alpha(T) = \sum_{l=1}^{L} R_\alpha(\tau_l) = R^{re}(T) + \alpha\left|\tilde{T}\right| \tag{1.17}$$

gdje je $\left|\tilde{T}\right| = L$ broj konačnih čvorova u pod stablu T od T_{max}. Gledajmo na $\alpha\left|\tilde{T}\right|$ kao na neki „kazneni“ izraz zbog veličine stabla, tako da $R_\alpha(T)$ može kazniti $R^{re}(T)$ za generiranje prevelikog stabla. Nadalje, za svaki α izaberemo ono pod stablo $T(\alpha)$ od T_{max} koje minimizira $R_\alpha(T)$

$$R_\alpha\big(T(\alpha)\big) = \min_T R_\alpha(T) \tag{1.18}$$

Ako $T(\alpha)$ zadovoljava prethodnu formulu, onda $T(\alpha)$ zovemo minimizirano pod stablo (tj. optimalno skraćeno pod stablo) od T_{max}. Za svaki α možemo pronaći više minimiziranih pod stabala od T_{max}.

Vrijednost α određuje veličinu stabla. Prema tome, kad je α vrlo mali, kaznena vrijednost od $\alpha\left|\tilde{T}\right|$ će biti također mala, dok će veličina minimiziranog pod stabla $T(\alpha)$ koji će biti određen sa $R^{re}\big(T(\alpha)\big)$ biti velika. Na primjer, ako stavimo da je $\alpha = 0$ i proširimo stablo T_{max} toliko veliko da svaki konačni čvor sadrži samo jednu vrijednost, tada svaki konačni čvor sadrži klasu jedne vrijednosti i $R^{re}(T_{max}) = 0$. Stoga, T_{max} minimizira $R_0(T)$. Povećanjem parametra složenosti α, minimizirano podstablo $T(\alpha)$ će imati sve manje konačnih čvorova. Za jako veliki α skraćujemo cijelo stablo T_{max} na samo korijenski čvor.

Kako doći od T_{max} do T_1? Pretpostavimo da čvor τ u stablu T_{max} ima konačne čvorove kćeri τ_L i τ_D. Tada je

$$R^{re}(\tau) \geq R^{re}(\tau_L) + R^{re}(\tau_D) \tag{1.19}$$

Primjer 4: Iz slike 5 vidim da čvor kojeg varijabla $pkhr$ dijeli na dvije kćeri ima 12 nule i 8 jedinica, lijeva kćer ima 7 nula i 0 jedinica, dok desna kćer ima 5 nula i 8 jedinica. Prema tome imamo $R^{re}(\tau) = 8/558 \geq R^{re}(\tau_L) + R^{re}(\tau_D) = (0+5)/558 = 5/558$

Ako je u prethodnoj formuli zadovoljena jednakost u čvoru τ, tada skraćujemo konačne čvorove τ_L i τ_D iz stabla. To radimo sve dok više ne postoji nijedno skraćivanje. Rezultat toga je stablo T_1.

Neka je τ bilo koji prolazni čvor od T_{max}. Neka je T_τ podstablo sa korijenskim čvorom τ, te neka je $\tilde{T}_\tau = \{\tau'_1, \tau'_2, \dots, \tau'_\tau\}$ skup konačnih čvorova podstabla T_τ. Tada je

$$R^{re}(T_\tau) = \sum_{\tau' \in \tilde{T}_\tau} R^{re}(\tau') = \sum_{l'=1}^{L_\tau} R^{re}(\tau'_{l'}) \tag{1.20}$$

Odnosno $R^{re}(\tau) > R^{re}(T_\tau)$.

Primjer 5: Iz slike 5 uzmimo da nam je prolazni čvor sa desne strane drugog dijeljenja stabla onaj koji ima podjelu stabla sa varijablom $posSE$. Taj čvor sadrži 43 nule i 182 jedinice. Neka je T_{posSE} podstablo sa tim konačnim čvorom. Vidimo da je formula zadovoljena: $R^{re}(posSE) = 43/558 > R^{re}\left(T_{posSE}\right) = (16 + 7 + 8 + 5 + 0)/558 = 36/558$.

Nadalje, neka je

$$R_\alpha(T_\tau) = R^{re}(T_\tau) + \alpha\left|\tilde{T}_\tau\right| \tag{1.21}$$

Sve dok je $R_\alpha(\tau) > R_\alpha(T_\tau)$, podstablo T_τ ima manju složenost nego njegov korijenski čvor τ pa se isplati zadržati T_τ. Iz primjera 5 zadržavamo T_{posSE} sve dok je $R^{re}_\alpha(\tau) = 43/558 + \alpha > 36/558 + 5\alpha = R^{re}_\alpha(T_\tau)$ ili $\alpha < 7/(558 * 4) = 0.0031362$

Uvrštavanjem formula (1.16) i (1.21) u prethodni uvjet dobivamo

$$\alpha < \frac{R^{re}(\tau) - R^{re}(T_\tau)}{\left|\tilde{T}_\tau\right| - 1} \tag{1.22}$$

Desna strana prethodne nejednakosti, koja je pozitivna, izračunava smanjenje R^{re} u odnosu na povećanje broja konačnih čvorova

Neka je $T_{max} = T_M$, gdje je M konačan prirodni broj.

Za $\tau \in T_M$ definiramo

$$g_M(\tau) = \frac{R^{re}(\tau) - R^{re}\left(T_{M,\tau}\right)}{\left|\tilde{T}_{M,\tau}\right| - 1}, \tau \notin \tilde{T}(\alpha_M) \tag{1.23}$$

gdje je $T_{M,\tau}$ jednak T_τ. Tada se $g_M(\tau)$ može razmatrati kao kritična vrijednost za α, sve dok je $g_M(\tau) > \alpha_M$, ali ne skraćujemo prolazne čvorova $\tau \in T_M$.

Definiramo najslabiji čvor $\tilde{\tau}_M$, kao čvor u T_M koji zadovoljava

$$g_M(\tilde{\tau}_M) = \min_{\tau \in T_M} g_M(\tau). \tag{1.24}$$

Dok se α povećava, $\tilde{\tau}_M$ je prvi čvor za koji je $R_\alpha(\tau) = R_\alpha(T_\tau)$, prema tome je $\tilde{\tau}_M$ poželjan za $T_{\tilde{\tau}_M}$. Postavimo da je $\alpha_{M-1} = g_M(\tilde{\tau}_M)$ i definiramo podstablo $T_{M-1} = T(\alpha_{M-1})$ iz T_M skraćivanjem podstabla $T_{\tilde{\tau}_M}$ od T_M tako da $\tilde{\tau}_M$ bude konačni čvor.

Za pronalazak T_{M-2} prvo tražimo najslabiji čvor $\tilde{\tau}_{M-1} \in T_{M-1}$ iz kritične vrijednosti

$$g_{M-1}(\tau) = \frac{R^{re}(\tau) - R^{re}\left(T_{M-1,\tau}\right)}{\left|\tilde{T}_{M-1,\tau}\right| - 1}, \tau \notin \tilde{T}(\alpha_{M-1}), \tau \in T(\alpha_{M-1}) \tag{1.25}$$

gdje je $T_{M-1,\tau}$ dio od T_τ koji se nalazi u T_{M-1}. Zatim postavimo

$$\alpha_{M-2} = g_{M-1}(\tilde{\tau}_{M-1}) = \min_{\tau \in T_{M-1}} g_{M-1}(\tau) \tag{1.26}$$

i definiramo pod stablo T_{M-2} iz T_{M-1} skraćivanjem podstabla $T_{\tilde{\tau}_{M-1}}$ od T_{M-1} tako da $\tilde{\tau}_{M-1}$ postane konačni čvor. Dalje radimo analogno konačan broj puta.

Kao što smo prethodno napomenuli, za svaki α možemo imati više minimiziranih podstabala. Problem se pojavljuje kad trebamo odlučiti koji od ponuđenih odabrati.

Za vrijednost parametra složenosti α, definiramo $T(\alpha)$ kao najmanje minimizirano stablo ako je uopće minimizirano stablo i ako zadovoljava sljedeći uvjet:

$$R_\alpha(\tau) = R_\alpha(T_\tau)\ tada\ je\ T \succ T(\alpha) \tag{1.27}$$

$T \succ T(\alpha)$ znači da je $T(\alpha)$ podstablo od T i ima manje konačnih čvorova nego T. Taj uvjet ujedno naglašava da je u bilo kojem slučaju $T(\alpha)$ najmanje stablo od svih stabala koji minimiziraju R_α. Prema tome za svaki α postoji jedinstveno najmanje minimizirano podstablo, što je dokazao L. Breiman u knjizi *„Classification and Regression Tree“*, (1984).

Gornja konstrukcija nam daje povećani slijed parametara složenosti,

$$0 = \alpha_M < \cdots < \alpha_1$$

koji odgovara konačnom slijedu uklopljenih pod stabala od T_{max}

$$T_{max} = T_M \succ \cdots \succ T_2 \succ T_1$$

gdje je $T_k = T(\alpha_k)$ jedinstveno najmanje minimizirano pod stablo za $\alpha \in [\alpha_k, \alpha_{k-1})$ i gdje je T_1 postablo koje sadrži samo korijenski čvor.

Počnemo sa T_M i povećavamo α sve dok $\alpha = \alpha_M$ odredi najslabiji čvor $\tilde{\tau}_M$, zatim skratimo podstablo $T_{\tilde{\tau}_M}$ sa tim čvorom kao korijenskim što nam daje T_{M-1}. Proceduru ponavljamo pronalaskom $\alpha = \alpha_{M-1}$ i najslabijeg čvora $\tilde{\tau}_{M-1}$ u T_{M-1}, te skraćivanjem pod stabla $T_{\tilde{\tau}_{M-1}}$ sa tim čvorom koji postaje korijen, iz čega dobivamo T_{M-2}. Taj proces skraćivanja ponavljamo sve dok ne dođemo do T_1. (vidi [1])

4.6. Zaustavljanje stabla

Lewis je u objasnio da se stablo nakon određenog broja koraka dijeljenja ipak treba zaustavit u nekom trenutku. Proces dijeljenja će biti zaustavljen: ako postoji samo jedan podatak u svakom čvoru djeteta, zatim ako svi podaci unutar svakog djeteta imaju identičnu distribuciju varijable koju predviđamo što onemogućava daljnje dijeljenje, te ako je korisnik ručno ograničio vanjski broj razina, odnosno dubinu maksimalnog stabla.

