INTELIGENTNI SISTEMI

as. ms Vladimir Jocović
as. ms Adrian Milaković

STABLA ODLUČIVANJA

06

„Random forests are like democracy—
many voices lead to better decisions..“
- Unknown author

STABLA ODLUČIVANJA

Šta su stabla odlučivanja?

Stabla odlučivanja su tip nadgledanih algoritama mašinskog učenja. Najčešće se koriste kod problema klasifikacije (mada mogu da se koriste i kod problema regresije).

Stablo odlučivanja je stablo u kojem je:

- Svakom unutrašnjem čvoru pridružen jedan ulazni parameter (atribut)
- Svakoj grani pridružena jedna vrednost ulaznog parametra čvora od kojeg grana počinje
- Svakom listu pridružen izlazni parametar u zavisnosti od vrednosti ulaznih parametara na datom putu kroz stablo, od korena do posmatranog lista

Ulazni parametri mogu da budu i kontinualnog i kategoričkog tipa.

Prema tipu izlazne vrednosti razlikujemo klasifikaciona i regresivna stabla.

STABLA ODLUČIVANJA

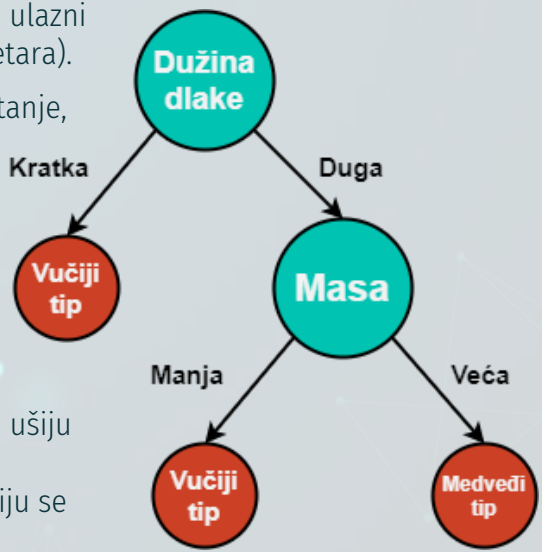
Npr. *Predviđanje tipa samojeda na osnovu nekoliko atributa.*

Stablo se koristi kao klasifikator tako što uzima ulazni primer, dat kao vektor atributa (ulaznih parametara).

Atribut koji je koreni čvor se interpretira kao pitanje, a odgovor se određuje na osnovu vrednosti tog atributa iz ulaznog primera.

Odgovor određuje koji čvor naslednik se posećuje.

Pitanja se ponavljaju dok se ne dođe do lista koji određuje izlaznu vrednost.



Pr. Samojed duge dlake, veće mase i zaobljenih ušiju se klasifikuje kao medveđi tip.
Pr. Samojed kratke dlage, veće mase i oštrih ušiju se klasifikuje kao vučiji tip.

STABLA ODLUČIVANJA

Izgradnja stabla (ID3 algoritam)

Stablo se gradi od korena ka listovima, tzv. pohlepnim (*greedy*) pristupom. Na početku sve instance prostora pripadaju istom skupu, nakon čega se prostor sukcesivno deli na podskupove. Za deljenje kažemo da je pohlepno zato što se pri svakom koraku najbolja podela određuje na osnovu stanja u posmatranom koraku, odnosno ne uzima se u obzir kako će se podela izvršiti u narednim koracima i koja bi podela mogla dovesti do boljih rezultata u narednim koracima.

U svakoj iteraciji se bira najbolji atribut na osnovu kojeg će stablo da se podeli u podskupove. Za svaku moguću vrednost izabranog atributa se prostor deli u podskupve, koji odgovaraju odabranoj vrednosti. Algoritam se rekurzivno ponavlja dok se ne dosegne unapred definisan kriterijum zaustavljanja ili dok se ne dostignu svi listovi.

STABLA ODLUČIVANJA

Kako odabrati najbolji atribut za podelu?

U zavisnosti od toga da li se gradi regresivno ili klasifikaciono stablo, kriterijum na osnovu kojeg se vrši podela može da bude različit.

Postoji nekoliko pristupa pri izboru najboljeg atributa za podelu:

- *Random* – Atribut se bira slučajnim izborom.
- *Least-values* – Bira se atribut sa najmanjim brojem mogućih vrednosti.
- *Most-values* – Bira se atribut sa najvećim brojem mogućih vrednosti.
- *Max-gain* – Bira se atribut koji je po nekom kriterijumu najbolji za podelu (ima najveću dobit informacija ili najmanji Gini indeks).

KLASIFIKACIONA STABLA

Šta je i kako se određuje dobit informacija?

Dobit informacija (*information gain*) računa razliku u entropiji pre i nakon podele skupa podataka na neki način.

Za podelu po svim atributima se računa dobit informacija i bira onaj atribut koji maksimizuje tu dobit.

Npr. Petar je zamislio broj od 1 do 100. Miloš ima pravo da Petru postavlja da/ne pitanja na osnovu kojih treba da pogodi broj koji je Petar zamislio. Koliko najmanje pitanja Miloš treba da postavi da bi pogodio broj koji je Petar zamislio?

