

Vaje 3:

1. naloga:

Klasifikator je na 4-razrednem problemu klasificiral pet testnih primerov. V spodnji tabeli je podana napovedana verjetnostna porazdelitev po starih razredih za vsakega od petih testnih primerov:

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

Izračunaj:

- a) klasifikacijsko točnost klasifikatorja,
- b) povprečno Brierjevo mero
- c) povprečno informacijsko vsebino odgovora, če je apriorna porazdelitev po razredih:
 $P(C_1) = 0.1, P(C_2) = 0.5, P(C_3) = 0.2$ in $P(C_4) = 0.2$.

Rešitev: (napišeš za vsakega če je pravilno klasificiran ali ne) + povprečje

a)

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.10	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

Prvi primer je nepravilno klasificiran, saj največja napoved ni ista kot razred (C₄ najbolj ustreza razredu C₁, namesto razredu C₄ kot bi moral).

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

Drugi primer je pravilno klasificiran, saj je napoved ista kot razred (razred testnega primera C₂ ustreza napovedi razreda C₂, saj ima tu najvišjo napoved (0.55)).

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.15	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

Tretji primer je pravilno klasificiran.

Četrти primer je pravilno klasificiran.

Peti primer je pravilno klasificiran.

$$\text{klasifikacijska točnost} = \frac{4}{5} \frac{(\text{št. pravilnih napovedi po razredih})}{(\text{vsi razredi})} = 0.8$$

b)

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

Prvi primer (razred testnega primera):

Ciljna distribucija:	C ₁	C ₂	C ₃	C ₄
	0	0	0	1

Napovedana distribucija:	C ₁	C ₂	C ₃	C ₄
	0.65	0.25	0.00	0.10

$$\text{Kvadratna napaka: } (0 - 0.65)^2 + (0 - 0.25)^2 + (0 - 0)^2 + (1 - 0.1)^2 = 1.295$$

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

Drugi primer:

Ciljna distribucija:	C ₁	C ₂	C ₃	C ₄
	0	1	0	0

Napovedana distribucija:	C ₁	C ₂	C ₃	C ₄
	0.20	0.55	0.25	0.00

$$\text{Kvadratna napaka: } (0 - 0.2)^2 + (1 - 0.55)^2 + (0 - 0.25)^2 + (0 - 0)^2 = 0.305$$

Tretji primer:

	C ₁	C ₂	C ₃	C ₄
Ciljna distribucija:	1	0	0	0
Napovedana distribucija:	0.75	0.00	0.25	0.00

$$\text{Kvadratna napaka: } (1 - 0.75)^2 + (0 - 0)^2 + (0 - 0.25)^2 + (0 - 0)^2 = 0.125$$

Četrti primer:

	C ₁	C ₂	C ₃	C ₄
Ciljna distribucija:	0	1	0	0
Napovedana distribucija:	0.25	0.50	0.00	0.25

$$\text{Kvadratna napaka: } (0 - 0.25)^2 + (1 - 0.5)^2 + (0 - 0)^2 + (0 - 0.25)^2 = 0.375$$

Peti primer:

	C ₁	C ₂	C ₃	C ₄
Ciljna distribucija:	0	0	1	0
Napovedana distribucija:	0.10	0.10	0.60	0.20

$$\text{Kvadratna napaka: } (0 - 0.1)^2 + (0 - 0.1)^2 + (1 - 0.6)^2 + (0 - 0.2)^2 = 0.22$$

$$Brier = \frac{1.295 + 0.305 + 0.125 + 0.375 + 0.22}{5} = 0.464$$

c)

$r^{(j)}$ pravilen razred j-tega testnega primera

$P(r^{(j)})$ apriorna verjetnost pravilnega razreda j-tega testnega primera

$P'(r^{(j)})$ aposteriorna verjetnost pravilnega razreda j-tega testnega primera

$$Inf_j = \begin{cases} -\log_2 P(r^{(j)}) + \log_2 P'(r^{(j)}), & P'(r^{(j)}) \geq P(r^{(j)}) \\ -\left(-\log_2(1 - P(r^{(j)})) + \log_2(1 - P'(r^{(j)}))\right), & P'(r^{(j)}) < P(r^{(j)}) \end{cases}$$

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

apriorna porazdelitev po razredih:

$P(C_1) = 0.1$
 $P(C_2) = 0.5$
 $P(C_3) = 0.2$
 $P(C_4) = 0.2$

$$Inf_j = \begin{cases} -\log_2 P(r^{(j)}) + \log_2 P'(r^{(j)}), & P'(r^{(j)}) \geq P(r^{(j)}) \\ -\left(-\log_2(1 - P(r^{(j)})) + \log_2(1 - P'(r^{(j)}))\right), & P'(r^{(j)}) < P(r^{(j)}) \end{cases}$$

Prvi primer:

aposteriorna ver. pravega razreda: $P'(C_4) = 0.1$

apriorna ver. pravega razreda: $P(C_4) = 0.2$

$$Inf_j: -(-\log_2(1-0.2)+\log_2(1-0.1))=-0.1699$$

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₃	0.20	0.55	0.25	0.00
C ₁	0.75	0.10	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

apriorna porazdelitev
po razredih:

P(C₁) = 0.1
P(C₂) = 0.5
P(C₃) = 0.2
P(C₄) = 0.2

$$Inf_j = \begin{cases} -\log_2 P(r^{(j)}) + \log_2 P'(r^{(j)}), & P'(r^{(j)}) \geq P(r^{(j)}) \\ -(-\log_2(1 - P(r^{(j)})) + \log_2(1 - P'(r^{(j)}))), & P'(r^{(j)}) < P(r^{(j)}) \end{cases}$$

Drugi primer:

aposteriori ver. pravega razreda: P'(C₂) = 0.55

apriorna ver. pravega razreda: P(C₂) = 0.5

$$Inf_2: -\log_2(0.5) + \log_2(0.55) = 0.1375$$

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.10	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

apriorna porazdelitev
po razredih:

P(C₁) = 0.1
P(C₂) = 0.5
P(C₃) = 0.2
P(C₄) = 0.2

$$Inf_j = \begin{cases} -\log_2 P(r^{(j)}) + \log_2 P'(r^{(j)}), & P'(r^{(j)}) \geq P(r^{(j)}) \\ -(-\log_2(1 - P(r^{(j)})) + \log_2(1 - P'(r^{(j)}))), & P'(r^{(j)}) < P(r^{(j)}) \end{cases}$$

Tretji primer:

aposteriori ver. pravega razreda: P'(C₁) = 0.75

apriorna ver. pravega razreda: P(C₁) = 0.1

$$Inf_1: -\log_2(0.1) + \log_2(0.75) = 2.9069$$

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

apriorna porazdelitev
po razredih:

P(C₁) = 0.1
P(C₂) = 0.5
P(C₃) = 0.2
P(C₄) = 0.2

$$Inf_j = \begin{cases} -\log_2 P(r^{(j)}) + \log_2 P'(r^{(j)}), & P'(r^{(j)}) \geq P(r^{(j)}) \\ -(-\log_2(1 - P(r^{(j)})) + \log_2(1 - P'(r^{(j)}))), & P'(r^{(j)}) < P(r^{(j)}) \end{cases}$$

Četrти primer:

aposteriori ver. pravega razreda: P'(C₂) = 0.5

apriorna ver. pravega razreda: P(C₂) = 0.5

$$Inf_4: -\log_2(0.5) + \log_2(0.5) = 0$$

Pravi razred testnega primera	Napoved po razredih			
	C ₁	C ₂	C ₃	C ₄
C ₄	0.65	0.25	0.00	0.10
C ₂	0.20	0.55	0.25	0.00
C ₁	0.75	0.00	0.25	0.00
C ₂	0.25	0.50	0.00	0.25
C ₃	0.10	0.10	0.60	0.20

apriorna porazdelitev
po razredih:

P(C₁) = 0.1
P(C₂) = 0.5
P(C₃) = 0.2

P(C₄) = 0.2

$$\ln f_j = \begin{cases} -\log_2 P(r^{(j)}) + \log_2 P'(r^{(j)}), & P'(r^{(j)}) \geq P(r^{(j)}) \\ -(-\log_2(1 - P(r^{(j)})) + \log_2(1 - P'(r^{(j)}))), & P'(r^{(j)}) < P(r^{(j)}) \end{cases}$$

Peti primer:

aposteriorna ver. pravega razreda: P'(C₂) = 0.6

apriorna ver. pravega razreda: P(C₂) = 0.2

$$\text{Inf}_5: -\log_2(0.2) + \log_2(0.6) = 1.585$$

$$Inf = \frac{-0.1699 + 0.1375 + 2.9069 + 0 + 1.585}{5} = 0.8919$$

Vaje 5:

2. naloga:

Naj bo A binarni atribut z vrednostima 0 in 1.