Stvoreno maksimalno stablo je često preistrenirano (eng. *overfit*), tj. stablo je osjetljivo u treniranom skupu za koji se smatra da se neće pojaviti u budućim nezavisnim grupama. Daljnja dijeljenja u stablu će najvjerojatnije biti preistrenirana nego ona od prije iako će jedan dio stabla možda trebati samo jednu ili dvije razine, dok druge strane stabla budu trebale više razina da bi uskladile pravu informaciju u skupu. (vidi [6])

4.7. Izbor najboljeg skraćenog stabla

U prethodnom dijelu smo pokazali kako konstruirati konačan niz smanjenih pod stabala $T_M, T_{M-1}, \ldots, T_2, T_1$ skraćivanjem čvorova od T_{max}. Postavlja se pitanje kada zaustaviti

skraćivanje i koje pod stablo iz priloženog niza odabrati kao najbolje skraćeno pod stablo.

Izenman tvrdi da izbor najboljeg takvog pod stabla ovisi o dobroj procjeni greške grupiranja $R(T_k)$ koja odgovara podstablu T_k. U knjizi *„Classification and Regression Tree“*, Breiman nudi procjenu greške korištenjem jedne od sljedećih dviju metoda. To su

- Nezavisnost testiranih podataka (eng. *independent test set*)
- Krosvalidacija (eng. *cross-validation*)

Ako je skup podataka s kojim radimo prevelik tada koristimo nezavisni test koji je jednostavan i učinkovito brz, te je ujedno najbolja metoda procjene. Za manji skup podataka koristimo krosvalidacijsku metodu. (vidi [1])

4.7.1. Nezavisnost testiranih podataka (eng. *independent test set*)

Prema [1] u skupu podataka $\mathcal{D}$ nasumično podijelimo vrijednosti podataka u skup za procjenu, tj. podataka za „učenje“ $\mathcal{L}$ i skup testiranih podataka $\mathcal{T}$, gdje je $\mathcal{D} = \mathcal{L} \cup \mathcal{T}$ i $\mathcal{L} \cap \mathcal{T} = \emptyset$. Pretpostavimo da postoji $n_{\mathcal{T}}$ vrijednosti u skupu testiranih podataka koji su izvučeni nezavisno iz iste temeljne distribucije kao učena opažanja.

Test se odvija na način da raširimo stablo T_{max} iz samo učenih podataka, skratimo ga od vrha da bismo dobili niz pod stabala $T_M \succ T_{M-1} \succ \cdots \succ T_1$, te zatim dodijelimo klasu svakom konačnom čvoru.

Zatim uzmemo svaku vrijednost iz skupa podataka za testiranje i ubacimo je u pod stablo T_k. Tada je svakoj vrijednosti u $\mathcal{T}$ pridružena jedna klasa od postojećih. Prema tome kako je sada klasa svake vrijednosti iz skupa $\mathcal{T}$ poznata, procjenjujemo $R(T_k)$ po $R^{ts}(T_k)$ formulom (1.21) za $\alpha = 0$, pa imamo $R^{ts}(T_k) = R^{re}(T_k)$ da je izmijenjena procjena dobivena korištenjem nezavisnih test podataka. Kada su posljedice grešaka identične kod svake klase, $R^{ts}(T_k)$ je proporcija svih testiranih vrijednosti koje su

pogrešno grupirane od strane T_k. Zatim se te procjene koriste za izbor najbolje skraćenog pod stabla T_* formulom

$$R^{ts}(T_*) = \min_k R^{ts}(T_k) \tag{1.28}$$

gdje je $R^{ts}(T_*)$ procijenjena greška grupiranja testiranih podataka.

Standardnu pogrešku od $R^{ts}(T)$ procjenjujemo tako da kada provučemo skup testiranih podataka kroz stablo T, vjerojatnost da smo krivo grupirali neko opažanje jest $p^* = R(T)$. Dakle imamo binomnu razdiobu uzorkovanja sa $n_{\mathcal{T}}$ Bernoulli-evih pokušaja i vjerojatnost uspjeha p^*. Ako je $p = R^{ts}(T)$ proporcija krivih grupiranja opažanja iz skupa $\mathcal{T}$, tada je p nepristran procjenitelj za p^* i varijanca od p je $p^*(1-p^*)/n_{\mathcal{T}}$. Zatim standardnu grešku od $R^{ts}(T)$ procjenjujemo formulom

$$\widehat{SE}\left(R^{ts}(T)\right) = \left\{\frac{R^{ts}(T)\,(1-R^{ts}(T))}{n_{\mathcal{T}}}\right\}^{\frac{1}{2}} \tag{1.29}$$

4.7.2. Krosvalidacija (eng. *cross-validation*)

Kod V-strukih krosvalidacija (*eng. V-fold cross-validation*) nasumično podijelimo podatke $\mathcal{D}$ u V disjunktnih pod skupova podjednake veličine takvih da je $\mathcal{D} = \bigcup_{v=1}^{V} \mathcal{D}_v$ gdje je $\mathcal{D}_v \cap \mathcal{D}_{v'} = \emptyset$ za $v \neq v'$, s popriličnom veličinom V od 5 do 10. Zatim formiramo V različitih skupova iz $\{\mathcal{D}_v\}$ koristeći $\mathcal{L}_v = \mathcal{D} - \mathcal{D}_v$ kao v-ti skup podataka za treniranje i $\mathcal{T}_v = \mathcal{D}_v$ kao v-ti skup podataka za testiranje, za $v = 1, \dots, V$. Ako skupovi $\{\mathcal{D}_v\}$ imaju isti broj opažanja, tada će svaki testirani skup imati $\left(\frac{V-1}{V}\right) \times 100$ vjerojatnost originalnog skupa podataka. Podaci iz $\mathcal{D}_v, v = 1, \dots, V$ su podaci za validaciju modela.

Dalje proširimo v-to stablo $T_{max}^{(v)}$ koristeći v-ti skup podataka za učenje, tj. treniranje $\mathcal{L}_v, v = 1, \dots, V$, te fiksiramo parametar složenosti α. Neka je $T^{(v)}(\alpha)$ najbolje skraćeno stablo od $T_{max}^{(v)}, v = 1, \dots, V$. Zatim svako opažanje u v-tom skupu testiranih podataka $\mathcal{T}_v$ provučemo kroz stablo $T^{(v)}(\alpha), v = 1, \dots, V$. Neka $n_{ij}^{(v)}(\alpha)$ označava broj j-te grupe opažanja testiranih podataka $\mathcal{T}_v$ koje su grupirane kao da su iz i-te grupe, za $i, j =$

$1, 2, \dots, K, v = 1, \dots, V$. Kako je $\mathcal{D} = \bigcup_{v=1}^{V} \mathcal{T}_v$ disjunktna suma, tada je ukupan broj opažanja iz j-tih grupa koje su grupirane kao da su iz i-te jednak $n_{ij}(\alpha) = \sum_{v=1}^{V} n_{ij}^{(v)}(\alpha)$, $i, j = 1, 2, \dots, K$. Ako stavimo da je broj opažanja u skupu $\mathcal{D}$ koji pripada j-toj grupi jednak $n_j, j = 1, 2, \dots, K$ i pretpostavimo da je greška grupiranja jednaka za sve grupe, tada za dani α imamo da je

$$R^{CV/V}(T(\alpha)) = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{K} n_{ij}(\alpha) \tag{1.30}$$

procijenjena greška grupiranja nad $\mathcal{D}$, gdje $T(\alpha)$ minimizira podstablo T_{max}.

Posljednji korak je pronaći pod stablo prave veličine. Breiman je predložio ocjenjivanje formule (1.22) sa nizom vrijednosti $\alpha'_k = \sqrt{\alpha_k \alpha_{k-1}}$, gdje je α'_k geometrijska sredina intervala $[\alpha_k, \alpha_{k-1})$ u kojem je $T(\alpha) = T_k$. Stavimo

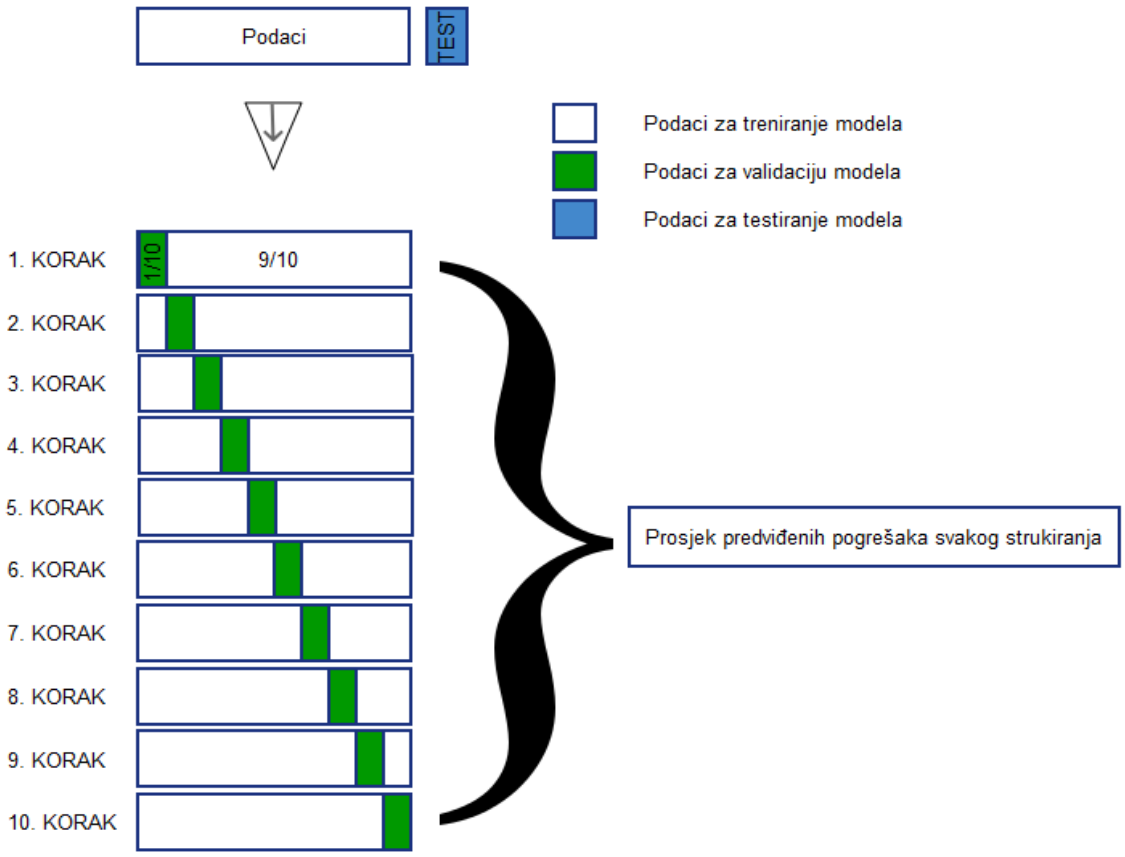
$$R^{CV/V}(T_k) = R^{CV/V}(T(\alpha'_k)) \tag{1.31}$$

zatim izaberemo najbolje skraćeno pod stablo T_* pravilom

$$R^{CV/V}(T_*) = \min_k R^{CV/V}(T_k) \tag{1.32}$$

i koristimo $R^{CV/V}(T_*)$ kao njenu procijenjenu grešku grupiranja. Izvođenje procjene standardne pogreške, iz krosvalidacijske procjene iz greške grupiranja, je naime dosta komplicirano.