Rešenje: 7 pitanja

Sa svakim da/ne pitanjem, optimalno bi trebalo odbaciti najviše ½ od preostalih brojeva ($\log_2 100 = 6.64$).

Ako je dat skup S veličine $|S|$, očekivan broj pokušaja radi određivanja konkretnog elementa je $\log_2 |S|$.

KLASIFIKACIONA STABLA

Šta je i kako se određuje entropija?

Entropija skupa primera, S, za neku binarnu klasifikaciju (ishod ima samo dve klase) je:

$$Entropy(S) = -p_1 * log_2(p_1) - p_2 * log_2(p_2)$$

gde je p_1 procenat pojavljivanja prve klase u skupu S, a p_2 procenat pojavljivanja druge klase u skupu S.

Ako svi primeri pripadaju istoj klasi, entropija je 0 (koristi se jednakost $0 * \log(0) = 0$).

Ako su primeri jednako raspoređeni ($p_1 = p_2 = 0.5$), entropija ima maksimalnu vrednost 1.

Entropiju možemo posmatrati kao prosečan broj bitova potreban za kodiranje klase primera iz skupa S, gde su češćim slučajevima kompresijom dodeljeni kraći kodovi.

Za probleme gde postoji više klasa ishoda, entropija se računa kao:

$$Entropy(S) = \sum_{i=1}^{c} -p_i * \log_2(p_i)$$

gde je *c* broj klasa.

KLASIFIKACIONA STABLA

Dobit informacija atributa *F* je očekivano smanjenje u entropiji koje proizilazi iz podele skupa S po ovom atributu.

$$Gain(S, F) = Entropy(S) - \sum_{v \in Values(F)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Drugi sabirak predstavlja sumu entropija svakog rezultujućeg podskupa nastalog iz podele skupa S po atributu F za svaku od vrednosti V, pomnoženu svojom relativnom veličinom u odnosu na početni skup S.

Npr. Skup podataka (2 banane, 2 kruške -> Entropija = 1):

[izdužen oblik, žut] -> banana; [izdužen oblik, zelen] -> banana;
[spljošten oblik, žut] -> kruška; [spljošten oblik, zelen] -> kruška;

Podela po obliku:
izdužen oblik: 2 banane, 0 kruški (Entropija = 0)
spljošten oblik: 0 banana, 2 kruške (Entropija = 0)
Gain = 1 – (0.5 * 0 + 0.5 * 0) = 1

Podela po boji:
izdužen oblik: 1 banana, 1 kruška (Entropija = 1)
spljošten oblik: 1 banana, 1 kruška (Entropija = 1)
Gain = 1 – (0.5 * 1 + 0.5 * 1) = 0

Bolja je podela po obliku!

KLASIFIKACIONA STABLA

Šta je i kako se određuje Gini indeks?

Gini indeks skupa primera, S, za neku binarnu klasifikaciju (ishod ima samo dve klase) je:

$$Gini(S) = 1 - p_1^2 - p_2^2$$

gde je p_1 procenat pojavljivanja prve klase u skupu S, a p_2 procenat pojavljivanja druge klase u skupu S.

Ako svi primeri pripadaju istoj klasi, indeks je 0.

Ako su primeri jednako raspoređeni ($p_1 = p_2 = 0.5$), indeks ima maksimalnu vrednost 0.5.

Za probleme gde postoji više klasa ishoda, Gini indeks se računa kao:

$$Gini(S) = 1 - \sum_{i=1}^{c} - p_i^2$$

Ukupan Gini indeks podele skupa na osnovu atributa *F* je dat formulom:

$$Gini(S, F) = \sum_{v \in Values(F)} \frac{|S_v|}{|S|} Gini(S_v)$$

Zadatak 1 – Životinje koje ležu jaja

★★★☆☆

Konstruisati stablo odlučivanja za određivanje da li životinja leže jaja na osnovu sledećeg skupa podataka i njihovih karakteristika.

Nezavisni atributi					Atribut odluke
Životinja	Toplokrvna	Pernata	Ima krzno	Pliva	Leže jaja
Noj	Da	Da	Ne	Ne	Da
Krokodil	Ne	Ne	Ne	Da	Da
Gavran	Da	Da	Ne	Ne	Da
Albatros	Da	Da	Ne	Ne	Da
Delfin	Da	Ne	Ne	Da	Ne
Koala	Da	Ne	Da	Ne	Ne

Zadatak 1 – Rešenje

Prvo ćemo rešiti zadatak korišćenjem Gini indeksa kao kriterijuma podele.