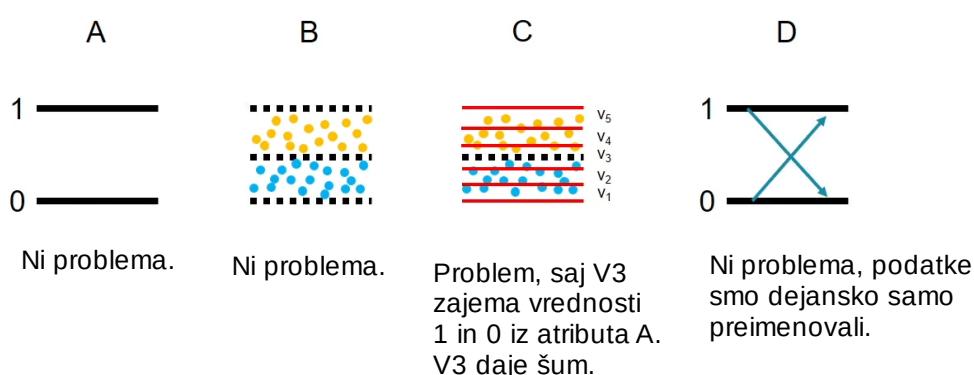
Naj bo B zvezni atribut, dobljen iz atributa A, tako da se vrednost 0 spremeni v naključno vrednost na intervalu [0,0.49] in vrednost 1 spremeni v naključno vrednost na intervalu [0.5,1].

Naj bo C diskretni atribut, dobljen iz atributa B z ekvidistančno diskretizacijo na 5 intervalov.

Naj bo D binarni atribut, dobljen iz atributa A tako da sta vrednosti 0 in 1 zamenjani.

Razvrsti vse stiri atribute po oceni kvalitete z gini indeksom. Razvrstitev argumentiraj!

Ugotavljanje z vizualizacijo:



Gini index precenjuje večvrednostne atribute, kar pomeni, da bosta atributa A in D ocenjena enako, najboljša, slabša (precenjena) pa C (ker je večvrednostni atribut in šuma v V3) in B (večvrednostni atribut). Katero mesto si delita je odvisno kaj gledamo da ima večji vpliv (oz. komu damo prednost), šum v atributu ali večvrednostni atribut. To je odvisno od konkretnje porazdelitve.

Rešitev:

Če gledamo torej po večvrednostnem atributu, bo najbolje ocenjen B (ogromno možnosti, precenjen), sledi C (5 možnosti, precenjen), sledita A in D (ocenjena enako).

Če gledamo pa po šumu, bo najbolje ocenjen C (šum v V3, precenjen), sledi B (večvrednostni atribut, precenjen), sledita A in D (ocenjena enako).

3. naloga:

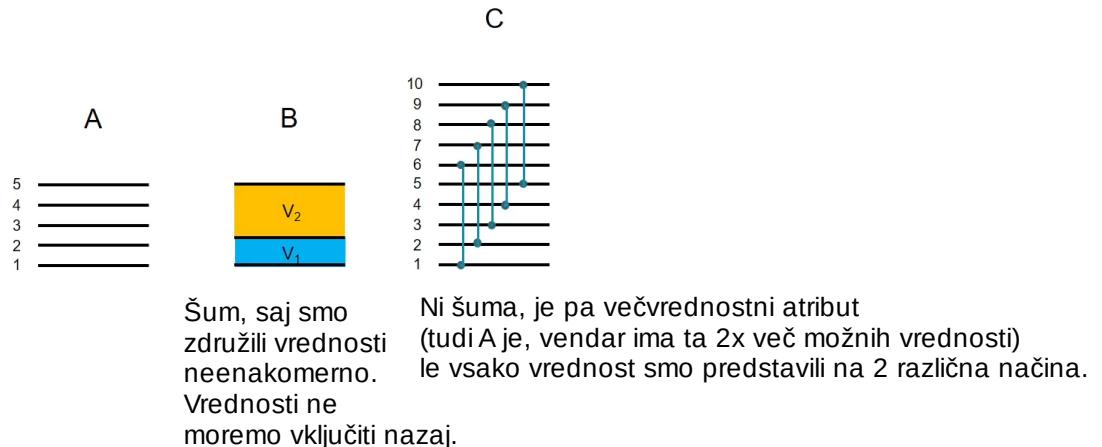
Naj bo A diskretni atribut s 5 vrednostmi (1,2,3,4,5).

Naj bo B binarni atribut, dobljen iz atributa A, tako da se zdruzijo prvi dve vrednosti v eno in preostale tri v drugo vrednost.

Naj bo C diskretni atribut, dobljen iz A tako, da se vsaka vrednost V naključno spremeni bodisi v isto vrednost V ali v vrednost V+5.

Razvrsti vse atribute po oceni kvalitete z informacijskim prispevkom. Razvrstitev argumentiraj!

Ugotavljanje z vizualizacijo:



Informacijski prispevek deluje le na diskretnih atributih. Na srečo so tukaj vsi diskretni.

Rešitev:

Najbolje bo ocenjen C (ker je večvrednostni atribut), slabše od njega bo ocenjen A, sledi B (zaradi šuma).

4. naloga:

Podani so naslednji klasifikacijski problemi na binarnih atributih A1 do A10.

- a) $C = |(A1 + A5 + A4) \text{ mod } 2|$
- b) $C = P(A1 > 0.5) * P(A2 > 0.1) > 0.3$
- c) $C = (A1 > 0.3 \text{ and } A2 > 0.1 \text{ or } A2 > 0.2 \text{ and } A5 = 0)$

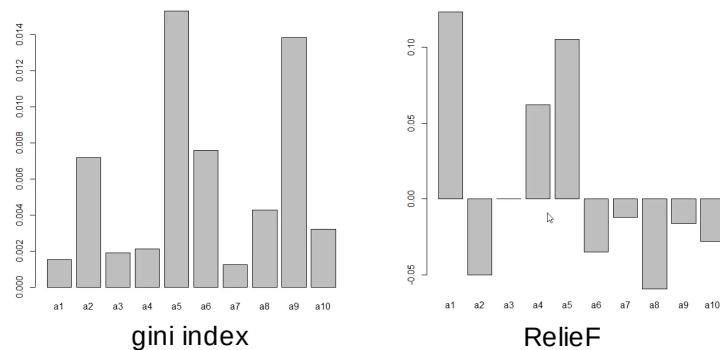
Pri tem $P(X)$ predstavlja verjetnost, da je X resnicen

Delno uredi attribute po pomembnosti, ce attribute ocenjujes z:

- oceno gini indeks
- oceno ReliefF

Ugotavljanje in rešitve:

- a) gini indeks, ker je kratkovidna ocena, bo ocenil attribute bolj kot ne naključno, kot nepomembne
 ReliefF, ker razzna attribute v relaciji, bo vedel da so ti attribute pomembni (A1, A5, A4), v primerjavi z A2, A3 itd.



A1	A4	A5	C
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

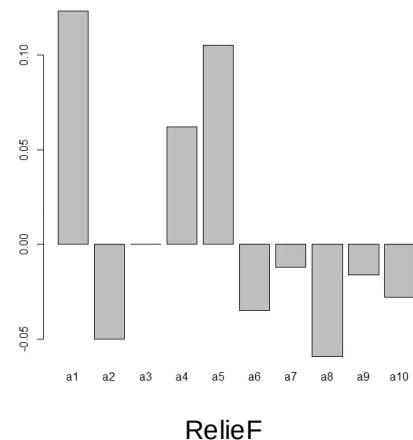
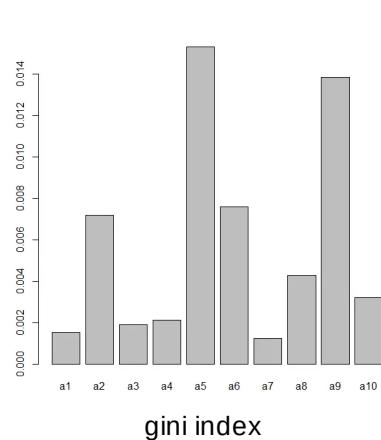
C pri A5 = 50% 0, 50% 1
 C pri A4 = 50% 0, 50% 1
 C pri A1 = 50% 0, 50% 1

b) $C = P(A1 > 0.5) * P(A2 > 0.1) > 0.3$

Kako beremo:

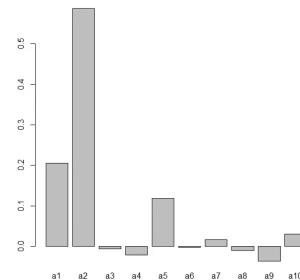
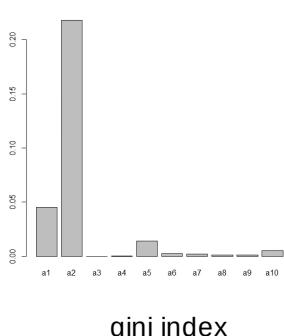
razred = verjetnost da je($A1 > 0.5$) * verjetnost($A2 > 0.1$) > 0.3
je enak (obe verjetnosti) večje kot konstanta 0.3

Razred je konstanten, atributi ne vplivajo na razred. Vsi atributi so nepomembni.



c)

Povezava ni tako izrazita kot pri a).