Najčešći način problema zastranjivanja ne-nezavisnosti sumanada iz (1.27) je da se ignorira ne-nezavisnost te da se umjesto toga uzima da su nezavisne, što u praksi odlično djeluje, gdje se najčešće uzima $V = 10$. Metoda jednog izostavljanja (eng. *the leave-one-out*) nije preporučljiva jer će rezultat pomoćnih stabala biti identičan stablu konstruiranom iz cijelog skupa podataka, pa ništa ne bi dobili iz tog postupka. (vidi [1])



Slika 6 10-struka krosvalidacija

4.8. Pravilo procijene pogreške

Kako bi se prevladale moguće nestabilnosti u odabiru najboljeg skraćenog pod stabla, Breiman predlaže alternativni način.

Neka $\hat{R}(T_*) = \min_k R(T_k)$ označava procijenjenu grešku grupiranja dobivenu iz nezavisnog testa podataka ili krosvalidacije. Tada biramo najmanje stablo T_{**} koje zadovoljava „1-SE“ uvjet:

$$\hat{R}(T_{**}) \leq \hat{R}(T_*) + \widehat{SE}\left(\hat{R}(T_*)\right) \tag{1.33}$$

Ovo pravilo daje bolje pod stablo jer kroz standardnu grešku odgovara na promjenjivost procjene krosvalidacije. (vidi [1])

5. Regresijska stabla

Već od prije smo upoznati sa linearnom regresijom čija je ideja izrada predviđanja. Tu je nezavisna varijabla *Y* modelirana kao linearna funkcija nezavisnih varijabli. Linearna regresija je globalni model sa jednom formulom koja sadrži cijeli skup podataka. Međutim, kada podaci imaju više značajki koje međusobno djeluju na ne-linearan način, tada je sastavljanje modela iznimno teško i komplicirano za objašnjenje rezultata. Da bi se olakšao način modeliranja podataka koriste se particije koje dijele podatke u manje dijelove s kojima se radi i takav se način modeliranja zove rekurzivna particija (eng. *recursive partitioning*).

U svom radu Izenman je dao objašnjenje algoritma za izradu modela korištenjem regresijskih stabala. Neka su nam podaci zadani u obliku $\mathcal{D} = \{(\boldsymbol{X}_i, Y_i), i = 1, 2, \dots, n\}$, gdje su Y_i mjerenja napravljena na kontinuiranoj varijabli Y, a $\boldsymbol{X}_i$ opažena mjerenja r-dimenzionalnog vektora $\boldsymbol{X}$. Pretpostavljamo da je Y u relaciji sa $\boldsymbol{X}$ višestrukom regresijom i da želimo koristiti metodu stabla za predviđanje varijable Y u odnosu na $\boldsymbol{X}$.

Konstrukcija regresijskog stabla slična je izgradnji klasifikacijskog i koristi se po dijelovima rekurzivna regresija. Dok je kod klasifikacijskih stabala, klasa konačnog čvora definirana kao klasa od više uvjeta svih opažanja u tom čvoru, gdje su veze nasumično odlučene, kod regresijskog stabla izlazna varijabla poprima konstantnu vrijednost $Y(\tau)$ u konačnom čvoru τ. Stoga se stablo može prikazati kao r-dimenzionalni histogram procijenjen površinskom regresijom, gdje je r broj ulaznih varijabli, $X_1, X_2, \dots, X_r$.

5.1. Vrijednost konačnog čvora

Kao što smo napomenuli u prethodnom dijelu, metodom regresijskog stabla želimo procijeniti vrijednost Y. Postavlja se pitanje kako pronaći $Y(\tau)$?

Znamo da je izmijenjena procjena predviđene pogreške jednaka (eng. *resubstitution estimate of prediction error*)

$$R^{re}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 \tag{2.1}$$

gdje je $\widehat{Y}_i = \hat{\mu}(\boldsymbol{X}_i)$ procijenjena vrijednost predviđene pogreške u $\boldsymbol{X}_i$. Da bi $\widehat{Y}_i$ bila konstanta u svakom čvoru ta pogreška treba biti oblika

$$\hat{\mu}(\boldsymbol{X}) = \sum_{\tau \epsilon \widehat{T}} Y(\tau)\mathbb{I}_{[\boldsymbol{X}\in\tau]} = \sum_{\boldsymbol{l=1}}^{\boldsymbol{L}} Y(\tau_l)\mathbb{I}_{[\boldsymbol{X}\in\tau_l]} \tag{2.2}$$

gdje je $\mathbb{I}_{[\boldsymbol{X}\in\tau_l]}$ jedinična funkcija koja poprima vrijednost jedan ako je $\boldsymbol{X} \in \tau_l$, inače nula.

Za $\boldsymbol{X_i} \in \tau_l$, $R^{re}(\hat{\mu})$ je minimalna ako uzmemo da je $\widehat{Y}_i = \bar{Y}(\tau_l)$ konstantna i poprima vrijednost prosjeka $\bar{Y}(\tau_l)$ od $\{Y_i\}$ za sve promatrane vrijednosti pridružene čvoru τ_l, odnosno

$$\bar{Y}(\tau_l) = \frac{1}{n(\tau_l)}\sum_{\boldsymbol{X_i}\in\tau_l} Y_i \tag{2.3}$$

gdje je $n(\tau_l)$ broj promatranih varijabli u čvoru τ_l, $l = 1, 2, \dots, L$.

Nadalje, izmijenjena procjena kod regresijskog stabla je oblika:

$$R^{re}(T) = \frac{1}{n}\sum_{l=1}^{L}\sum_{\boldsymbol{X_i}\in\tau_l}(Y_i - \bar{Y}(\tau_l))^2 = \sum_{l=1}^{L} R^{re}(\tau_l) \tag{2.4}$$

$$R^{re}(\tau_l) = \frac{1}{n}\sum_{\boldsymbol{X_i}\in\tau_l}(Y_i - \bar{Y}(\tau_l))^2 = p(\tau_l)s^2(\tau_l) \tag{2.5}$$

gdje je $s^2(\tau_l)$ pristrani uzorak varijanci od svih vrijednosti Y_i u čvoru τ_l, te $p(\tau_l) = n(\tau_l)/n$ omjer opaženih vrijednost u čvoru τ_l. Prema tome vrijedi $R^{re}(T) = p(\tau_l)s^2(\tau_l)$.

5.2. Strategija dijeljenja čvora

Slično kao i kod klasifikacijskih stabala, trebamo odrediti način dijeljenja svakog prolaznog čvora u stablu. U čvoru $\tau \epsilon \widehat{T}$ odaberemo dijeljenje koje daje najveće

smanjenje u izmijenjenoj procjeni $R^{re}(T)$. Zbog podjele τ na čvorove kćeri τ_L i τ_D, reduciranje od $R^{re}(\tau)$ je dano sa

$$\Delta R^{re}(\tau) = R^{re}(\tau) - R^{re}(\tau_L) - R^{re}(\tau_D) \tag{2.6}$$

najboljim dijeljenjem u onom τ koji maksimizira $\Delta R^{re}(\tau)$. Rezultat takve podjele jest da dobivena najbolja podjela razvrsta promatrane vrijednosti ovisno o tome da li Y ima manju ili veću vrijednost.

Pronalazak takvih čvorova τ_L i τ_D je ekvivalentno minimiziranju izraza $R^{re}(\tau_L) + R^{re}(\tau_D)$. Zatim formulu $R^{re}(\tau_l) = p(\tau_l)s^2(\tau_l)$ svodimo na pronalazak tih čvorova da bi odredili

$$\min_{\tau_L,\tau_D}\{ p(\tau_L)s^2(\tau_L) + p(\tau_D)s^2(\tau_D)\} \tag{2.7}$$

gdje su $p(\tau_L)$ i $p(\tau_D)$ omjeri promatranih vrijednosti u čvoru τ.

5.3. Skraćivanje stabla

Ideja skraćivanja regresijskog stabla je analogna skraćivanju klasifikacijskog. Vodi se metodom da prvo proširimo stablo na T_{max}, na način da dijelimo čvorove sve dok svaki čvor sadrži manje podataka od promatranih, tj. sve dok je $n_\tau \leq n_{min}$ za svaki čvor $\mathcal{T}$, gdje najčešće stavljamo da je $n_{min} = 5$.

Zatim definiramo mjeru složenosti kao:

$$R_\alpha(T) = R^{re}(T) + \alpha\left|\tilde{T}\right| \tag{2.8}$$

gdje je $\alpha > 0$ parametar složenosti.

Koristimo $R_\alpha(T)$ kao kriterij za dijeljenje čvora u smislu kada i kako ćemo dijeliti, te kao rezultat dobijemo niz podstabala

$$T_{max} = T_M \succ T_{M-1} \succ \cdots \succ T_1$$

i povezani niz parametara složenosti

$$0 = \alpha_M < \alpha_{M-1} < \cdots < \alpha_1$$

gdje je je za $\alpha \in [\alpha_k, \alpha_{k-1})$, T_k najmanje minimizirano stablo od T_{max}.

5.4. Izbor najboljeg skraćenog stabla

Izbor najboljeg stabla svodi se na procjenu $R(T_k)$ koristeći test nezavisnosti ili krosvalidaciju čiji su algoritmi prikazani u prethodnoj cjelini.

Koristeći nezavisni skup testiranih podataka, $\mathcal{T}$, procjena od $R(T_k)$ je dana sljedećom formulom

$$R^{ts}(T_k) = \frac{1}{n_{\mathcal{T}}} \sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{T}} (Y_i - \hat{\mu}_k(\boldsymbol{X}_i))^2 \tag{2.9}$$

gdje je $n_{\mathcal{T}}$ broj promatranih podataka u testiranom skupu i $\hat{\mu}_k(\boldsymbol{X}_i)$ procijenjena funkcija predviđanja povezana sa pod stablom T_k.