Atribut „toplokrvni“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „toplokrvni“ DA je veličine 5 (noj, gavran, albatros, delfin, koala). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 3 (noj, gavran, albatros), a vrednosti NE je 2 (delfin, koala). Gini indeks takvog podskupa je:

$$Gini(3\ DA, 2\ NE) = 1 - (3/5)^2 - (2/5)^2 = 0.48$$

Podskup skupa S u kojem je vrednost atributa „toplokrvni“ NE je veličine 1 (krokodil). Samo vrednost DA za izlazni atribut „leže jaja“ se u takvom podskupu nalazi. Gini indeks takvog podskupa je:

$$Gini(1\ DA, 0\ NE) = 1 - (1/1)^2 - (0/1)^2 = 0$$

Sada računamo ukupan Gini indeks za podelu po atributu „toplokrvna“:

$$Gini(S, \text{toplokrvna}) = 5/6 * 0.48 + 1/6 * 0 = 0.4$$

Zadatak 1 – Rešenje

Atribut „pernata“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „pernata“ DA je veličine 3 (noj, gavran, albatros). Samo vrednost DA za izlazni atribut „leže jaja“ se u takvom podskupu nalazi. Gini indeks takvog podskupa je:

$$Gini(3\ DA, 0\ NE) = -(3/3)^2 - (0/3)^2 = 0$$

Podskup skupa S u kojem je vrednost atributa „pernata“ NE je veličine 3 (krokodil, delfin, koala). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 1 (krokodil), a vrednosti NE je 2 (delfin, koala). Gini indeks takvog podskupa je:

$$Gini(1\ DA, 2\ NE) = 1 - (1/3)^2 - (2/3)^2 = 0.44444$$

Sada računamo Gini indeks za podelu po atributu „pernata“:

$$Gini(S, pernata) = 3/6 * 0 + 3/6 * 0.44444 = 0.22222$$

Zadatak 1 – Rešenje

Atribut „ima krzno“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „ime krzno“ DA je veličine 1 (koala). Samo vrednost NE za izlazni atribut „leže jaja“ se u takvom podskupu nalazi. Gini indeks takvog podskupa je:

$$Gini(0\ DA, 1\ NE) = 1 - (0/1)^2 - (1/1)^2 = 0$$

Podskup skupa S u kojem je vrednost atributa „ima krzno“ NE je veličine 5 (noj, krokodil, gavran, albatros, delfin). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 4 (noj, krokodil, gavran, albatros), a vrednosti NE je 1 (delfin). Gini indeks takvog podskupa je:

$$Gini(4\ DA, 1\ NE) = 1 - (4/5)^2 - (1/5)^2 = 0.32$$

Sada računamo Gini indeks za podelu po atributu „ima krzno“:

$$Gini(S, \text{ima krzno}) = 1/6 * 0 + 5/6 * 0.32 = 0.26667$$

Zadatak 1 – Rešenje

Atribut „pliva“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „pliva“ DA je veličine 2 (krokodil, delfin). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 1 (krokodil), a vrednosti NE je 1 (delfin). Gini indeks takvog podskupa je:

$$Gini(1\ DA, 1\ NE) = 1 - (1/2)^2 - (1/2)^2 = 0.5$$

Podskup skupa S u kojem je vrednost atributa „pliva“ NE je veličine 4 (noj, gavran, albatros, koala). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 3 (noj, gavran, albatros), a vrednosti NE je 1 (koala). Gini indeks takvog podskupa je:

$$Gini(3\ DA, 1\ NE) = 1 - (3/4)^2 - (1/4)^2 = 0.375$$

Sada računamo Gini indeks za podelu po atributu „pliva“:

$$Gini(S, pliva) = 2/6 * 0.5 + 4/6 * 0.375 = 0.41667$$

Zadatak 1 – Rešenje

Tražimo najbolju podelu.

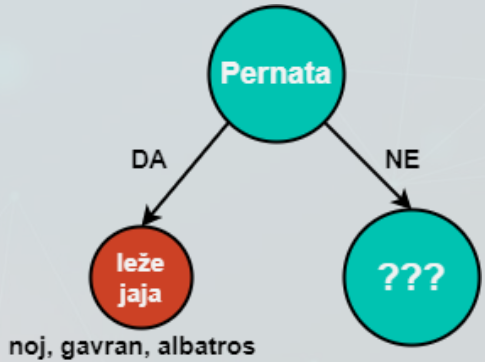
$$Gini(S, \text{toplokrvna}) = 5/6 * 0.48 + 1/6 * 0 = 0.4$$

$$Gini(S, \text{pernata}) = 3/6 * 0 + 3/6 * 0.44444 = 0.22222$$

$$Gini(S, \text{ima krzno}) = 1/6 * 0 + 5/6 * 0.32 = 0.26667$$

$$Gini(S, \text{pliva}) = 2/6 * 0.5 + 4/6 * 0.375 = 0.41667$$

Najmanji Gini indeks ima podela po atributu „pernata“ pa taj atribut biramo za koreni čvor i skup podataka delimo na osnovu tog atributa.



Zadatak 1 – Rešenje

Nakon podele ulaznog skupa podataka po atributu „pernata“, dobijamo sledeći podskup.