A1	A2	A5	C
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

Vaje 7:

1. naloga:

Zaporniki se na otoku lahko prosto gibljejo. Do kopnega je 8 km in pogosto lahko vidimo v morju morske pse. Nekateri zaporniki so poskušali pobegniti s plavanjem ponoči, a samo nekaterim uspe doseči kopno. V tabeli so opisi nekaterih poskusov pobega.

Dober plavalec	Dežuje	Polna luna	Podnevi videni morski psi	So ga opazili pazniki	Je dosegel kopno
DA	DA	DA	DA	DA	DA
NE	NE	DA	NE	NE	DA
DA	NE	NE	NE	DA	NE
NE	NE	DA	NE	DA	NE
DA	NE	NE	DA	NE	NE
DA	DA	DA	DA	NE	NE

Uporabi naivni Bayes z m-oceno ($m=5$) in oceni verjetnost, da je bil zapornik dober plavalec, če ni dosegel kopna med polno luno ter so ga med pobegom opazili pazniki.

Rešitev:

Pri naivnem Bayes-u atribute, ki niso omenjeni v oceni verjetnosti ignoriramo.

$$P(\text{plavalec} = \text{"DA"} | \text{kopno} = \text{"NE"}, \text{luna} = \text{"DA"}, \text{pazniki} = \text{"DA"}) = ?$$

pogojna verjetnost

$$P(\text{plavalec} = \text{"DA"}) * \frac{P(\text{plavalec} = \text{"DA"} | \text{kopno} = \text{"NE"})}{P(\text{plavalec} = \text{"DA"})} * \frac{P(\text{plavalec} = \text{"DA"} | \text{luna} = \text{"DA"})}{P(\text{plavalec} = \text{"DA"})} * \frac{P(\text{plavalec} = \text{"DA"} | \text{pazniki} = \text{"DA"})}{P(\text{plavalec} = \text{"DA"})}$$

apriorna verjetnost

št. primerov iz razreda "DA"
(iz iskanega razreda)

$$P(\text{plavalec} = \text{"DA"}) = \frac{N_k + 1}{N + n_0} = \frac{4 + 1}{6 + 2} = 0.625$$

apriorna verjetnost

št. učnih primerov št. razredov

št. primerov iz razreda "DA" z vrednostjo atributa kopno = "NE"

apriorna verjetnost razreda "DA"

$$P(\text{plavalec} = \text{"DA"} | \text{kopno} = \text{"NE"}) = \frac{N_{k,i} + m * P_k}{N_i + m} = \frac{3 + 5 * 0.625}{4 + 5} = 0.6805$$

pogojna verjetnost razreda "DA" pri vrednosti atributa kopno = "NE"

št. primerov z vrednostjo atributa kopno = "NE"

parameter m-ocene
(podan v besedilu naloge)

$$P(\text{plavalec} = \text{"DA"} | \text{luna} = \text{"DA"}) = \frac{N_{k,i} + m * P_k}{N_i + m} = \frac{2 + 5 * 0.625}{4 + 5} = 0.5694$$

$$P(\text{plavalec} = \text{"DA"} | \text{pazniki} = \text{"DA"}) = \frac{N_{k,i} + m * P_k}{N_i + m} = \frac{2 + 5 * 0.625}{3 + 5} = 0.6406$$

$$\begin{aligned} P(\text{plavalec} = \text{"DA"}) * \frac{P(\text{plavalec} = \text{"DA"} | \text{kopno} = \text{"NE"})}{P(\text{plavalec} = \text{"DA"})} * \frac{P(\text{plavalec} = \text{"DA"} | \text{luna} = \text{"DA"})}{P(\text{plavalec} = \text{"DA"})} * \frac{P(\text{plavalec} = \text{"DA"} | \text{pazniki} = \text{"DA"})}{P(\text{plavalec} = \text{"DA"})} = \\ = 0.625 * \frac{0.6805}{0.625} * \frac{0.5694}{0.625} * \frac{0.6406}{0.625} = 0.6354 \end{aligned}$$

to ni verjetnost, je le ocena!

Če nas zanimajo dejanske verjetnosti, je to oceno potrebno normalizirati:

izračunamo nasprotno oceno verjetnosti

$$\begin{aligned} P(\text{plavalec} = \text{"NE"}) * \frac{P(\text{plavalec} = \text{"NE"} | \text{kopno} = \text{"NE"})}{P(\text{plavalec} = \text{"NE"})} * \frac{P(\text{plavalec} = \text{"NE"} | \text{luna} = \text{"DA"})}{P(\text{plavalec} = \text{"NE"})} * \frac{P(\text{plavalec} = \text{"NE"} | \text{pazniki} = \text{"DA"})}{P(\text{plavalec} = \text{"NE"})} = \\ = 0.375 * \frac{0.3194}{0.375} * \frac{0.4306}{0.375} * \frac{0.3594}{0.375} = 0.3515 \end{aligned}$$

$$\begin{aligned} P(\text{plavalec} = \text{"DA"} | \text{kopno} = \text{"NE"}, \text{luna} = \text{"DA"}, \text{pazniki} = \text{"DA"}) &= \frac{0.6354}{0.6354 + 0.3515} = 0.6438 \\ P(\text{plavalec} = \text{"NE"} | \text{kopno} = \text{"NE"}, \text{luna} = \text{"DA"}, \text{pazniki} = \text{"DA"}) &= \frac{0.3515}{0.6354 + 0.3515} = 0.3562 \end{aligned}$$

Na izpitu bo pisalo ali nas zanimajo le ocene verjetnosti ali dejanske verjetnosti oz. ali rezultat normaliziramo.

2. naloga:

Ko je novopečeni gobar prišel iz gozda, je nesel polno košaro gob, čeprav ni vedel, ali so užitne ali ne. Prosil je izkušenega gobarja, naj jih razdeli v tri košare: v prvo vse užitne, v drugo vse strupene (ki so seveda neužitne) in v tretjo nestrupene a neužitne gobe. Novopečeni gobar je sestavil učno množico tako, da je za vsako gobo zapisal barvo klobuka in barvo beta, nato pa je preštel gobe iz vsake podskupine. Za vsako košaro je dobil po eno tabelo:

UŽITNE	Rdeč klobuk	Rjav klobuk	Bel klobuk
Bel bet	0	50	10
Rjav bet	10	15	5

NEUŽITNE STRUPENE	Rdeč klobuk	Rjav klobuk	Bel klobuk
Bel bet	30	5	20
Rjav bet	0	10	0

NEUŽITNE NESTRUPENE	Rdeč klobuk	Rjav klobuk	Bel klobuk
Bel bet	0	0	0
Rjav bet	5	10	0

Kakšna je verjetnost, da je goba z rjavim betom in rdečim klobukom užitna, če naivni Bayes uporablja m-oceno in je $m = 10$?

Rešitev:

$$P(\text{goba} = \text{"užitna"} | \text{bet} = \text{"rjav"}, \text{klobuk} = \text{"rdeč"}) = ?$$

$$P(\text{goba} = \text{"užitna"}) * \frac{P(\text{goba} = \text{"užitna"} | \text{bet} = \text{"rjav"})}{P(\text{goba} = \text{"užitna"})} * \frac{P(\text{goba} = \text{"užitna"} | \text{klobuk} = \text{"rdeč"})}{P(\text{goba} = \text{"užitna"})}$$

$$P(\text{goba} = \text{"užitna"}) = \frac{90+1}{170+3} = 0.526$$

vsota elementov v užitni tabeli
 apriorna verjetnost
 skupno število primerov, dobimo tako da seštejemo vse verjetnosti v teh treh tabelah
 št. različnih razredov (užitne, neužitne strupene, neužitne nestrupene)

$$P(\text{goba} = \text{"užitna"} | \text{bet} = \text{"rjav"}) = \frac{30+10 * 0.526}{55+10} = 0.5425$$

užitne gobe z rjavim bet-om
 apriorna verjetnost
 seštejemo vse primere, ki imajo rjav bet, ne glede na to v kateri tabeli oz. razredu so
 m

$$P(\text{goba} = \text{"užitna"} | \text{klobuk} = \text{"rdeč"}) = \frac{10+10 * 0.526}{45+10} = 0.2774$$

št. vseh užitnih gob z rdečimi klobuki
 št. vseh gob z rdečimi klobuki

$$P(\text{goba} = \text{"užitna"}) * \frac{P(\text{goba} = \text{"užitna"} | \text{bet} = \text{"rjav"})}{P(\text{goba} = \text{"užitna"})} * \frac{P(\text{goba} = \text{"užitna"} | \text{klobuk} = \text{"rdeč"})}{P(\text{goba} = \text{"užitna"})} =$$

$$= 0.526 * \frac{0.5425}{0.526} * \frac{0.2774}{0.526} = 0.2861$$

ocena verjetnosti

Nasprotne ocene verjetnosti (za normalizacijo):

$$P(\text{goba} = \text{"užitna"}) * \frac{P(\text{goba} = \text{"užitna"} | \text{bet} = \text{"rjav"})}{P(\text{goba} = \text{"užitna"})} * \frac{P(\text{goba} = \text{"užitna"} | \text{klobuk} = \text{"rdeč"})}{P(\text{goba} = \text{"užitna"})} = 0.2861$$

$$P(\text{goba} = \text{"neužitna"}) * \frac{P(\text{goba} = \text{"neužitna"} | \text{bet} = \text{"rjav"})}{P(\text{goba} = \text{"neužitna"})} * \frac{P(\text{goba} = \text{"neužitna"} | \text{klobuk} = \text{"rdeč"})}{P(\text{goba} = \text{"neužitna"})} = 0.0925 * \frac{0.245}{0.0925} * \frac{0.1077}{0.0925} = 0.2853$$

$$P(\text{goba} = \text{"strupena"}) * \frac{P(\text{goba} = \text{"strupena"} | \text{bet} = \text{"rjav"})}{P(\text{goba} = \text{"strupena"})} * \frac{P(\text{goba} = \text{"strupena"} | \text{klobuk} = \text{"rdeč"})}{P(\text{goba} = \text{"strupena"})} = 0.3815 * \frac{0.2125}{0.3815} * \frac{0.6148}{0.3815} = 0.3425$$