Nadalje, kod procijene od $R(T_k)$ koristeći V-struke krosvalidacije, prvo konstruiramo minimalnu grešku složenosti pod stabala $T^{(v)}(\alpha), v = 1, 2, , \dots, V$ sa parametrom α. Zatim postavimo $\alpha'_k = \sqrt{\alpha_k \alpha_{k-1}}$ i neka nam $\hat{\mu}_k^{(v)}(\boldsymbol{X})$ označava procjenjenu funkciju predviđanja povezanu sa podstablom $T^{(v)}(\alpha'_k)$. Procijena od $R(T_k)$ je dana formulom

$$R^{CV/V}(T_k) = \frac{1}{n} \sum_{v=1}^{V} \sum_{(\boldsymbol{X}_i, Y_i) \in \mathcal{T}^v} (Y_i - \hat{\mu}_k^{(v)}(\boldsymbol{X}_i))^2 \tag{2.10}$$

U primjeni se najčešće uzima $V = 10$ za procjenu krosvalidacije gdje dijelimo skup podataka za procijenu u 10 pod skupova. Zatim koristimo 9 tih pod skupova za širenje i skraćivanje stabla, te zatim koristimo izostavljene podatke za testiranje rezultata stabla.

Za dani niz pod stabala $\{T_k\}$, odaberemo najmanje podstablo T_{**} za koje vrijedi

$$\hat{R}(T_{**}) \leq \hat{R}(T_*) + \widehat{SE}\left(\hat{R}(T_*)\right) \tag{2.11}$$

gdje je $\hat{R}(T_*) = \min_k \hat{R}(T_k)$ procijenjena greška predviđanja dobivena korištenjem nezavisnih testiranih podataka $\left(R^{ts}(T_*)\right)$ ili krosvalidacije $\left(R^{CV/V}(T_*)\right)$. (vidi [1])

6. Dodatni pristupi

6.1. Višestruki pristupi

Dosta primijenjenih radova je provedeno na konstrukciji klasifikacijskih stabala kod višestrukog pristupa, pogotovo kod onih gdje je varijabla izlaza binarna. U takvim primjerima, mjera homogenosti u čvoru τ za binarnu varijablu je generalizirana funkcijom skalarne vrijednosti matričnog argumenta. Najčešće se koristi $-\log|\boldsymbol{V}_\tau|$, gdje je $\boldsymbol{V}_\tau$ primjer kovarijacijske matrice unutar čvora sa m binarnih odgovora u čvoru τ i izvorni čvor kvadratnog oblika u $\boldsymbol{V}$, gdje je kovarijacijska matrica izvedena iz korijenskog čvora.

Složenost stabla T je tada definirana kao

$$R_\alpha(T) = \sum_{l=1}^{L} R_\alpha(\tau_l) = R^{re}(T) + \alpha\left|\tilde{T}\right| \tag{3.1}$$

gdje je $R^{re}(T)$ mjera homogenosti unutar čvora sumirana preko svih konačnih čvorova.

Kad se radi sa višestrukim pristupom, iz dosadašnje primjene se jasno vidi da dostupna količina podataka za konstrukciju stabla treba biti znatno velika. (vidi [1])

6.2. Stabla doživljenja (eng. survival trees)

Metode bazirane na stablima za analizu cenzuriranih preživjelih podataka su postale vrlo korisni alati u biomedicinskim istraživanjima pomoću kojih određuju prognostičke faktore za predviđena preživljavanja. Takva stabla zovemo stabla doživljenja (*eng. survival trees*). Preživljeni podaci koje koristimo se najčešće gledaju kao smrtni podaci, ali se ujedno mogu generalizirati kao vrijeme dešavanja nekoj događaja. Cenzurirane podatke doživljenja dobijemo kada pacijent doživi kraj promatranog perioda kojeg proučavamo, napusti promatranje prerano ili tijekom promatranja umre od bolesti koja nije vezana za promatrani studij.

Naime, tijekom korištenja takve metode za analizu, potrebno je odabrati kriterij kojim bi dijelili donesene odluke. Postoji više takvih kriterija i oni se mogu podijeliti u dva tipa ovisno o tome dali se opredijelimo na korištenje mjerenja homogenosti unutar čvora ili između čvorova. Mnoge primjene takvih metoda, parametarski bazirane, obično sadrže negativni oblik log-likelihood funkcije gubitka, koji se razlikuju u korištenju spomenute funkcije, te po tome kako opisuju model za vjerojatnost promatranih podataka unutar čvorova. (vidi [1])

6.3. Višestruki prilagodljivi regresijski splajnovi – MARS (eng. - multivariate adaptive *regression splines*)

Rekurzivna podjela koja se koristi u konstrukciji regresijskih stabala je generalizirana na fleksibilne klase ne-parametarskog regresijskog modela kojeg zovemo MARS (eng. – *multivariate adaptive regression splines*).

U MARS metodi, Y je povezan sa $\boldsymbol{X}$ modelom $Y = \mu(\boldsymbol{X}) + \epsilon$, gdje greška ϵ ima očekivanje nula. Regresijska funkcija, $\mu(\boldsymbol{X})$, je težinski zbroj od L osnovnih funkcija, odnosno

$$\mu(\boldsymbol{X}) = \beta_0 + \sum_{l=1}^{L} \beta_l B_l(\boldsymbol{X}) \tag{3.2}$$

gdje je B_l, l -ta osnovna funkcija, a

$$B_l(\boldsymbol{X}) = \prod_{m=1}^{M_l} \Phi_{lm}(X_{q(l,m)}) \tag{3.3}$$

produkt od M_l glatkih funkcija $\{\Phi_{lm}(X)\}$, gdje je M_l konačan broj i $q(l,m)$ indeks koji ovisi o l-toj osnovnoj funkciji i m-toj savitljivoj funkciji. Prema tome, za svaki l, $B_l(\boldsymbol{X})$ može sadržavati jednu savitljivu funkciju ili može biti produkt od dvije ili više takvih funkcija. Ujedno se nijedna ulazna varijabla ne smije pojavljivati više od jednom u produktu. Tako savitljive funkcije su najčešće linearnog oblika

$$\Phi_{lm}(X) = (X - \mathrm{t}_{lm})_+, \Phi_{l+1,m}(X) = (\mathrm{t}_{lm} - X)_+ \tag{3.4}$$

gdje je t_{lm} čvor od $\Phi_{lm}(X)$ koji se pojavljuje na jednoj od promatranih vrijednosti od $X_{q(l,m)}, m = 1, 2, \dots, M_l, l = 1, 2, \dots, L$. U prethodnoj funkciji vrijedi $(x)_+ = \max(0, x)$. Ako vrijedi $B_l(\boldsymbol{X}) = \mathbb{I}_{[\boldsymbol{X} \in \tau_l]}$ i $\beta_l = Y(\tau_l)$ tada je regresijska funkcija (3.2) ekvivalentna predviđenoj vrijednosti regresijskog stabla u formuli (2.2).

Proces započinje unošenjem slobodnog koeficijenta $\beta_0, (B_0(\boldsymbol{X}) = 1)$ u model, te zatim u svakom sljedećem koraku dodajemo jedan par uvjeta oblika kao što su u (3.4), odnosno izabiremo ulaznu varijablu i čvor minimiziranjem kriterija za grešku sume kvadrata

$$ESS(L) = \sum_{i=1}^{n} (y_i - \mu_L(x_i))^2 \tag{3.5}$$

gdje za dani L, $\mu_L(x_i)$ evoluira u $\boldsymbol{X} = x_i$. (vidi [1])

6.4. Bagging i Boosting

Bootstrap agregacija ili baging je tehnika koju je predložio Breiman i dobivena je od kombinacije engleskih riječi *Bootstrap* i *aggregating*. Tehnika se bavi stvaranjem višestrukih sličnih skupova podataka, ponavljanjem CART analize na svakom skupu, zatim sakuplja dobivene rezultata, preračunava stablo te povezuje i uspoređuje statistike dobivene sakupljenim rezultatima. Ova se tehnika inače koristi kao krosvalidacijska metoda za veća stabla koje korisnik želi skratiti gdje različite verzije istog stabla imaju različite greške grupiranja. Općenito, ova tehnika poboljšava rezultate nestabilnih stabala, ali ujedno može smanjiti svojstva stabilnih. Korištenjem procedure „ipred“ u R-u ova se tehnika lako implementira.

Boosting ima sličan pristup kao i bagging. Ona smanjuje grešku grupiranja rekurzivnim modelom CART analize. Kod ove tehnike klasifikatori se iterativno stvaraju težinskim vrijednostima primjera, gdje se težine prilagođavaju u svakoj iteraciji baziranoj na slučajevima pogreške grupiranja u prethodnom koraku. Ova se tehnika najčešće koristi na podacima koji imaju veliku grešku grupiranja jer su ne poučni, tj. rezultati su slabe

značajnosti i podaci su slabo povezani. Boosting se također može implementirati kroz neke procedure. (vidi [3])

6.5. Nedostajući podaci (eng. missing values)

Kod klasifikacijskih i regresijskih statističkih analiza javljaju se problemi sa vrijednostima koje nedostaju u testiranom skupu podataka, ali srećom postoji nekoliko načina kako riješiti problem nedostatka kad se koristi metoda stabla.

Jedan od mogućih načina jest da ubacimo skup podataka s nedostajućim vrijednostima u stablo konstruirano od cjelovitog skupa podataka i gledamo koliko daleko ide. Ako varijabla s nedostajućom vrijednosti nije sudjelovala u izgradnji stabla, tada će podaci koje smo ubacili upasti u odgovarajući konačni čvor, te zatim možemo klasificirati podatak ili predvidjeti njegovu vrijednost Y. Međutim, ako podatak ne može upasti nigdje dalje od posebnog prolaznog čvora τ, jer dijeljenje kod τ uključuje upravo tu varijablu sa nedostajućom vrijednosti, možemo zaustaviti podatak u čvoru τ ili prisiliti sve podatke sa nedostajućim vrijednostima za tu varijablu da upadnu u isti dječji čvor.