Životinja	Toplokrvna	Ima krzno	Pliva	Leže jaja
Krokodil	Ne	Ne	Da	Da
Delfin	Da	Ne	Da	Ne
Koala	Da	Da	Ne	Ne

Za atribut „toplokrvna“ računamo:

$$Gini(0\ DA, 2\ NE) = 1 - (0/2)^2 - (2/2)^2 = 0$$

$$Gini(1\ DA, 0\ NE) = 1 - (1/1)^2 – (0/1)^2 = 0$$

$$Gini(S', \text{toplokrvna}) = 2/3 * 0 + 1/3 * 0 = 0$$

Zadatak 1 – Rešenje

Za atribut „ima krzno“ računamo:

$$Gini(0\ DA, 1\ NE) = 1 - (0/1)^2 - (1/1)^2 = 0$$

$$Gini(1\ DA, 1\ NE) = 1 - (1/2)^2 - (1/2)^2 = 0.5$$

$$Gini(S', \text{ima krzno}) = 1/3 * 0 + 2/3 * 0.5 = 0.33333$$

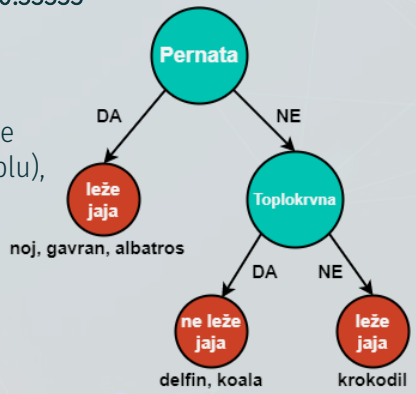
Za atribut „pliva“ računamo:

$$Gini(1\ DA, 1\ NE) = 1 - (1/2)^2 - (1/2)^2 = 0.5$$

$$Gini(0\ DA, 1\ NE) = 1 - (0/1)^2 - (1/1)^2 = 0$$

$$Gini(S', \text{pliva}) = 2/3 * 0.5 + 1/3 * 0 = 0.33333$$

Najbolji Gini indeks ima podela po atributu „toplokrvna“ pa taj čvor biramo za sledeći čvor. Kako skup podataka ne može dalje da se deli (stigli smo do listova po svim putanjama u stablu), konstrukcija stabla je završena.



Zadatak 1 – Rešenje

Sada ćemo isti zadatak da rešimo korišćenjem dobiti informacija kao kriterijuma podele.

U ulaznom skupu S imamo dve vrednosti za atribut „leže jaja“: DA i NE.

Ulazni skup S sadrži četiri reda čija je vrednost izlaznog atributa DA i dva reda čija je vrednost NE. Računamo entropiju takvog skupa.

$$Entropy(4\ DA, 2\ NE) = -\ 4/6 * \log_2(4/6) - 2/6 * \log_2(2/6) = 0.91829$$

Sada je potrebno pronaći dobit informacija za sve atribute: toplokrvna, pernata, ima krzno, pliva (atribut „životinja“ nam ne daje nikakve informacije). Atribut sa najvećom dobiti informacija će biti atribut na osnovu kojeg vršimo prvu podelu ulaznog skupa.

Zadatak 1 – Rešenje

Atribut „toplokrvni“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „toplokrvni“ DA je veličine 5 (noj, gavran, albatros, delfin, koala). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 3 (noj, gavran, albatros), a vrednosti NE je 2 (delfin, koala). Entropija takvog podskupa je:

$$\textit{Entropy}(3\ \text{DA}, 2\ \text{NE}) = -\ 3/5 * \log_2(3/5) - 2/5 * \log_2(2/5) = 0.97095$$

Podskup skupa S u kojem je vrednost atributa „toplokrvni“ NE je veličine 1 (krokodil). Samo vrednost DA za izlazni atribut „leže jaja“ se u takvom podskupu nalazi. Entropija takvog podskupa je:

$$\textit{Entropy}(1\ \text{DA}, 0\ \text{NE}) = -\ 1/1 * \log_2(1/1) - 0/1 * \log_2(0/1) = 0$$

Sada računamo dobit informacija za podelu po atributu „toplokrvna“:

$$\textit{Gain}(\text{S}, \text{toplokrvna}) = 0.91829 - (5/6 * 0.97095 + 1/6 * 0) = 0.10916$$

Zadatak 1 – Rešenje

Atribut „pernata“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „pernata“ DA je veličine 3 (noj, gavran, albatros). Samo vrednost DA za izlazni atribut „leže jaja“ se u takvom podskupu nalazi. Entropija takvog podskupa je:

$$Entropy(3\ DA, 0\ NE) = -\ 3/3 * \log_2(3/3) - 0/3 * \log_2(0/3) = 0$$

Podskup skupa S u kojem je vrednost atributa „pernata“ NE je veličine 3 (krokodil, delfin, koala). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 1 (krokodil), a vrednosti NE je 2 (delfin, koala). Entropija takvog podskupa je:

$$Entropy(1\ DA, 2\ NE) = -\ 1/3 * \log_2(1/3) - 2/3 * \log_2(2/3) = 0.91829$$

Sada računamo dobit informacija za podelu po atributu „pernata“:

$$Gain(S, pernata) = 0.91829 - (3/6 * 0 + 3/6 * 0.91829) = 0.45914$$

Zadatak 1 – Rešenje

Atribut „ima krzno“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „ime krzno“ DA je veličine 1 (koala). Samo vrednost NE za izlazni atribut „leže jaja“ se u takvom podskupu nalazi. Entropija takvog podskupa je:

$$Entropy(0\ DA, 1\ NE) = -\ 0/1 * \log_2(0/1) - 1/1 * \log_2(1/1) = 0$$

Podskup skupa S u kojem je vrednost atributa „ima krzno“ NE je veličine 5 (noj, krokodil, gavran, albatros, delfin). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 4 (noj, krokodil, gavran, albatros), a vrednosti NE je 1 (delfin). Entropija takvog podskupa je:

$$Entropy(4\ DA, 1\ NE) = -\ 4/5 * \log_2(4/5) - 1/5 * \log_2(1/5) = 0.7219$$

Sada računamo dobit informacija za podelu po atributu „ima krzno“:

$$Gain(S, \text{ima krzno}) = 0.91829 - (1/6 * 0 + 5/6 * 0.7219) = 0.3167$$

Zadatak 1 – Rešenje

Atribut „pliva“ ima dve vrednosti: DA i NE.

Podskup skupa S u kojem je vrednost atributa „pliva“ DA je veličine 2 (krokodil, delfin). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 1 (krokodil), a vrednosti NE je 1 (delfin). Entropija takvog podskupa je:

$$Entropy(1\ DA, 1\ NE) = -1/2 * \log_2(1/2) - 1/2 * \log_2(1/2) = 1$$

Podskup skupa S u kojem je vrednost atributa „pliva“ NE je veličine 4 (noj, gavran, albatros, koala). Učestalost pojavljivanja vrednosti DA za izlazni atribut „leže jaja“ u takvom podskupu je 3 (noj, gavran, albatros), a vrednosti NE je 1 (koala). Entropija takvog podskupa je:

$$Entropy(3\ DA, 1\ NE) = -3/4 * \log_2(3/4) - 1/4 * \log_2(1/4) = 0.81127$$

Sada računamo dobit informacija za podelu po atributu „pliva“:

$$Gain(S, pliva) = 0.91829 - (2/6 * 1 + 4/6 * 0.81127) = 0.04411$$

Zadatak 1 – Rešenje

Ostale dobiti smo izračunali ranije. Tražimo najveću od njih.

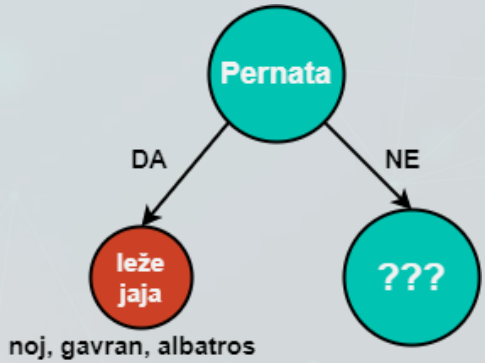
$$Gain(S, \text{toplokrvna}) = 0.91829 - (5/6 * 0.97095 + 1/6 * 0) = 0.10916$$

$$Gain(S, \text{pernata}) = 0.91829 - (3/6 * 0 + 3/6 * 0.91829) = 0.45914$$

$$Gain(S, \text{ima krzno}) = 0.91829 - (1/6 * 0 + 5/6 * 0.7219) = 0.3167$$

$$Gain(S, \text{pliva}) = 0.91829 - (2/6 * 1 + 4/6 * 0.81127) = 0.04411$$

Najveću dobit ima podela po atributu „pernata“ pa taj atribut biramo za koreni čvor i skup podataka delimo na osnovu tog atributa.



Zadatak 1 – Rešenje

Nakon podele ulaznog skupa podataka po atributu „pernata“, dobijamo sledeći podskup.

Životinja	Toplokrvna	Ima krzno	Pliva	Leže jaja
Krokodil	Ne	Ne	Da	Da
Delfin	Da	Ne	Da	Ne
Koala	Da	Da	Ne	Ne

Sada ponavljamo ceo postupak sa novim skupom S’. Računamo entropiju takvog skupa:

$$Entropy(1\ DA, 2\ NE) = -\ 1/3 * \log_2(1/3) - 2/3 * \log_2(2/3) = 0.91829$$

Za atribut „toplokrvna“ računamo:

$$Entropy(0\ DA, 2\ NE) = -\ 0/2 * \log_2(0/2) - 2/2 * \log_2(2/2) = 0$$

$$Entropy(1\ DA, 0\ NE) = -\ 1/1 * \log_2(1/1) - 0/1 * \log_2(0/1) = 0$$

$$Gain(S', toplokrvna) = 0.91829 - (2/3 * 0 + 1/3 * 0) = 0.91829$$

Zadatak 1 – Rešenje

Za atribut „ima krzno“ računamo:

$$Entropy(0\ DA, 1\ NE) = -\ 0/1 * \log_2(0/1) - 1/1 * \log_2(1/1) = 0$$

$$Entropy(1\ DA, 1\ NE) = -\ 1/2 * \log_2(1/2) - 1/2 * \log_2(1/2) = 1$$

$$Gain(S', \text{ima krzno}) = 0.91829 - (1/3 * 0 + 2/3 * 1) = 0.25162$$

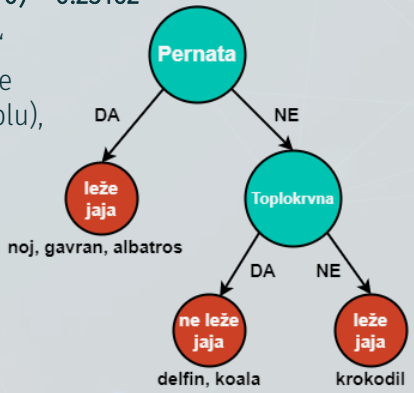
Za atribut „pliva“ računamo:

$$Entropy(1\ DA, 1\ NE) = -\ 1/2 * \log_2(1/2) - 1/2 * \log_2(1/2) = 1$$

$$Entropy(0\ DA, 1\ NE) = -\ 0/1 * \log_2(0/1) - 1/1 * \log_2(1/1) = 0$$

$$Gain(S', \text{pliva}) = 0.91829 - (2/3 * 1 + 1/3 * 0) = 0.25162$$

Najbolju dobit informacija ima podela po atributu „toplokrvna“ pa taj čvor biramo za sledeći čvor. Kako skup podataka ne može dalje da se deli (stigli smo do listova po svim putanjama u stablu), konstrukcija stabla je završena.



Iako je dobijeno isto stablo, u opštem slučaju to ne mora da važi. Gini indeks generalno brže formira stablo zbog lakšeg matematičkog računa, dok dobit informacija daje za nijansu bolje rezultate prilikom testiranja. Češće nam je bitnija brzina treniranja modela, pa je Gini indeks danas popularniji izbor.

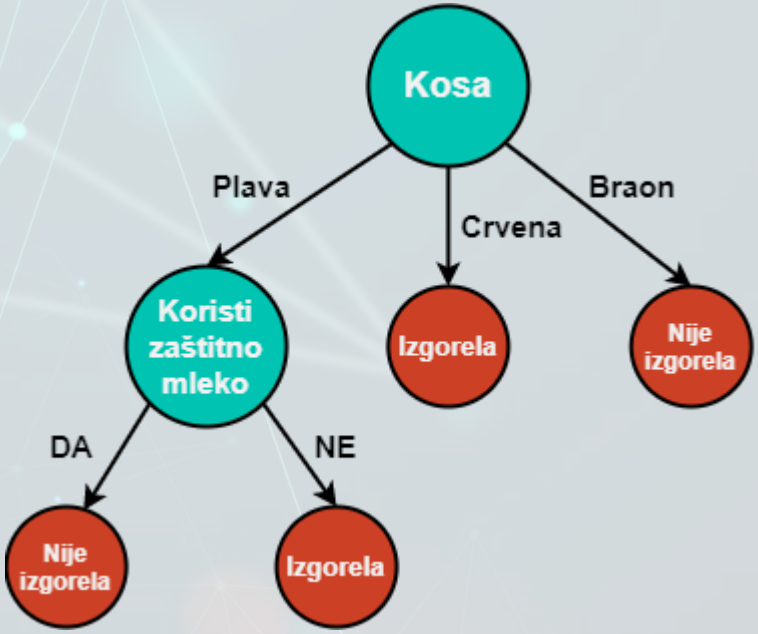
Zadatak za samostalnu vežbu – Sunčanje

★★★☆☆

Konstruisati stablo odlučivanja za određivanje da li će osoba izgoreti u toku sunčanja na osnovu sledećeg skupa podataka i njihovih karakteristika.

Ime	Kosa	Visina	Masa	Koristi zaštitnu kremu	Izgoreo/la
Aleksa	Plava	Prosečna	Laka	Ne	Da
Bojan	Plava	Visoka	Prosečna	Da	Ne
Ceca	Braon	Niska	Prosečna	Da	Ne
Darko	Plava	Niska	Prosečna	Ne	Da
Ema	Crvena	Prosečna	Teška	Ne	Da
Filip	Braon	Visoka	Teška	Ne	Ne
Goran	Braon	Prosečna	Teška	Ne	Ne
Helena	Plava	Niska	Laka	Da	Ne

Zadatak za samostalnu vežbu – Rešenje



Zadatak 2 – Simpsonovi

★★★☆☆

Konstruisati stablo odlučivanja za određivanje da li je osoba muškog ili ženskog pola na osnovu sledećeg skupa podataka i njihovih karakteristika.

Ime	Dužina kose	Masa	Starost	Pol
Homer	0	113	36	M
Mardž	30	68	34	Ž
Bart	5	40	10	M
Lisa	15	35	8	Ž
Megi	10	9	1	Ž
Ejb	3	77	70	M
Selma	20	72	41	Ž
Oto	30	81	38	M
Krasti	15	90	45	M

Zadatak 2 – Simpsonovi

★★★☆☆

Nakon konstrukcije stabla, izvršiti testiranje nad sledećim skupom podataka i izračunati tačnost, preciznost i odziv.