Če nas zanima zaprt sistem dogodkov:

$$0.2861 + 0.2853 + 0.3425 = 0.9139$$

Normalizacija:

$$P(\text{goba} = \text{"užitna"} | \text{bet} = \text{"rjav"}, \text{klobuk} = \text{"rdeč"}) = \frac{0.2861}{0.9139} = 0.3131$$

$$P(\text{goba} = \text{"neužitna"} | \text{bet} = \text{"rjav"}, \text{klobuk} = \text{"rdeč"}) = \frac{0.2853}{0.9139} = 0.3122$$

$$P(\text{goba} = \text{"strupena"} | \text{bet} = \text{"rjav"}, \text{klobuk} = \text{"rdeč"}) = \frac{0.3425}{0.9139} = 0.3748$$

Tukaj bi zaprt sistem dogodkov prišel 1.

Naloga 3:

V morju je potapljač naletel na razbitino gusarske ladje in v notranjosti je bil zaboj poln cekinov. Ker je bil zaboj pretežak, je vzel s seboj samo 101 naključno izbranih cekinov. Na obali je pregledal cekine in ugotovil, da jih je 50 srebrnih, 30 bronastih in 20 zlatih, enega pa je izgubil. Pri srebrnih je ugotovil, da ima polovica vtisnjen simbol Črnega gusarja, pri bronastih tretjina in pri zlatih tri četrtine. Ker je posumil, da so nekateri ponarejeni, jih je še stehtal. Od srebrnih jih je bila petina lažjih (ponarejenih) in vsi so imeli vtisnjen simbol gusarja, od bronastih ni bil nobeden ponarejen in od zlatih polovica in vsi so imeli vtisnjen simbol gusarja. Naj bo to naša učna množica (2 atributa: vrsta cekina, simbol gusarja; 2 razreda: pravi/ponarejen; 100 učnih primerov).

- Nariši odločitveno drevo, ki vsebuje celotno zgoraj opisano informacijo, torej v vsakem listu mora biti napisano število cekinov, ki ustreza temu listu in verjetnost pravilnega odgovora. Pri gradnji drevesa za izbiro atributa uporabi informacijski prispevek,
- Z odločitvenim drevesom klasificiraj 101. cekin, ki ga je potapljač kasneje našel na obali, kjer ga je bil izgubil. Cekin je bil srebrn z vtisnjem simbolom gusarja.

Rešitev:

a) Iz podatkov zgradimo odločitveno drevo.

Vrsta	Simbol	Razred ponaredek / pravi	
srebrn 50	DA 25	10	15
	NE 25	0	25
bronast 30	DA 10	0	10
	NE 20	0	20
zlat 20	DA 15	10	5
	NE 5	0	5



Niso vsi primeri v istem razredu, zato
poiščemo najbolj ustrezeno razbitje učne
množice

$$H_R = - \sum_k p_k \log p_k \quad \text{logaritem je log}_2, \text{ ne log}_10$$

$$H_{R|A} = - \sum_j p_j \sum_k p_{kj} \log p_{kj}$$

$$\text{InfGain}(A) = H_R - H_{R|A}$$

Informacijski prispevek za atribut vrsta:

$$P(\text{Razred} = \text{"ponaredek"}) = 20/100 \quad \begin{matrix} \text{apriorna verjetnost razreda ponaredek} \\ \text{navadna relativna frekvenca} \end{matrix}$$

$$P(\text{Razred} = \text{"pravi"}) = 80/100$$

$$H(\text{Razred}) = -(20/100 * \log_2(20/100) + 80/100 * \log_2(80/100)) = 0.722$$

$$P(\text{Vrsta} = \text{"srebrn"}) = 50/100$$

$$P(\text{Razred} = \text{"ponaredek"} | \text{Vrsta} = \text{"srebrn"}) = 10/50$$

$$P(\text{Razred} = \text{"pravi"} | \text{Vrsta} = \text{"srebrn"}) = 40/50$$

$$P(\text{Vrsta} = \text{"bronast"}) = 30/100$$

$$P(\text{Razred} = \text{"ponaredek"} | \text{Vrsta} = \text{"bronast"}) = 0/30$$

$$P(\text{Razred} = \text{"pravi"} | \text{Vrsta} = \text{"bronast"}) = 30/30$$

$$P(Vrsta="zlat") = 20/100$$

$$P(Razred="ponarededek" | Vrsta="zlat") = 10/20$$

$$P(Razred="pravi" | Vrsta="zlat") = \textcolor{red}{10/20}$$

$$\begin{aligned} 0 * \log_2(0) &= 0 \\ 1 * \log_2(1) &= 0 \end{aligned}$$

$$\begin{aligned} H(Razred | Vrsta) = & -(50/100 * (10/50 * \log_2(10/50) + 40/50 * \log_2(40/50)) + \\ & 30/100 * (0/30 * \log_2(0/30) + 30/30 * \log_2(30/30)) + \\ & 20/100 * (10/20 * \log_2(10/20) + 10/20 * \log_2(10/20))) = \textcolor{red}{0.561} \end{aligned}$$

$$InfGain(Vrsta) = 0.722 - 0.561 = \textcolor{red}{0.161}$$

Informacijski prispevek za atribut simbol:

$$P(Simbol="DA") = 50/100$$

$$P(Razred="ponarededek" | Simbol="DA") = 20/50$$

$$P(Razred="pravi" | Simbol="DA") = \textcolor{red}{30/50}$$

$$P(Simbol="NE") = 50/100$$

$$P(Razred="ponarededek" | Simbol="NE") = 0/50$$

$$P(Razred="pravi" | Simbol="NE") = \textcolor{red}{50/50}$$

$$\begin{aligned} H(Razred | Simbol) = & -(50/100 * (20/50 * \log_2(20/50) + 30/50 * \log_2(30/50)) + \\ & 50/100 * (0/50 * \log_2(0/50) + 50/50 * \log_2(50/50))) = 0.485 \end{aligned}$$

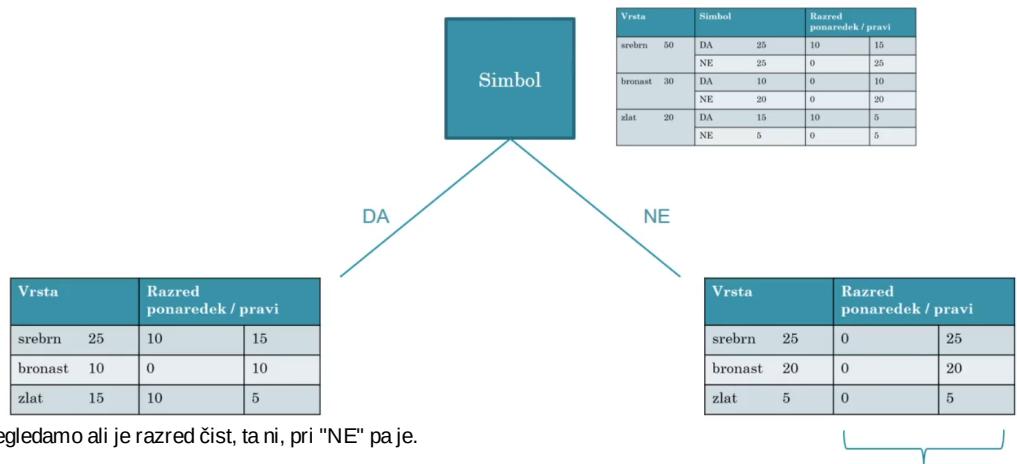
$$InfGain(Simbol) = 0.722 - 0.485 = \textcolor{red}{0.237}$$

Izbira atributa:

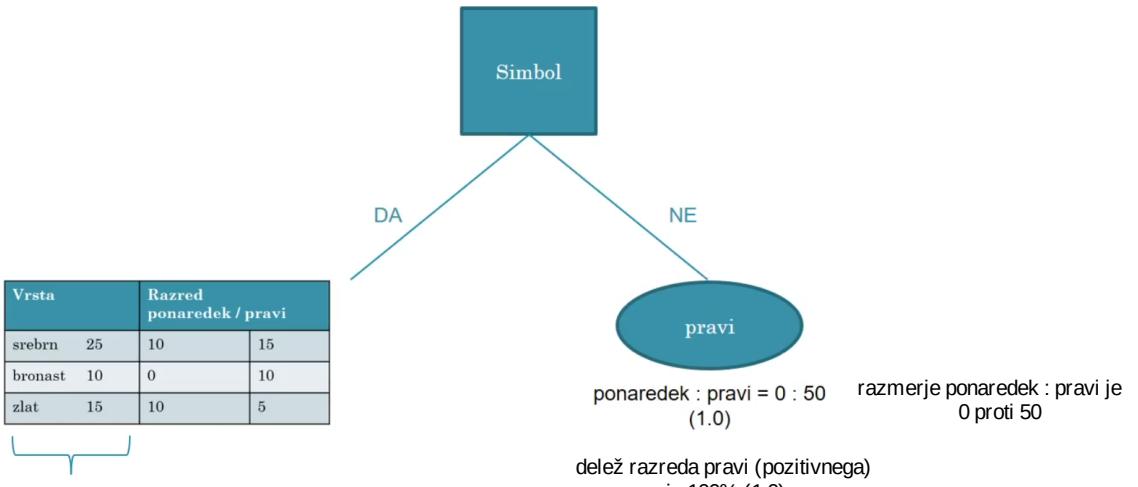
$$InfGain(Vrsta) = 0.722 - 0.561 = 0.161$$

$$InfGain(Simbol) = 0.722 - 0.485 = 0.237 \quad \leftarrow \quad \text{izberemo}$$

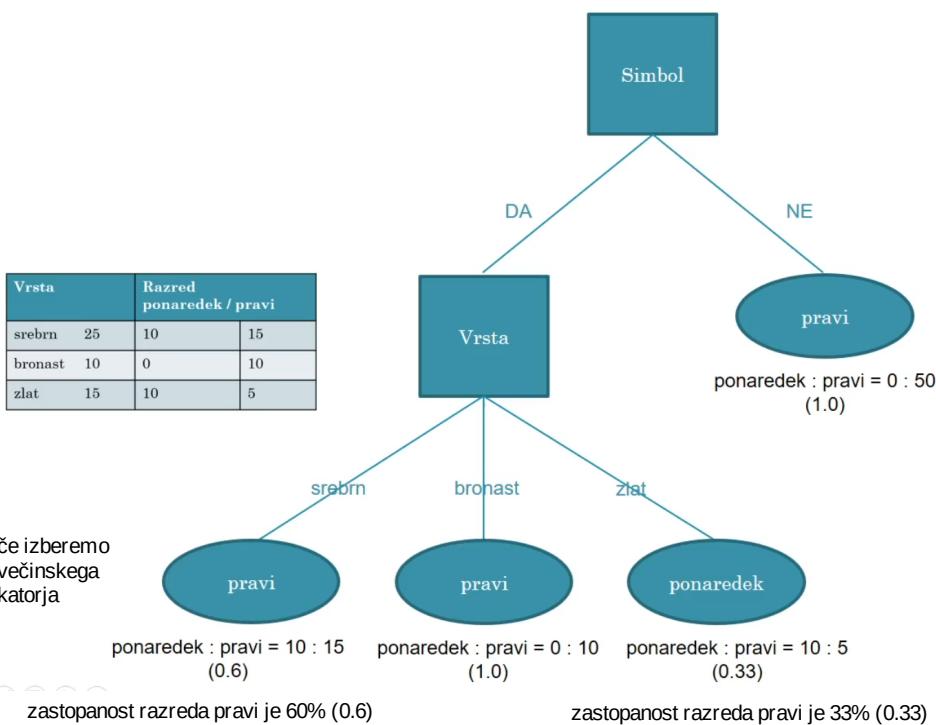
Gradnja drevesa (oz. poddreves):



vsi primeri so iz razreda "pravi",
zato takoj naredimo list

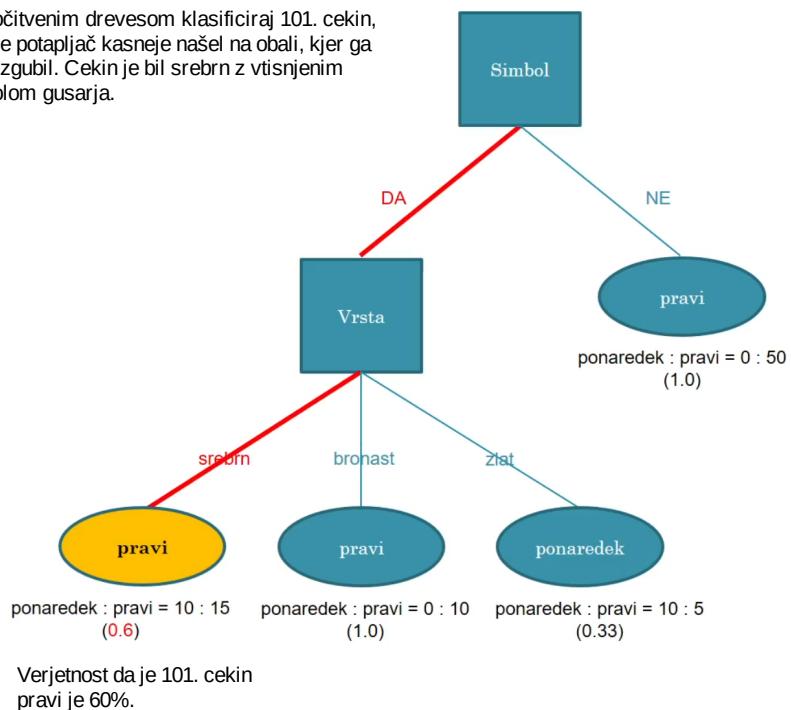


ostal je en sam atribut ("Vrsta"),
zato takoj naredimo notranje
vozlišče



b)

Z odločitvenim drevesom klasificiraj 101. cekin, ki ga je potapljač kasneje našel na obali, kjer ga je bil izgubil. Cekin je bil srebrn z vtisnjениm simbolom gusarja.



Naloga 4:

Da bi zagovili likvidnost banke, so začeli zbirati podatke o posojilojemalcih in o tem, ali je bilo posojilo vrnjeno. Zbrali so podatke v spodnji tabeli.

Posojilo vrnjeno	Spol	Trenutno zakreditiran	Znesek > 100K	V kazenskem postopku
NE	MOŠKI	NE	DA	DA
DA	MOŠKI	NE	DA	NE
NE	ŽENSKA	DA	NE	NE
DA	ŽENSKA	DA	DA	NE
NE	MOŠKI	DA	NE	NE
DA	MOŠKI	NE	NE	DA
NE	MOŠKI	NE	DA	DA
DA	MOŠKI	DA	DA	NE

Direktor banke se je odločil, da bo napovedoval, ali bo posojilo vrnjeno, kar s pomočjo odločitvenega drevesa, ki ga je zgradil z uporabo gini-indeksa iz zgornje podatkovne baze o posojilojemalcih.

a) Nariši direktorjevo odločitveno drevo.

b) Kakšna je napoved za direktorjevo ženo, ki ni v kazenskem postopku, nikoli ni imela kreditov in bi si sposodila 20 000 evrov?

Rešitev:

$$Gini(A) = \sum_j p_{.j} \sum_k p_{k|j}^2 - \sum_k p_k^2$$

$$\sum_k p_k^2$$

$$P(Vrnjeno = "DA") = 4/8$$

$$P(Vrnjeno = "NE") = 4/8$$

$$\begin{aligned} \sum_{k,j} p_{k,j}^2 & P(Spol = "MOŠKI") = 6/8 \\ \sum_j p_{.,j} & P(Vrnjeno = "DA" | Spol = "MOŠKI") = 3/6 \\ & P(Vrnjeno = "NE" | Spol = "MOŠKI") = 3/6 \end{aligned}$$

$$\begin{aligned} \sum_{k,j} p_{k,j}^2 & P(Spol = "ŽENSKA") = 2/8 \\ \sum_j p_{.,j} & P(Vrnjeno = "DA" | Spol = "ŽENSKA") = 1/2 \\ & P(Vrnjeno = "NE" | Spol = "ŽENSKA") = 1/2 \end{aligned}$$

$$Gini(Spol) = \frac{6}{8} * \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2 \right) + \frac{2}{8} * \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) - \left(\left(\frac{4}{8}\right)^2 + \left(\frac{4}{8}\right)^2 \right) = 0$$

$P(Vrnjeno = "DA")$

$P(Vrnjeno = "NE")$

$$Gini(Zakreditiran) = \frac{4}{8} * \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right) + \frac{4}{8} * \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right) - \left(\left(\frac{4}{8}\right)^2 + \left(\frac{4}{8}\right)^2 \right) = 0$$

$P(Zakreditiran = "DA")$

$P(Vrnjeno = "DA" | Zakreditiran = "DA")$

$P(Vrnjeno = "NE" | Zakreditiran = "DA")$

$P(Vrnjeno = "DA" | Zakreditiran = "NE")$

$P(Vrnjeno = "NE" | Zakreditiran = "NE")$

$$Gini(Znesek) = \frac{5}{8} * \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) + \frac{3}{8} * \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right) - \left(\left(\frac{4}{8}\right)^2 + \left(\frac{4}{8}\right)^2 \right) = 0.033$$