Da bi bolje modelirali i izgradili stablo sa nedostajućim podacima Breiman u svojoj knjizi CART, 1984 je predložio metodu zamjene podjele (eng. *surrogate splits*). Ideja metode je da u danom čvoru τ koristimo onu varijablu koja najbolje predviđa željenu podjelu kao zamjensku varijablu na kojoj se treba dijeliti čvor τ. Ako varijabla koja najbolje dijeli čvor τ sadrži nedostajuću vrijednost u podatku koji puštamo u konstruirano stablo, tada koristimo zamjensku varijablu, tj. iduću onu koja najbolje dijeli čvor pretpostavljajući da ta zamjenska varijabla sadrži cjelovite vrijednosti.

Ako se nedostajući podatak pojavi za unosnu varijablu sa L razina, tada uvodimo dodatni nivo „nedostajući“ ili „NA“. Tada varijabla ima $L + 1$ razina. (vidi [1])

6.6. Softverski paketi

Postoji nekoliko raznovrsnih softvera pomoću kojih možemo riješiti problem klasifikacijskih i regresijskih stabala odlučivanja. Najčešći softveri koji se danas koriste i svima su dostupni su: S-plus i R program koji se ujedno koristi u ovom radu za prikaz primjera i grafički prikaz stabla. Alternativni softveri koji se također koriste za problem klasifikacijskih i regresijskih stabala su: SAS Enterprise Miner, SPSS Classification Trees, Statistica i Systat. Međutim, implementacija tih programa nije ista, neke od razlika koje se mogu uočiti su: funkcija nečistoće (funkcija entropije), kriterij dijeljenja, pravilo zaustavljanja.

Treba napomenuti da se prikaz koda u navedenim programima različito prikazuje. Kod većine programa je tekstualni zapis, dok je kod ostalih grafički prikaz spojenih uvjeta i pravila.

7. Primjena CART analize u medicini

7.1. R Studio

U izradi ovog diplomskog rada korišten je R Studio, integrirani razvojni dio (IDE) R softvera. (vidi [5])

R je besplatni softver namijenjen za statističko računarstvo i grafički prikaz kojeg koriste statistički stručnjaci i studenti diljem svijeta. R Studio uključuje konzolu, urednik sa istaknutom sintaksom koja podržava direktno izvršavanje koda, alate za crtanje, otklanjanje grašaka i radni prostor za upravljanje. Osim direktnog programiranja naredbi u radnom prostoru, korisnik se može pozvati na već dostupne pakete i projekte koji sadrže razne gotove visoko provjerene funkcije koje su napisali vrhunski stručnjaci iz R Studio tima. U ovom se radu koriste paketi „RPART“ i „RATTLE“ za bolji rezultat i prikaz podataka.

Iz gore navedenih paketa koriste se „rpart“ i „fancyRpartPlot“ funkcije. Obje funkcije služe za prikaz klasifikacijskog i regresijskog stabla, a razlikuju se u samom prikazu rezultata. „Rpart“ vraća klasično stablo gdje se u svakom čvoru nalazi numerički zapis grupiranih podataka, dok „fancyRpartPlot“ vraća zapis u postotku, tj. s kolikom će vjerojatnosti pacijent upasti u određenu klasu ovisno o zadovoljenim uvjetima u prethodnim čvorovima, te ujedno daje ljepši i snalažljiviji prikaz stabla.

- rpart(formula, data, weights, subset, na.action = na.rpart, method, model = FALSE, x = FALSE, y = TRUE, parms, control, cost, ...)

- fancyRpartPlot(model, main="", sub, palettes, ...)

7.2. Rpart (eng. Recursion partitioning and Regression Trees)

```
1. X<-read.csv("podaci.csv",sep=";")
2. head(X)
```

	bhr	basebp	basedp	pkhr	sbp	dp	dose	maxhr	mphr.b.	mbp	dpmaxd	dobdose
1	92	103	9476	114	86	9804	40	100	74	121	12100	40
2	62	139	8618	120	158	18960	40	120	82	158	18960	40
3	62	139	8618	120	157	18840	40	120	82	157	18840	40
4	93	118	10974	118	105	12390	30	118	72	105	12390	30
5	89	103	9167	129	173	22317	40	129	69	176	22704	40

	age	gender	baseEF	dobEF	chestpain	posECG	equivecg	restwma	posSE	hxofHT
1	85	0	27	32	1	1	1	0	1	1
2	73	0	39	40	1	1	0	0	1	1
3	73	0	39	40	1	1	0	0	1	1
4	57	1	42	57	1	1	1	0	1	1
5	34	0	45	57	0	1	0	1	1	1

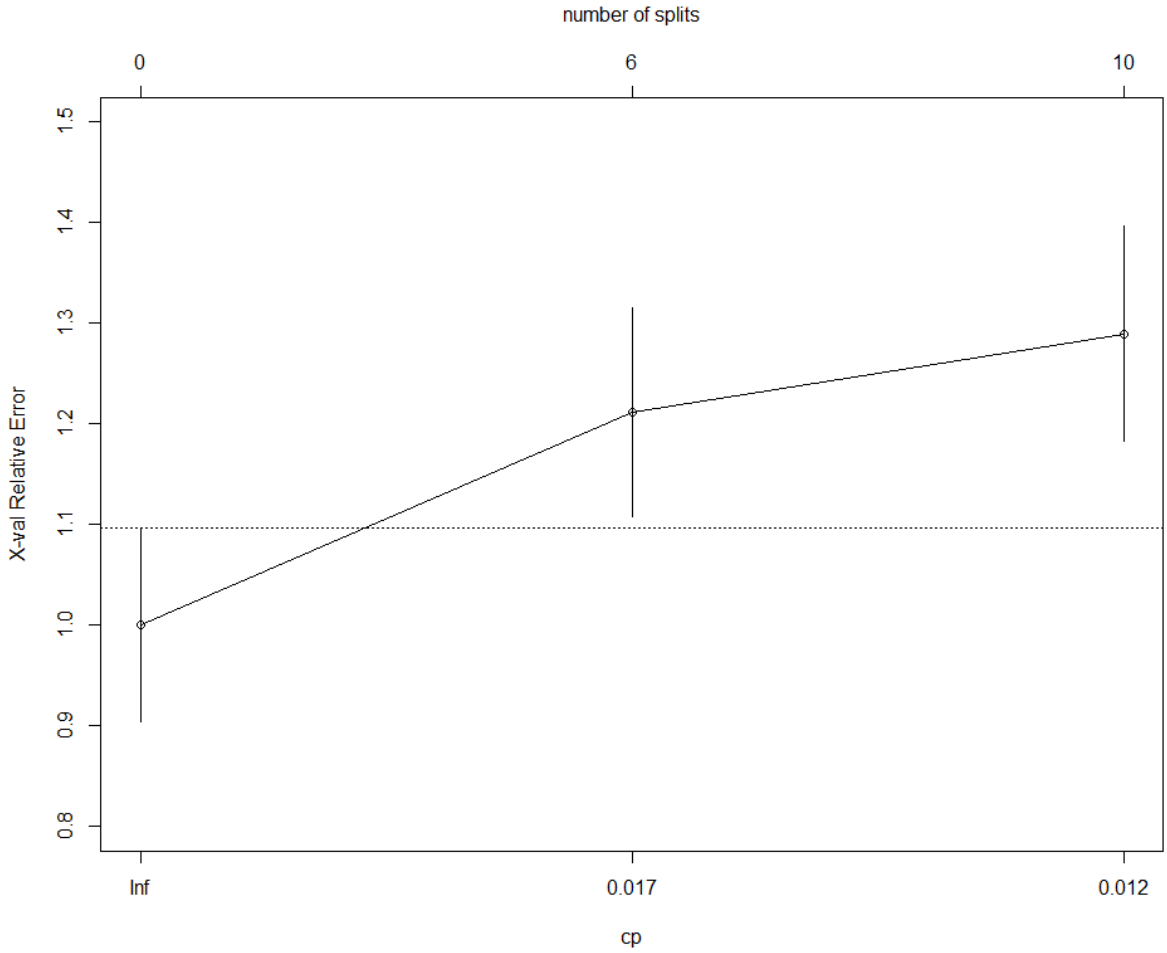
	hxofdm	hxofcig	hxofMI	hxofPTCA	hxofCABG	anyy
1	1	1	0	0	1	1
2	0	1	0	1	1	0
3	0	1	0	1	1	0
4	1	1	1	1	1	1
5	1	1	1	1	1	1

```
3. stablo<-rpart(any~.,data=X,parms = list(split='information'),method=
c("class“) ,control = rpart.control(xval=10))
n= 558

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

 1) root 558 90 1 (0.16129032 0.83870968)
   2) restwma< 0.5 301 75 1 (0.24916944 0.75083056)
     4) dobEF< 52.5 76 32 1 (0.42105263 0.57894737)
       8) hxofHT< 0.5 55 27 0 (0.50909091 0.49090909)
        16) mbp< 128.5 9  1 0 (0.88888889 0.11111111) *
```

```
        17) mbp>=128.5 46 20 1 (0.43478261 0.56521739)
          34) dobEF< 35.5 10  3 0 (0.70000000 0.30000000) *
          35) dobEF>=35.5 36 13 1 (0.36111111 0.63888889)
            70) hxofcig>=0.75 11  4 0 (0.63636364 0.36363636) *
            71) hxofcig< 0.75 25  6 1 (0.24000000 0.76000000) *
       9) hxofHT>=0.5 21  4 1 (0.19047619 0.80952381) *
     5) dobEF>=52.5 225 43 1 (0.19111111 0.80888889)
      10) posSE< 0.5 89 27 1 (0.30337079 0.69662921)
        20) mphr.b.< 82.5 48 20 1 (0.41666667 0.58333333)
          40) hxofdm< 0.5 20  8 0 (0.60000000 0.40000000)
            80) pkhr>=118.5 7  0 0 (1.00000000 0.00000000) *
            81) pkhr< 118.5 13  5 1 (0.38461538 0.61538462) *
          41) hxofdm>=0.5 28  8 1 (0.28571429 0.71428571) *
        21) mphr.b.>=82.5 41  7 1 (0.17073171 0.82926829) *
      11) posSE>=0.5 136 16 1 (0.11764706 0.88235294) *
   3) restwma>=0.5 257 15 1 (0.05836576 0.94163424) *
```

```
4. printcp(stablo)

Classification tree:
rpart(formula = any ~ ., data = X, method = c("class"), parms = list(sp
lit = "information"),
    control = rpart.control(xval = 10))

Variables actually used in tree construction:
[1] dobEF   hxofcig hxofdm  hxofHT  mbp     mphr.b. pkhr    posSE   res
twma

Root node error: 90/558 = 0.16129

n= 558

        CP nsplit rel error xerror     xstd
1 0.019444      0   1.00000 1.0000 0.096535
2 0.014815      6   0.84444 1.2111 0.104058
3 0.010000     10   0.76667 1.2889 0.106508
```

```
5. plotcp(stablo,upper = c("splits"))
```