Ime	Dužina kose	Masa	Starost	Pol
Barni	5	120	33	M
Apu	15	74	32	M
Ned	8	77	45	M
Edna	15	74	40	Ž

Zadatak 2 – Rešenje

Skup podataka transformišemo i prilagođavamo radu algoritma.

Ime	Kosa veća od 12	Masa veća od 75	Starost veća od 40	Pol
Homer	NE	DA	NE	M
Mardž	DA	NE	NE	Ž
Bart	NE	NE	NE	M
Lisa	DA	NE	NE	Ž
Megi	NE	NE	NE	Ž
Ejb	NE	DA	DA	M
Selma	DA	NE	DA	Ž
Oto	DA	DA	NE	M
Krasti	DA	DA	DA	M

Zadatak 2 – Rešenje

Kako deliti skup kada je atribut kontinualnog tipa?

Jedna varijanta: naći granične vrednosti na osnovu koje bi se skup podelio na podskupove takve da je dobit informacija za takvu podelu najbolja moguća (problem – koliko „sitno“ tražiti granične vrednosti?).

Npr. Dužina kose

Isprobavamo vrednosti od 0.5 do 29.5. Koji korak uzeti? Ne previše mali, a ne ni previše veliki.

Naći sve granične vrednosti u kojima se menja izlazni atribut „pol“ (problem preobučavanja).

Druga varijanta: naći graničnu vrednost takvu da se podelom skupa dobiju podskupovi sa jednakim (ili skoro jednakim) brojem elemenata (bolje performanse – potencijalno lošija podela).

Npr. za dužinu kose bismo mogli da izaberemo graničnu vrednost 12.

Zadatak 2 – Rešenje

Računamo entropiju početnog skupa:

$$Entropy(4\ Ž, 5\ M) = -\ 4/9 * \log_2(4/9) - 5/9 * \log_2(5/9) = 0.9911$$

Za atribut „dužina kose“ biramo graničnu vrednost 12 i računamo:

$$<= 12 : Entropy(1\ Ž, 3\ M) = -\ 1/4 * \log_2(1/4) - 3/4 * \log_2(3/4) = 0.8113$$

$$> 12 : Entropy(3\ Ž, 2\ M) = -\ 3/5 * \log_2(3/5) - 2/5 * \log_2(2/5) = 0.9710$$

$$Gain(S, \text{dužina kose} <= 12) = 0.9911 - (4/9 * 0.8113 + 5/9 * 0.9710) = 0.0911$$

Da li bi nam izbor neke druge granične vrednosti dao veću dobit?

Za atribut „masa“ biramo graničnu vrednost 75 i računamo:

$$<= 75 : Entropy(4\ Ž, 1\ M) = -\ 4/5 * \log_2(4/5) - 1/5 * \log_2(1/5) = 0.7219$$

$$> 75 : Entropy(0\ Ž, 4\ M) = -\ 0/4 * \log_2(0/4) - 4/4 * \log_2(4/4) = 0$$

$$Gain(S, \text{masa} <= 75) = 0.9911 - (5/9 * 0.7219 + 4/9 * 0) = 0.5900$$

Da li bi nam izbor neke druge granične vrednosti dao veću dobit?

Zadatak 2 – Rešenje

Za atribut „starost“ biramo graničnu vrednost 40 i računamo:

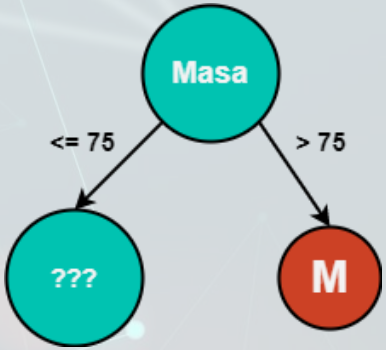
$<= 40$: *Entropy*(3 Ž, 3 M) = $- 3/6 * \log_2(3/6) - 3/6 * \log_2(3/6) = 1$

> 40 : *Entropy*(1 Ž, 2 M) = $- 1/3 * \log_2(1/3) - 2/3 * \log_2(2/3) = 0.9183$

Gain(S, starost <= 40) = $0.9911 - (6/9 * 1 + 3/9 * 0.9183) = 0.0183$

Da li bi nam izbor neke druge granične vrednosti dao veću dobit?

Najveću dobit nam daje podela po atributu „masa“.

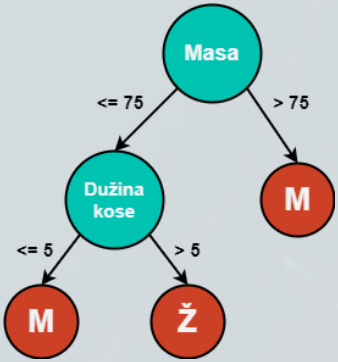


Zadatak 2 – Rešenje

Nastavljajući proceduru, podelu možemo na dalje izvršiti po atributu „dužina kose“ sa granicom 5 čime dolazimo do kraja.

Vrši se testiranje nad sledećim skupom podataka.

Ime	Dužina kose	Masa	Starost	Pol	Predikcija
Barni	5	120	33	M	M
Apu	15	74	32	M	Ž
Ned	8	77	45	M	M
Edna	15	74	40	Ž	Ž



Barni, Ned i Edna su ispravno klasifikovani, dok je Apu pogrešno klasifikovan.