$P(Znesek = "DA")$

$P(Vrnjeno = "DA" | Znesek = "DA")$

$P(Vrnjeno = "NE" | Znesek = "DA")$

$P(Vrnjeno = "DA" | Znesek = "NE")$

$P(Vrnjeno = "NE" | Znesek = "NE")$

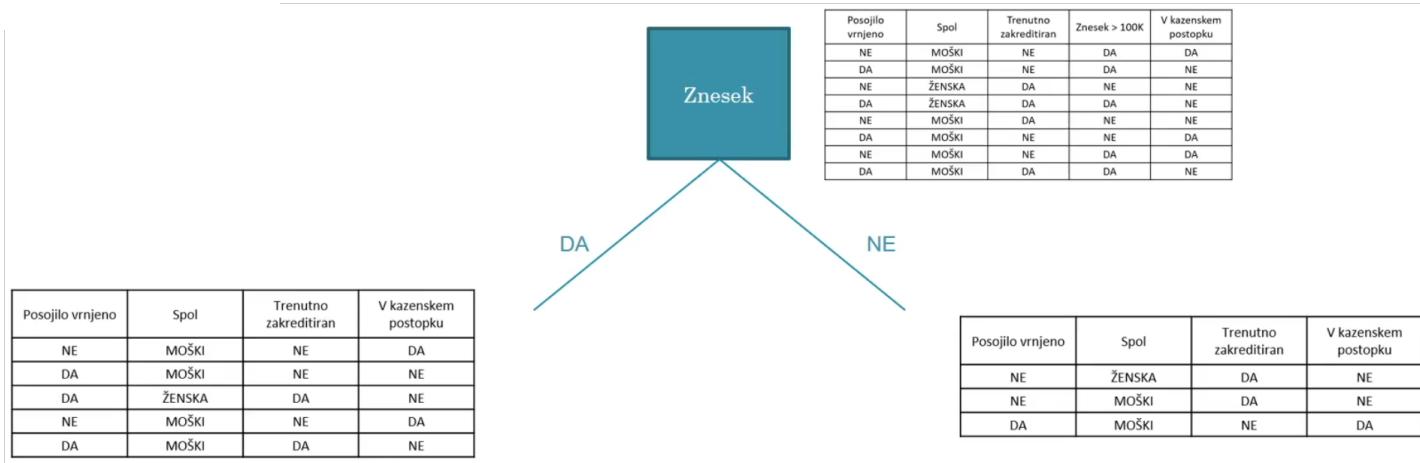
$$Gini(Kazenski) = \frac{3}{8} * \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right) + \frac{5}{8} * \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) - \left(\left(\frac{4}{8}\right)^2 + \left(\frac{4}{8}\right)^2 \right) = 0.033$$

$P(Kazenski = "DA")$

...

$Gini(Spol) = 0$
 $Gini(Zakreditiran) = 0$
 $Gini(Znesek) = 0.033$ ← izberemo
 $Gini(Kazenski) = 0.033$ ni pomembno katerega od obeh 0.033 izberemo

Naredimo drevo:



Pregledamo če imamo čist razred, vendar ga nimamo.

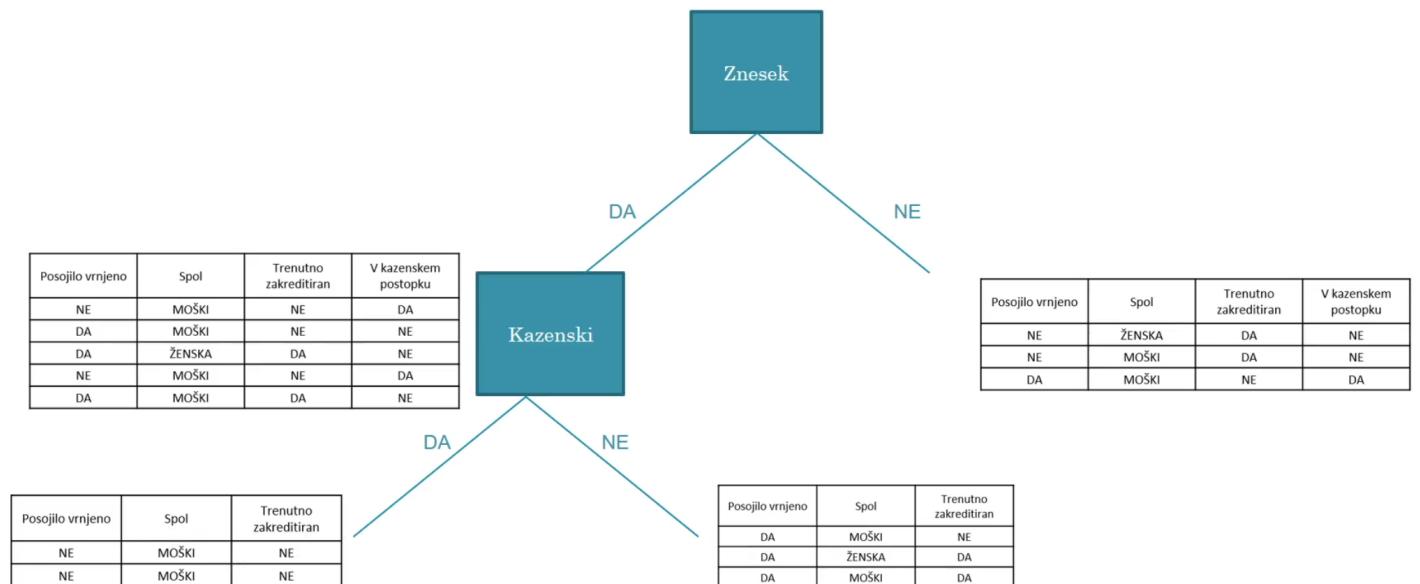
Pregledamo če je razred čist, vendar ni.

Za oba razreda moramo torej ponovno narediti gini index.

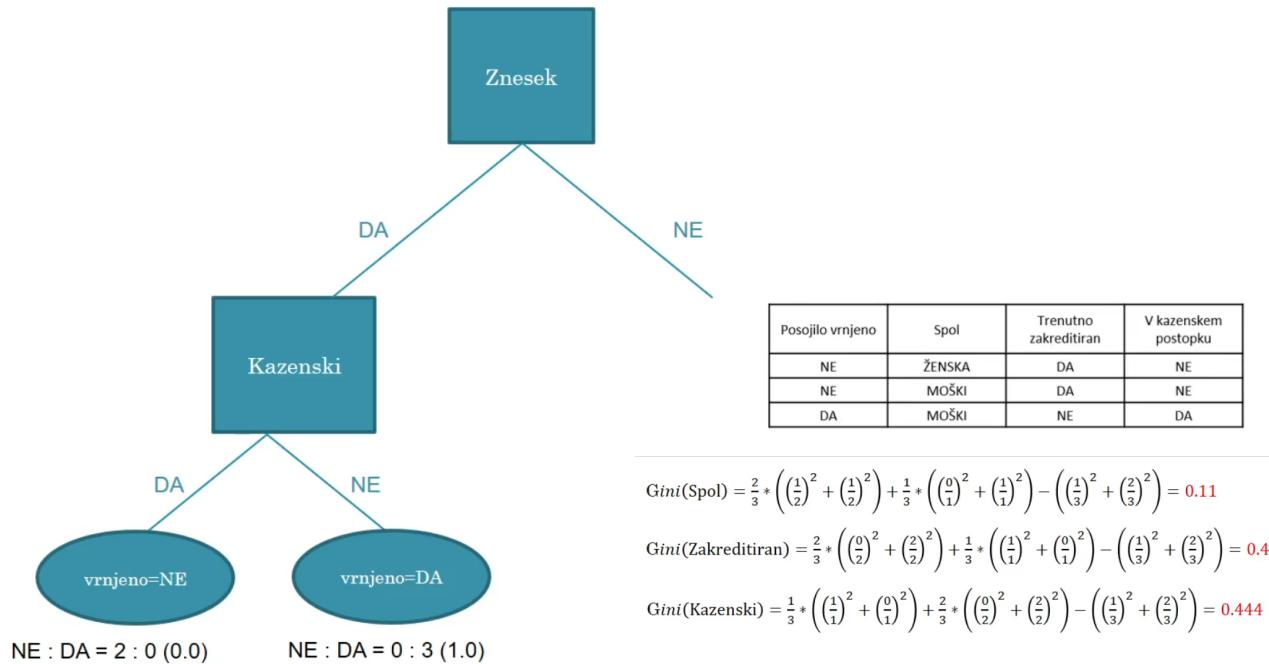
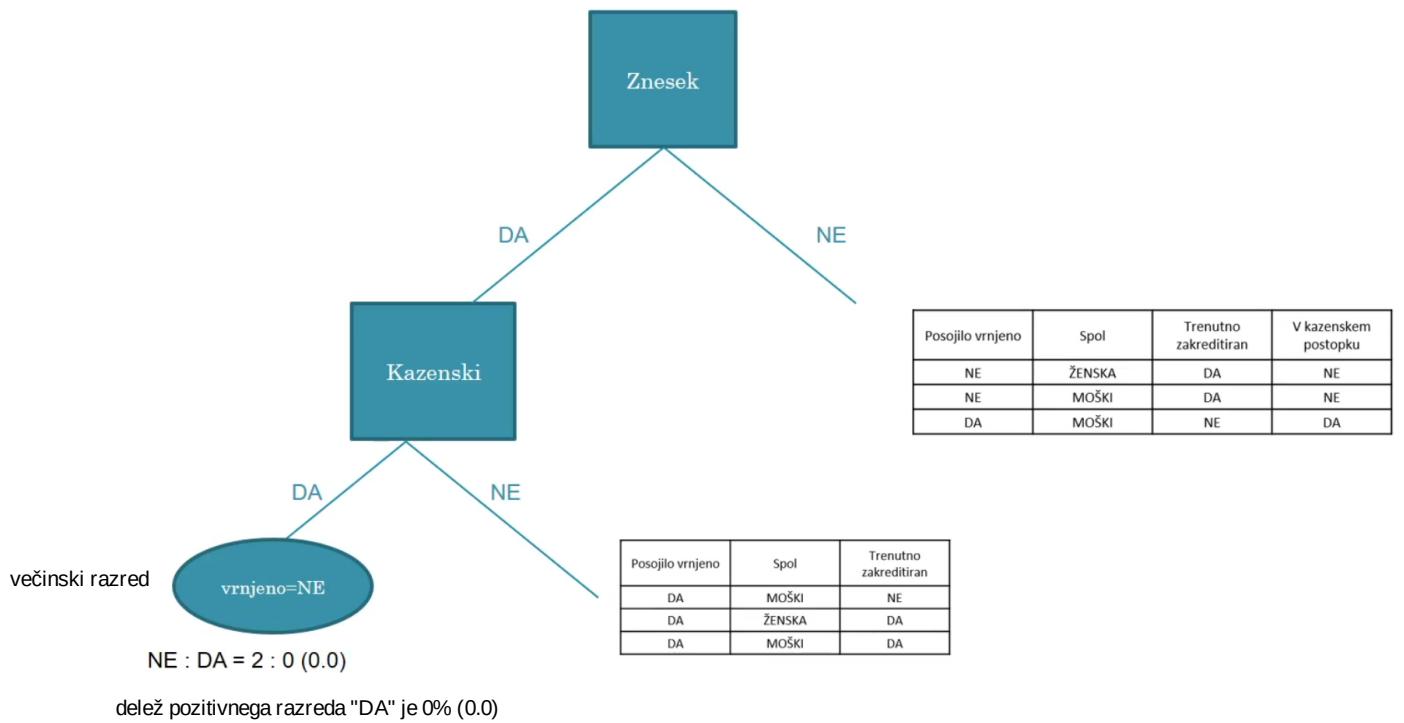
$$Gini(Spol) = \frac{4}{5} * \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 \right) + \frac{1}{5} * \left(\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2 \right) - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) = 0.08$$