Slika 7 Odnos između parametra složenosti, greške grupiranja i broja dijeljenja stabla kao rezultat krosvalidacije

```
6. summary(stablo)
Call:
rpart(formula = any ~ ., data = X, method = c("class"), parms = list(sp
lit = "information"),
    control = rpart.control(xval = 10))
  n= 558

          CP nsplit rel error   xerror       xstd
1 0.01944444      0 1.0000000 1.000000 0.09653495
2 0.01481481      6 0.8444444 1.211111 0.10405832
3 0.01000000     10 0.7666667 1.288889 0.10650758
```

```
Variable importance
restwma   dobEF  baseEF   posSE mphr.b.    pkhr   maxhr  hxofMI      dp
     17      14      11       9       7       6       6       4       3
    bhr     mbp  hxofHT  gender  basedp hxofcig  hxofdm dpmaxdo  basebp
      3       3       3       3       2       2       2       2       2
    sbp
      1

Node number 1: 558 observations,    complexity param=0.01944444
  predicted class=1  expected loss=0.1612903  P(node) =1
    class counts:    90   468
   probabilities: 0.161 0.839
  left son=2 (301 obs) right son=3 (257 obs)
  Primary splits:
      restwma < 0.5     to the left,  improve=20.369730, (0 missing)
      posSE   < 0.5     to the left,  improve=19.916360, (0 missing)
      dobEF   < 52.5    to the left,  improve=18.891090, (0 missing)
      baseEF  < 53.5    to the left,  improve=13.267660, (0 missing)
      hxofMI  < 0.5     to the left,  improve= 8.016245, (0 missing)
  Surrogate splits:
      baseEF < 56.5    to the left,  agree=0.704, adj=0.358, (0 split)
      dobEF  < 69.5    to the left,  agree=0.703, adj=0.354, (0 split)
      posSE  < 0.5     to the left,  agree=0.654, adj=0.249, (0 split)
      hxofMI < 0.5     to the left,  agree=0.640, adj=0.218, (0 split)
      gender < 0.5     to the left,  agree=0.611, adj=0.156, (0 split)

Node number 2: 301 observations,    complexity param=0.01944444
  predicted class=1  expected loss=0.2491694  P(node) =0.5394265
    class counts:    75   226
   probabilities: 0.249 0.751
  left son=4 (76 obs) right son=5 (225 obs)
  Primary splits:
      dobEF  < 52.5    to the left,  improve=7.498062, (0 missing)
      posSE  < 0.5     to the left,  improve=5.770169, (0 missing)
      hxofHT < 0.5     to the left,  improve=4.594196, (0 missing)
      baseEF < 36.5    to the left,  improve=4.170991, (0 missing)
      bhr    < 89.5    to the left,  improve=3.983749, (0 missing)
  Surrogate splits:
      baseEF  < 41      to the left,  agree=0.904, adj=0.618, (0 split)
      maxhr   < 64.5    to the left,  agree=0.757, adj=0.039, (0 split)
      mphr.b. < 41.5    to the left,  agree=0.754, adj=0.026, (0 split)
```

```
      pkhr    < 64.5    to the left,  agree=0.751, adj=0.013, (0 split)
      dpmaxdo < 26268.5 to the right, agree=0.751, adj=0.013, (0 split)

Node number 3: 257 observations
  predicted class=1  expected loss=0.05836576  P(node) =0.4605735
    class counts:    15   242
   probabilities: 0.058 0.942

Node number 4: 76 observations,    complexity param=0.01944444
  predicted class=1  expected loss=0.4210526  P(node) =0.1362007
    class counts:    32    44
   probabilities: 0.421 0.579
  left son=8 (55 obs) right son=9 (21 obs)
  Primary splits:
      hxofHT < 0.5     to the left,  improve=3.388671, (0 missing)
      age    < 76.5    to the left,  improve=2.025083, (0 missing)
      bhr    < 90      to the left,  improve=2.013627, (0 missing)
      pkhr   < 121     to the left,  improve=1.894037, (0 missing)
      basedp < 14680   to the left,  improve=1.385555, (0 missing)
  Surrogate splits:
      basebp < 106.5   to the right, agree=0.75, adj=0.095, (0 split)
      basedp < 6354    to the right, agree=0.75, adj=0.095, (0 split)

Node number 5: 225 observations,    complexity param=0.01481481
  predicted class=1  expected loss=0.1911111  P(node) =0.4032258
    class counts:    43   182
   probabilities: 0.191 0.809
  left son=10 (89 obs) right son=11 (136 obs)
  Primary splits:
      posSE   < 0.5     to the left,  improve=5.882423, (0 missing)
      basedp  < 15132   to the right, improve=4.305910, (0 missing)
      dp      < 15706   to the left,  improve=3.395283, (0 missing)
      mphr.b. < 82.5    to the left,  improve=3.002242, (0 missing)
      posECG  < 0.5     to the left,  improve=2.500691, (0 missing)
  Surrogate splits:
      maxhr   < 144.5   to the right, agree=0.636, adj=0.079, (0 split)
      pkhr    < 148.5   to the right, agree=0.631, adj=0.067, (0 split)
      posECG  < 0.5     to the left,  agree=0.631, adj=0.067, (0 split)
      mphr.b. < 97.5    to the right, agree=0.622, adj=0.045, (0 split)
      baseEF  < 62.5    to the right, agree=0.622, adj=0.045, (0 split)

Node number 8: 55 observations,    complexity param=0.01944444
```

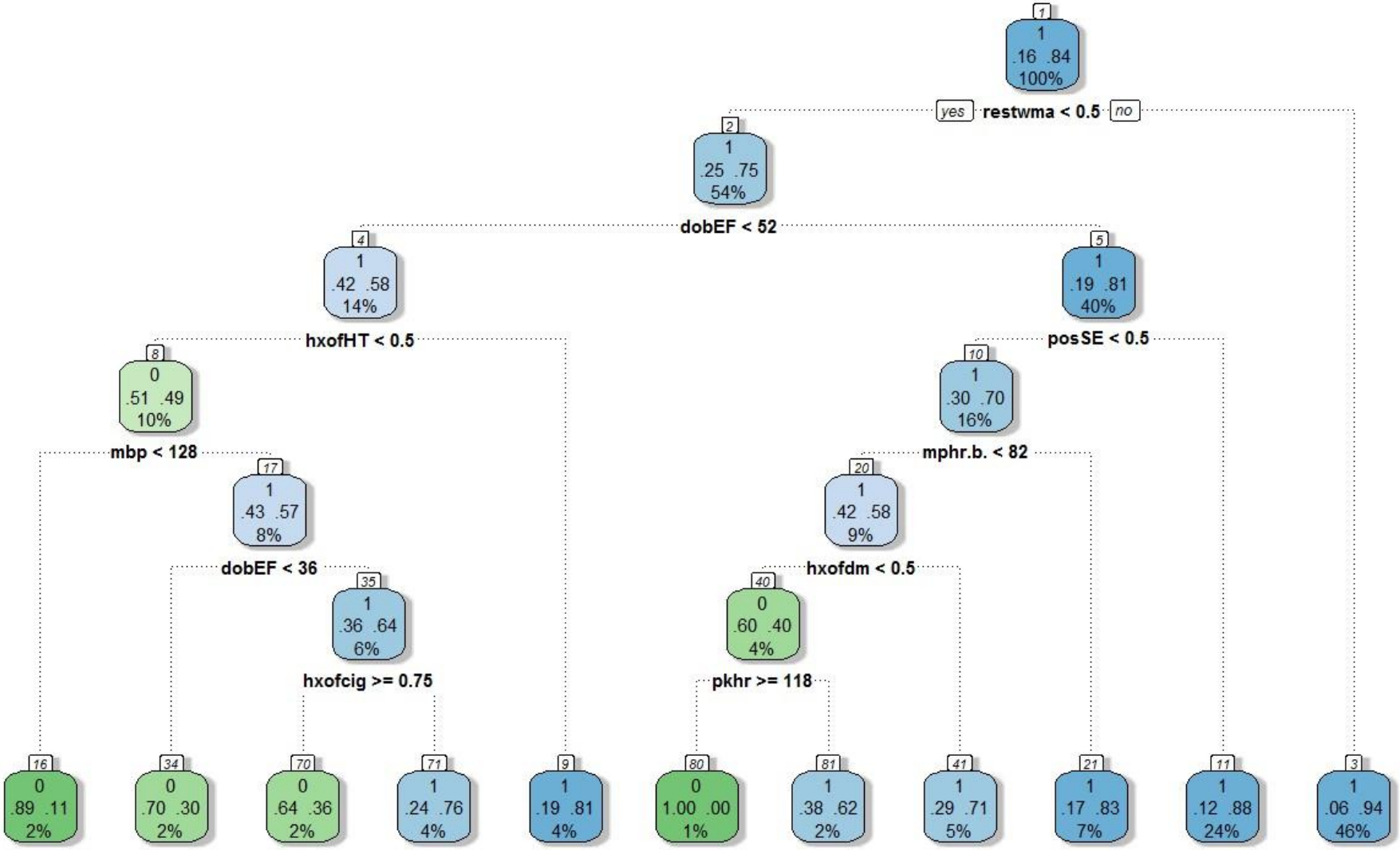
```
  predicted class=0  expected loss=0.4909091  P(node) =0.09856631
    class counts:    28    27
   probabilities: 0.509 0.491
  left son=16 (9 obs) right son=17 (46 obs)
  Primary splits:
      mbp     < 128.5   to the left,  improve=3.482166, (0 missing)
      dpmaxdo < 13085   to the left,  improve=2.788386, (0 missing)
      basebp  < 142     to the left,  improve=2.675554, (0 missing)
      dobEF   < 34      to the left,  improve=2.138985, (0 missing)
      age     < 76.5    to the left,  improve=1.919693, (0 missing)
  Surrogate splits:
      sbp     < 111     to the left,  agree=0.891, adj=0.333, (0 split)
      dp      < 11394.5 to the left,  agree=0.855, adj=0.111, (0 split)
      dpmaxdo < 12909.5 to the left,  agree=0.855, adj=0.111, (0 split)

Node number 9: 21 observations
  predicted class=1  expected loss=0.1904762  P(node) =0.03763441
    class counts:     4    17
   probabilities: 0.190 0.810

Node number 10: 89 observations,    complexity param=0.01481481
  predicted class=1  expected loss=0.3033708  P(node) =0.1594982
    class counts:    27    62
   probabilities: 0.303 0.697
  left son=20 (48 obs) right son=21 (41 obs)
  Primary splits:
      mphr.b. < 82.5    to the left,  improve=3.278607, (0 missing)
      hxofdm  < 0.5     to the left,  improve=3.190602, (0 missing)
      dp      < 15760   to the left,  improve=2.778147, (0 missing)
      basedp  < 13899   to the right, improve=2.743452, (0 missing)
      maxhr   < 98.5    to the left,  improve=1.844902, (0 missing)
  Surrogate splits:
      maxhr   < 121.5   to the left,  agree=0.899, adj=0.780, (0 split)
      pkhr    < 119.5   to the left,  agree=0.876, adj=0.732, (0 split)
      dpmaxdo < 18632   to the left,  agree=0.742, adj=0.439, (0 split)
      dp      < 16395   to the left,  agree=0.730, adj=0.415, (0 split)
      bhr     < 76.5    to the left,  agree=0.629, adj=0.195, (0 split)

Node number 11: 136 observations
  predicted class=1  expected loss=0.1176471  P(node) =0.2437276
    class counts:    16   120
   probabilities: 0.118 0.882
```

```
Node number 16: 9 observations
  predicted class=0  expected loss=0.1111111  P(node) =0.01612903
    class counts:     8     1
   probabilities: 0.889 0.111

Node number 17: 46 observations,    complexity param=0.01944444
  predicted class=1  expected loss=0.4347826  P(node) =0.08243728
    class counts:    20    26
   probabilities: 0.435 0.565
  left son=34 (10 obs) right son=35 (36 obs)
  Primary splits:
      dobEF   < 35.5    to the left,  improve=1.837733, (0 missing)
      bhr     < 90      to the left,  improve=1.535026, (0 missing)
      dpmaxdo < 18829   to the right, improve=1.374506, (0 missing)
      sbp     < 165     to the right, improve=1.293556, (0 missing)
      hxofcig < 0.75    to the right, improve=1.235064, (0 missing)
  Surrogate splits:
      basebp  < 116     to the left,  agree=0.891, adj=0.5, (0 split)
      baseEF  < 27.5    to the left,  agree=0.848, adj=0.3, (0 split)
      dp      < 12400   to the left,  agree=0.804, adj=0.1, (0 split)
      mphr.b. < 49.5    to the left,  agree=0.804, adj=0.1, (0 split)
      dpmaxdo < 12245   to the left,  agree=0.804, adj=0.1, (0 split)

Node number 20: 48 observations,    complexity param=0.01481481
  predicted class=1  expected loss=0.4166667  P(node) =0.08602151
    class counts:    20    28
   probabilities: 0.417 0.583
  left son=40 (20 obs) right son=41 (28 obs)
  Primary splits:
      hxofdm < 0.5     to the left,  improve=2.389495, (0 missing)
      maxhr  < 122     to the right, improve=2.204730, (0 missing)
      basebp < 120.5   to the left,  improve=2.080597, (0 missing)
      pkhr   < 122     to the right, improve=2.080597, (0 missing)
      dose   < 37.5    to the left,  improve=1.765492, (0 missing)
  Surrogate splits:
      baseEF < 54      to the left,  agree=0.708, adj=0.30, (0 split)
      dobEF  < 62.5    to the left,  agree=0.708, adj=0.30, (0 split)
      basebp < 137.5   to the right, agree=0.688, adj=0.25, (0 split)
      basedp < 13899   to the right, agree=0.667, adj=0.20, (0 split)
      mbp    < 159.5   to the right, agree=0.667, adj=0.20, (0 split)
```

```
Node number 21: 41 observations
  predicted class=1  expected loss=0.1707317  P(node) =0.0734767
    class counts:     7    34
   probabilities: 0.171 0.829

Node number 34: 10 observations
  predicted class=0  expected loss=0.3  P(node) =0.01792115
    class counts:     7     3
   probabilities: 0.700 0.300

Node number 35: 36 observations,    complexity param=0.01944444
  predicted class=1  expected loss=0.3611111  P(node) =0.06451613
    class counts:    13    23
   probabilities: 0.361 0.639
  left son=70 (11 obs) right son=71 (25 obs)
  Primary splits:
      hxofcig < 0.75    to the right, improve=2.558675, (0 missing)
      basedp  < 10968   to the right, improve=2.555107, (0 missing)
      mbp     < 149.5   to the right, improve=1.858522, (0 missing)
      sbp     < 168     to the right, improve=1.372376, (0 missing)
      dpmaxdo < 18933.5 to the right, improve=1.208915, (0 missing)
  Surrogate splits:
      mphr.b. < 66      to the left,  agree=0.778, adj=0.273, (0 split)
      basebp  < 169.5   to the right, agree=0.722, adj=0.091, (0 split)
      dp      < 15194   to the left,  agree=0.722, adj=0.091, (0 split)
      dpmaxdo < 15450   to the left,  agree=0.722, adj=0.091, (0 split)
      baseEF  < 46.5    to the right, agree=0.722, adj=0.091, (0 split)

Node number 40: 20 observations,    complexity param=0.01481481
  predicted class=0  expected loss=0.4  P(node) =0.03584229
    class counts:    12     8
   probabilities: 0.600 0.400
  left son=80 (7 obs) right son=81 (13 obs)
  Primary splits:
      pkhr    < 118.5   to the right, improve=4.798614, (0 missing)
      mphr.b. < 73      to the right, improve=3.110445, (0 missing)
      maxhr   < 112.5   to the right, improve=2.295753, (0 missing)
      dose    < 37.5    to the left,  improve=2.249692, (0 missing)
      bhr     < 69.5    to the right, improve=1.726092, (0 missing)
  Surrogate splits:
      maxhr   < 112.5   to the right, agree=0.95, adj=0.857, (0 split)
      bhr     < 70.5    to the right, agree=0.90, adj=0.714, (0 split)
```

```
      mphr.b. < 73      to the right, agree=0.90, adj=0.714, (0 split)
      basedp  < 11017   to the right, agree=0.80, adj=0.429, (0 split)
      dp      < 19324   to the right, agree=0.80, adj=0.429, (0 split)

Node number 41: 28 observations
  predicted class=1  expected loss=0.2857143  P(node) =0.05017921
    class counts:     8    20
   probabilities: 0.286 0.714

Node number 70: 11 observations
  predicted class=0  expected loss=0.3636364  P(node) =0.01971326
    class counts:     7     4
   probabilities: 0.636 0.364

Node number 71: 25 observations
  predicted class=1  expected loss=0.24  P(node) =0.04480287
    class counts:     6    19
   probabilities: 0.240 0.760

Node number 80: 7 observations
  predicted class=0  expected loss=0  P(node) =0.0125448
    class counts:     7     0
   probabilities: 1.000 0.000

Node number 81: 13 observations
  predicted class=1  expected loss=0.3846154  P(node) =0.02329749
    class counts:     5     8
   probabilities: 0.385 0.615
```

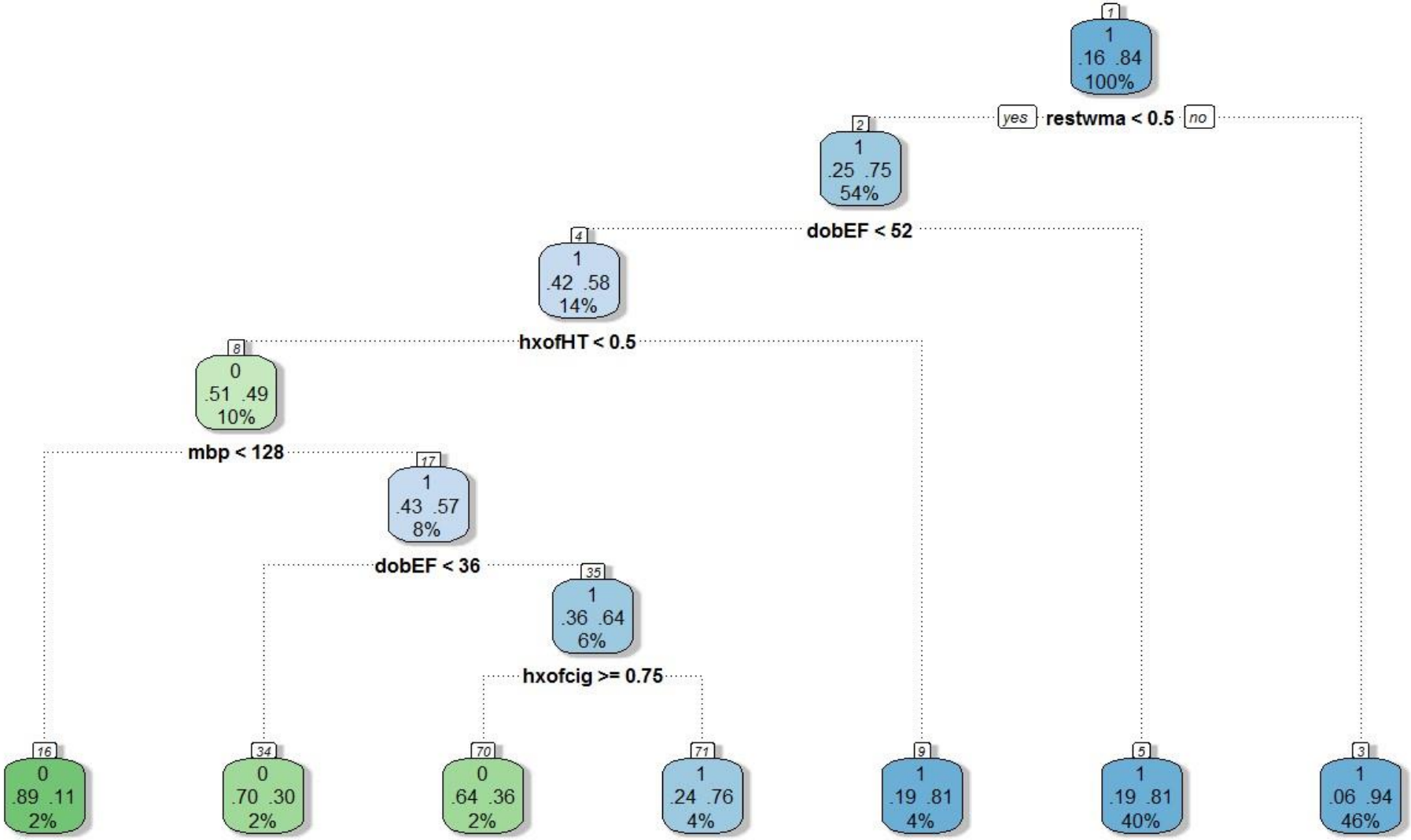
6. fancyRpartPlot(stablo)