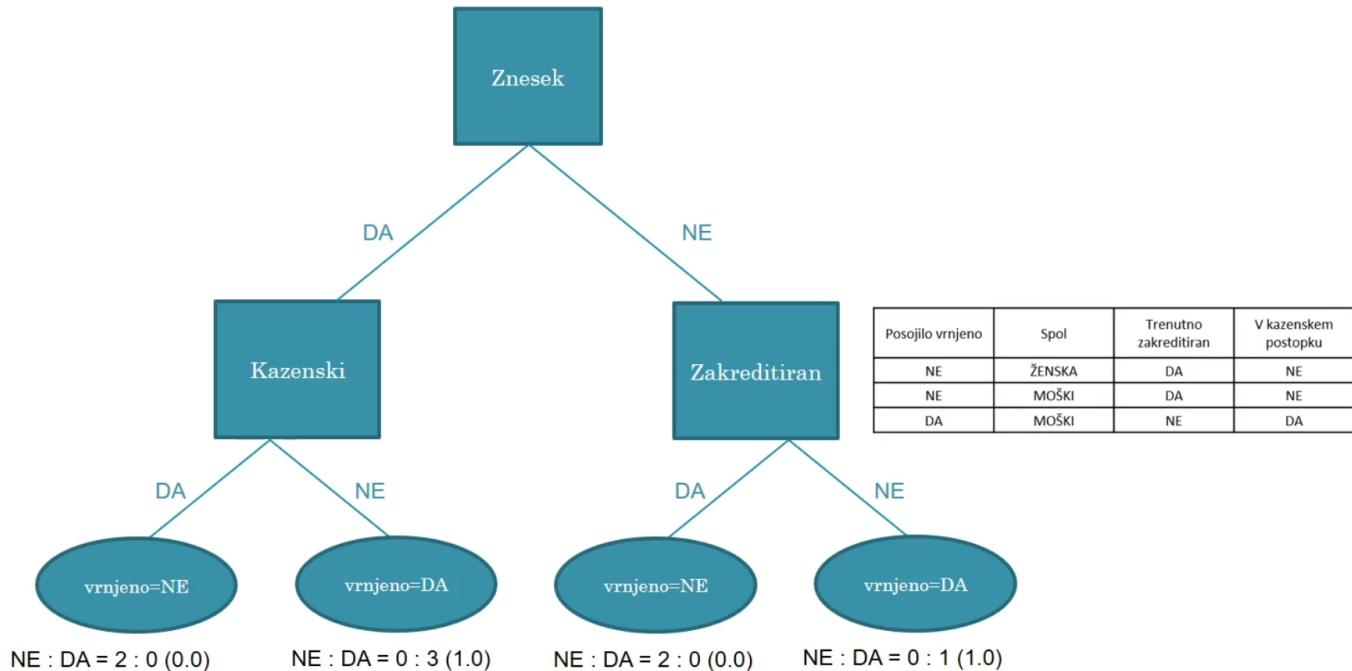
$$Gini(Zakreditiran) = \frac{2}{5} * \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2 \right) + \frac{3}{5} * \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2 \right) - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) = 0.213$$

$$Gini(Kazenski) = \frac{2}{5} * \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2 \right) + \frac{3}{5} * \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2 \right) - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) = 0.48$$
 ← izberemo

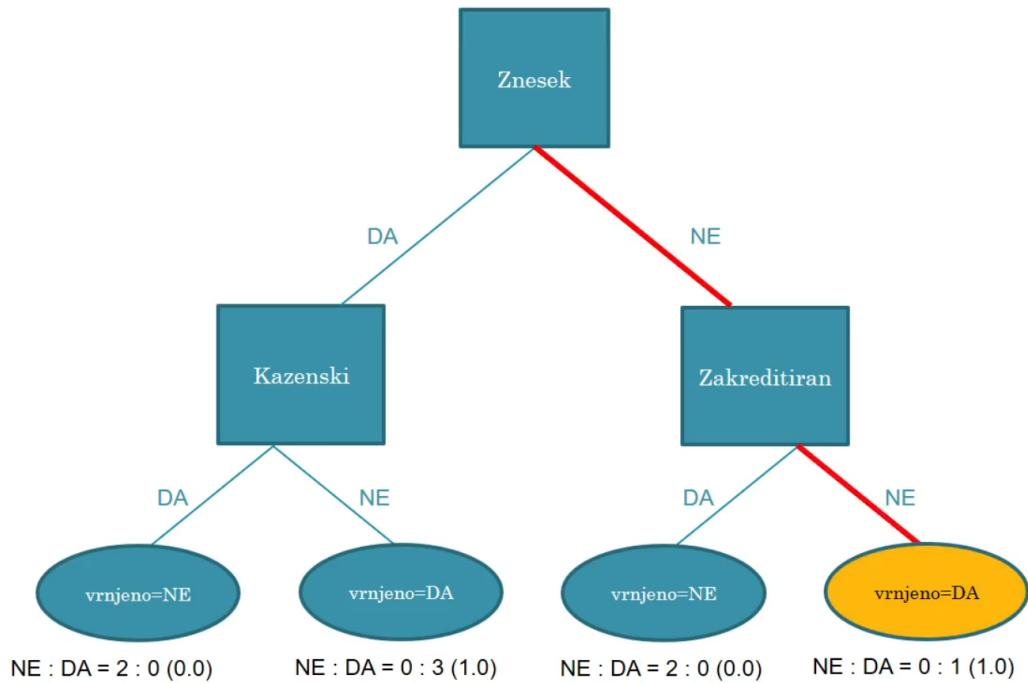


Pregledamo če imamo čist razred, in ga imamo.



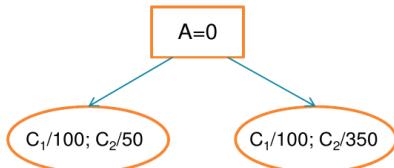


b) Kakšna je napoved za direktorjevo ženo, ki ni v kazenskem postopku, nikoli ni imela kreditov in bi si sposodila 20 000 evrov?



Naloga 5:

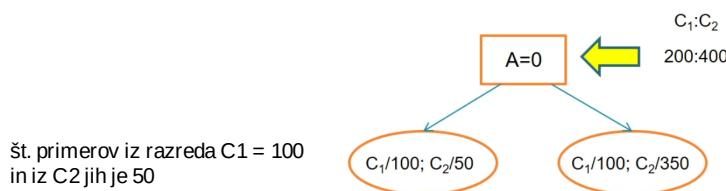
Spodaj je odločitveno drevo s frekvencami primerov v listih za razreda C1 in C2:



Izračunaj pričakovano klasifikacijsko točnost drevesa z m-oceno, če je:

- a) = 0,
- b) = 1000

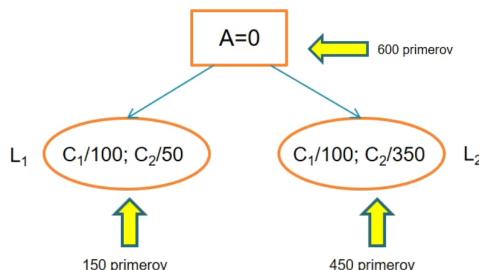
Rešitev:



Število primerov: $200 + 400 = 600$

$$\text{Apriorna verjetnost razredov: } p_a(C_1) = \frac{200}{600} = 0.33$$

$$p_a(C_2) = \frac{400}{600} = 0.67$$



a):

$$p(C_1|L_1) = \frac{r+m*p_a}{n+m} = \frac{100+0*0.33}{150+0} = 0.67 \quad p(C_2|L_1) = \frac{r+m*p_a}{n+m} = \frac{350+0*0.67}{450+0} = 0.78$$



Statična točnost lista = verjetnost klasifikacije v pravilni razred

$$\text{Pričakovana točnost odločitvenega drevesa: } \frac{150}{600} * 0.67 + \frac{450}{600} * 0.78 = 0.75$$

verjetnost, da pridemo v list L₁, pomnožena s točnostjo lista L₁

verjetnost, da pridemo v list L₂, pomnožena s točnostjo lista L₂

b) $m = 1000$

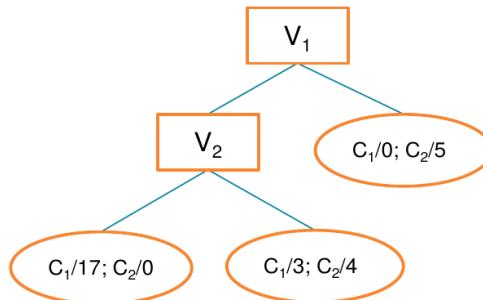
$$p(C_1|L_1) = \frac{r+m*p_a}{n+m} = \frac{100+1000*0.33}{150+1000} = 0.37 \quad p(C_2|L_1) = \frac{r+m*p_a}{n+m} = \frac{350+1000*0.67}{450+1000} = 0.7$$

Pričakovana točnost odločitvenega drevesa: $\frac{150}{600} * 0.37 + \frac{450}{600} * 0.7 = 0.62$

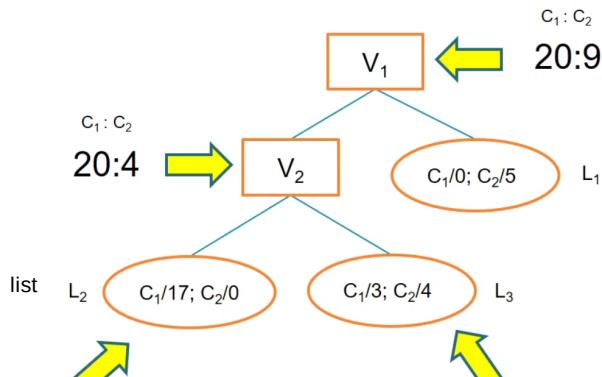
Naloga 6:

Podano je odločitveno drevo za klasifikacijo v razrede z naslednjimi apriornimi verjetnostmi razredov: $p_a(C1)=0.6$ in $p_a(C2)=0.4$.

Obreži podano odločitveno drevo s postopkom minimizacije napake in vrednostjo $m = 10$.



Rešitev:



Statična točnost lista L_2 :

$$p(C_1|L_2) = \frac{r+m*p_a}{n+m} = \frac{17+10*0.6}{17+10} = 0.852$$

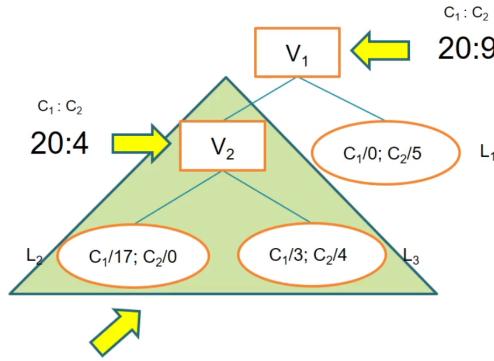
Statična točnost lista L_3 :

$$p(C_2|L_3) = \frac{r+m*p_a}{n+m} = \frac{4+10*0.4}{7+10} = 0.471$$

uporabljen je večinski razred (C1)

uporabljen je večinski razred (C2)

Izračunajmo še vzvratno točnost vozlišča V2.



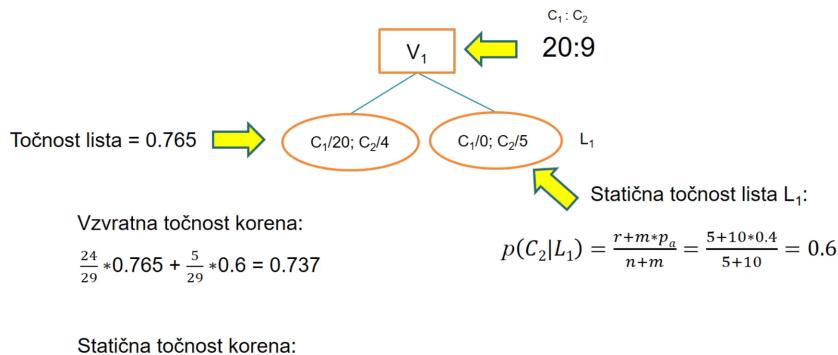
Vzvratna točnost v notranjem vozlišču V_2 :

$$\frac{17}{24} * 0.852 + \frac{7}{24} * 0.471 = 0.741$$

Statična točnost notranjega vozlišča V_2 :

$$p(C_1|V_2) = \frac{r+m*p_a}{n+m} = \frac{20+10*0.6}{24+10} = 0.765$$

Ker je statična točnost notranjega vozlišča (0.765) večja od vzvratne točnosti (0.741) → drevo porežemo.



Vzvratna točnost korena:

$$\frac{24}{29} * 0.765 + \frac{5}{29} * 0.6 = 0.737$$

$$p(C_2|L_1) = \frac{r+m*p_a}{n+m} = \frac{5+10*0.4}{5+10} = 0.6$$

Statična točnost korena:

$$p(C_1|V_1) = \frac{r+m*p_a}{n+m} = \frac{20+10*0.6}{29+10} = 0.67$$

Tokrat je vzvratna točnost (0.737) večja od statične točnosti (0.67) → drevesa ne porežemo.

Naloga 7:

Klasifikator je na 4-razrednem problemu dosegel naslednje rezultate, predstavljene z matriko zmot. V vsaki celici je vneseno ustrezno število_testnih_primerov/cena_napačne_klasifikacije:

Pravi razred	Napovedani razred			
	C ₁	C ₂	C ₃	C ₄
C ₁	12/0	0/1	4/2	4/2
C ₂	5/1	12/0	2/1	5/3
C ₃	5/10	0/3	20/0	15/2
C ₄	8/2	0/1	2/4	46/0

Izračunaj:

- klasifikacijsko točnost klasifikatorja,
- pričakovano točnost večinskega klasifikatorja (predpostavi, da je verjetnostna distribucija po razredih v testni množici enaka distribuciji v učni množici),
- povprečno ceno napačne klasifikacije.

Rešitev:

Testna množica vsebuje:

$$12 + 4 + 4 = 20 \text{ primerov iz razreda } C_1$$

$$5 + 12 + 2 + 5 = 24 \text{ primerov iz razreda } C_2$$

$$5 + 20 + 15 = 40 \text{ primerov iz razreda } C_3$$

$$8 + 2 + 46 = 56 \text{ primerov iz razreda } C_4$$

Skupaj je to $20 + 24 + 40 + 56 = 140$ primerov

a) klasifikacijska točnost klasifikatorja

$$(12 + 12 + 20 + 46) / 140 = 0.643$$

b) pričakovana točnost večinskega klasifikatorja

večinski razred je C_4 , zato je točnost več. klas. $= 56 / 140 = 0.4$

c) povprečna cena napačne klasifikacije

$$\begin{aligned} & (\\ & 12*0 + 0*1 + 4*2 + 4*2 + \\ & 5*1 + 12*0 + 2*1 + 5*3 + \\ & 5*10 + 0*3 + 20*0 + 15*2 + \\ & 8*2 + 0*1 + 2*4 + 46*0 \\ &) / 140 = 1.014 \end{aligned}$$