Slika 8 Klasifikacijsko stablo za predviđanje srčanog događa dobiveno korištenjem funkcije „fancyRpartPlot“

```
7. cpp<-stablo$cptable[which.min((stablo$cptable[,"xerror"]-stablo$cptable[
,"rel error"]-stablo$cptable[,"xstd"])<0 ),"CP"]

 [1] 0.01481481

8. pruned<-prune(stablo,cp=cpp)
n= 558

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

 1) root 558 90 1 (0.16129032 0.83870968)
   2) restwma< 0.5 301 75 1 (0.24916944 0.75083056)
     4) dobEF< 52.5 76 32 1 (0.42105263 0.57894737)
       8) hxofHT< 0.5 55 27 0 (0.50909091 0.49090909)
        16) mbp< 128.5 9  1 0 (0.88888889 0.11111111) *
        17) mbp>=128.5 46 20 1 (0.43478261 0.56521739)
          34) dobEF< 35.5 10  3 0 (0.70000000 0.30000000) *
          35) dobEF>=35.5 36 13 1 (0.36111111 0.63888889)
            70) hxofcig>=0.75 11  4 0 (0.63636364 0.36363636) *
            71) hxofcig< 0.75 25  6 1 (0.24000000 0.76000000) *
       9) hxofHT>=0.5 21  4 1 (0.19047619 0.80952381) *
     5) dobEF>=52.5 225 43 1 (0.19111111 0.80888889) *
   3) restwma>=0.5 257 15 1 (0.05836576 0.94163424) *
```

9. fancyRpartPlot(pruned)



Slika 9 Skraćeno klasifikacijsko stablo sa parametrom složenost cp=0.014815 dobiveno korištenjem funkcije „fancyRpartPlot“

U prva dva reda koda prikazano je učitavanje i prikaz nekoliko od 558 podataka koje promatramo. Nadalje u idućem koraku koristimo rpart funkciju na učitanim podacima u kojoj za podjelu čvorova koristimo funkciju entropije kao funkciju nečistoće te metodu klasifikacijskog stabla jer je varijabla izlaza kategorijska. Kao rezultat na poziv funkcije dobijemo listu čvorova sa njihovim sadržajem i rezultat krosvalidacije korištenjem 10-strukih koraka. Svaki čvor sadrži informaciju o rednom broju čvora, varijablu kojom se taj čvor dijeli, količinu podataka koji se nalaze u tom čvoru, klasu koju poprima taj čvor – 0 ili 1 ovisno o dominaciji broja pacijenata. Zatim vjerojatnost događaja izlazne varijable gdje se prvi broj odnosi na vjerojatnost da će pacijent kao rezultat imati nulu, odnosno s kolikom će vjerojatnosti imati neki srčani događaj od četiri koja promatramo. Drugi broj se odnosi na vjerojatnost događaja jedinice, tj. postotak da pacijent neće imati nijedan od 4 moguća događaja. Čvor označen zvjezdicom (*) određuje konačni čvor, tj. list.

Rpart svojim pozivom napravi prošireno stablo tako da svaki prolazni čvor sadrži minimalno deset podataka te ujedno računa parametar složenosti na način da napravi krosvalidaciju s zadanim korakom. Nadalje, funkcija „printcp“ vraća razine stabla s vrijednostima koje vidimo u tablici 5: parametar složenosti, broj dijeljenja stabla, „rel error“ - očekivana suma kvadrata koja je jednaka greški grupiranih podataka korištenih za izradu stabla, tj. $1 - R^2$, „xerror“ – greška grupiranja dobivena iz krosvalidacije i standardnu pogrešku, tj. „xstd“.

	CP	nsplit	rel error	xerror	xstd
1	0.019444	0	1.00000	1.0000	0.096535
2	0.014815	6	0.84444	1.2111	0.104058
3	0.010000	10	0.76667	1.2889	0.106508

Tablica 1 Tablica krosvalidacije za izbor parametra složenosti

Iz Tablice 5 uočimo da greška grupiranja kod krosvalidacije ima vrijednost 1.2111, jer je to najmanja vrijednost za koju vrijedi uvjet da je $rel\,error + xstd < xerror$, stoga parametar složenosti ima vrijednost 0.014815 koju dalje koristimo za skraćivanje stabla.

Nadalje, peta linija daje vizualni prikaz odnosa između parametra složenosti, greške grupiranja te broja podjela stabla iz koje možemo potvrditi odluku iz prethodne tablice.

„Summary“ funkcija, tj. šesta linija koda detaljnije opisuje rezultate „rpart“ funkcije gdje prikazuje preciznu podjelu svakog čvora prikazanog u vizualizaciji stabla. Uostalom pokazuje vrijednost parametra složenosti za broj podjela čvorova, te vrijednosti relativnih grešaka.

„FancyRpartPlot“, vizualna funkcija, pruža prikaz stabla sa čvorovima koji ga dijele iz kojeg uočavamo mogući najbolji put za procjenu događaja srčanog događaja. Kao što vidimo iz slike 8 put se sastoji od idućih varijabli i uvjeta redom: $\{restwma < 0.5, dobEF < 52, hxofHT < 0.5, mbp < 128\}$ gdje je procjenjena vjerojatnost srčanog događaja jednaka 89%.

Za kraj, skratimo stablo koristeći funkciju „prune“ parametrom dobivenog stabla i novo dobivenim parametrom složenosti. Iz slike 9 uočimo da se grane stabla skraćuju sa povećanjem parametra složenosti, tj. smanjenjem greške grupiranja stabla iz krosvalidacije.

Kao rezultat CART analize uočavamo da varijabla *restwma* koja najbolje predviđa srčani događaj prva prepreka u stablu za dijeljenje prisutnosti i nedostatku te varijable. Za 301 osobu koje imaju negativnu vrijednost za *restwma*, iduća značajna varijabla za predviđanje događaja je *dobEF* sa vrijednosti manjom od 52. Sljedeće varijable dijeljenje su redom: *hxofHT* s negativnom vrijednosti i *mbp* vrijednosti manjom od 128, što nam daje ukupno 9 pacijenata od cijele populacije koju promatramo. Od njih devetero, 89% pacijenata je imalo srčani događaj nasuprot vjerojatnosti od 16% iz cijele populacije.

8. Literatura

[1] Alan Julian Izenman, *Modern Multivariate Statistical Techniques: Regression, Classification, and Monifold Learning*, Springer Science + Business Media, LLC, 223 Spring Street, New York, NY 10013, USA, 2008.

[2] Alan Garfinkel, *Prognostical Value of Dobutamine Stress Echocardiography in Predicting Cardiac Events in Patients With Known or Suspected Coronary Artery Disease*, Journal of the American College of Cardiology, Elsevier Science Inc, 1999

[3] Jake Morgan, *Classification and Regression Tree Analysis*, School of Public Health, Techical Reports Available: Statistical Methods for Health Policy and Management Research, 2013

[4] Leonard Gordon, *Using classification and Regression Trees (CART) in SAS® Enterprise Miner™ for Applications in Public Health,* SAS Institute Inc.,2013

[5] R Studio, https://www.rstudio.com/

[6] Roger J. Lewis, *An Introduction to Classification and Regression Tree (CART) Analysis,* Presented at the 2000 Annual Meeting of the Society for Academic Emergencyc Medicine in San Francisco, California

[7] Izvor podataka, www.stat.ucla.edu/projects/datasets/cardiac-explanation.html, UCLA Department of Statistics (2012)

9. Sažetak

Klasifikacijsko i regresijsko stablo odlučivanja, kao savršen statistički alat, koristi se najčešće u istraživačkim medicinskim problemima čiji je cilj predvidjeti vjerojatnost varijable izlaza, bila ona kategorijska ili kontinuirana. Premda nam CART analiza vraća rezultate u obliku vjerojatnosti u više konačnih čvorova, na samom korisniku je zadaća da procijeni koliko su ti rezultati značajni za daljnju uporabu.

U ovom radu je okvirno opisan algoritam za CART analizu kojim se definira, proširuje i skraćuje stablo koristeći neke od metoda kao što su nezavisnost podataka, krosvalidacija i funkcija nečistoće. Ujedno sadrži i definirane dodatne pristupe stablima koji služe za bolji rad sa podacima i samim stablom.

Na kraju samog rada se nalazi primjena CART analize na medicinskim podacima u R softveru, gdje se koristi klasifikacijsko stablo kao alat za izradu modela za predviđanje srčanog događaja na pacijentima sa mogućom koronarnom bolesti.

Prednosti korištenja CART analize tiču se njene ne-parametarske i ne-linearne prirode. Ne zahtijeva pretpostavku distribucije podataka te funkcionalnu formu za prediktore što im omogućava da identificiraju složene interakcije među podacima. To je ujedno i najlakše korištena statistička tehnika, jer ne zahtijeva puno znanja iz statistike i prati proces odlučivanja koji ljudi koriste za donošenje odluka.

10. Summary

Classification ad regression decision trees, as perfect statistical tool, is often used in medical research problems that aims to predict value of target variable, whether it be categorical or continuous. Although, CART returns probabilistic analysis results in more terminal nodes, but it's on the user to evaluate how are these results important for further use.

This work describes a framework for CART analysis algorithm, which defines, extends and shorts trees, using some of the methods such as independence set of data, cross validation and impurity function. It also contains additional approaches to trees that serve for better operations on data.

At the end of this work reader can find application of CART analysis on the medical data in R software, where is used classification tree tool to develop a model. We use that model for predicting cardiac events in patients with known or suspected coronary artery disease.

Advantages of using CART are related to its non-parametric and non-linear nature. It does not require the assumption of distribution of data and functional form for predictors which allows them to identify the complex interactions among data. It's also the most easily used statistical technique because it does not require much knowledge of statistics and monitor deciding process which people use to make final decisions.

11. Životopis

Rođen sam 24. travnja 1992. godine u Imotskom. Nakon osnovne škole (OŠ Zmijavci) upisujem Prirodoslovno – Matematičku Gimnaziju u Imotskom. Maturirao sam 2011. godine i iste godine upisao Prirodoslovno – Matematički fakultet, Matematički odsjek u Zagrebu. Nakon završenog Preddiplomskog sveučilišnog studija Matematike; smjer: nastavnički, 2014. godine upisujem Diplomski sveučilišni studij Matematičke statistike na istom odsjeku.