Tačnost modela je stoga 75%.

Preciznost za klasu M je 100%, dok je odziv 67%.

Preciznost za klasu Ž je 50%, dok je odziv 100%.

REGRESIVNA STABLA

Šta je i kako se određuje standardno kvadratno odstupanje?

Ako je dat skup A veličine |A| sa ulaznim vrednostima x, tada je srednja vrednost skupa A

$$x_{sr} = \frac{\sum x}{|A|}$$

Standardno kvadratno odstupanje S(A), skupa A je:

$$S(A) = \sqrt{\frac{\sum(x - x_{sr})^2}{|A|}}$$

Koeficijent varijacije CV(A), skupa A je:

$$CV(A) = \frac{S(A)}{X_{sr}} * 100\%$$

REGRESIVNA STABLA

Šta je i kako se određuje standardno kvadratno odstupanje?

Standardno odstupanje se koristi pri izboru atributa za podelu.

Kako regresivna stabla imaju kontinualnu izlaznu vrednost, potrebna je vrednost na osnovu koje se određuje kriterijum zaustavljanja konstrukcije stabla. U tu svrhu se koristi koeficijent varijacije, veličina skupa ili oba.

Srednja vrednost se koristi za vrednost listova.

Kombinovano standardno kvadratno odstupanje za dva atributa (ulazni X i izlazni T) koje se koristi kao entropija kod klasifikacionih stabala se računa na sledeći način:

$$S(T, X) = \sum_{c \in X} P(c) S(c)$$

gde je *c* jedna od vrednosti atributa X, *P(c)* učestalost pojavljivanja vrednosti *c* u skupu, a *S(c)* standardno kvadratno odstupanje podskupa glavnog skupa dobijenog uzimanjem onih redova čije su vrednosti atributa X jednake *c*.

REGRESIVNA STABLA

Šta je i kako se određuje redukcija standardnog odstupanja?

Redukcija standardnog odstupanja se dobija kao razlika standardnih odstupanja pre i posle podele po nekom atributu i odgovara očekivanoj dobiti kod klasifikacionih stabala.

Atribut koji se koristi za podelu se bira kao onaj sa maksimalnom redukcijom.

Proces se iterativno ponavlja dok nije ispunjen kriterijum zaustavljanja. Uglavnom je kriterijum zaustavljanja maksimalan koeficijent varijacije koji podskup može da ima ili maksimalan broj instanci u podskupu.

Zadatak 3 – Igranje golfa

★★★★☆

Konstruisati stablo odlučivanja za određivanje koliko vremena će društvo igrati golf na osnovu sledećeg skupa podataka i njihovih karakteristika. Kriterijumi zaustavljanja su 10% za koeficijent varijacije i 3 za veličinu skupa.

Vreme	Temperatura	Vlažnost	Vetar	Trajanje igre
Kiša	Vruće	Visoka	Ne	25
Kiša	Vruće	Visoka	Da	30
Oblačno	Vruće	Visoka	Ne	46
Sunčano	Blago	Visoka	Ne	45
Sunčano	Hladno	Normalna	Ne	52
Sunčano	Hladno	Normalna	Da	23
Oblačno	Hladno	Normalna	Da	43
Kiša	Blago	Visoka	Ne	35
Kiša	Hladno	Normalna	Ne	38
Sunčano	Blago	Normalna	Ne	46
Kiša	Blago	Normalna	Da	48
Oblačno	Blago	Visoka	Da	52
Oblačno	Vruće	Normalna	Ne	44
Sunčano	Blago	Visoka	Da	30

Zadatak 3 – Rešenje

Veličina skupa |trajanje igre| = 14

Srednja vrednost izlaza x_{sr} = 39.8

Standardno odstupanje izlaza S(trajanje igre) = 9.32

Koeficijent varijacije izlaza CF(trajanje igre) = 23%

I veličina skupa i koeficijent varijacije su veći od kriterijuma zaustavljanja te se algoritam nastavlja.

Sada je potrebno pronaći standardno odstupanje za sve podskupove dobijene podelom svakog od atributa: vreme, temperatura, vlažnost, vetar. Atribut sa najvećom redukcijom će biti atribut na osnovu kojeg vršimo prvu podelu ulaznog skupa.

Trajanje igre
25
30
46
45
52
23
43
35
38
46
48
52
44
30

Zadatak 3 – Rešenje

Atribut „vreme“ ima tri vrednosti: Oblačno, Kiša i Sunčano.

Podskup skupa S u kojem je vrednost atributa „vreme“ Oblačno je veličine 4.
Srednja vrednost izlazne vrednosti takvog podskupa je 46.25.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 3.49.

Podskup skupa S u kojem je vrednost atributa „vreme“ Kiša je veličine 5.
Srednja vrednost izlazne vrednosti takvog podskupa je 35.2.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 7.78.

Podskup skupa S u kojem je vrednost atributa „vreme“ Sunčano je veličine 5.
Srednja vrednost izlazne vrednosti takvog podskupa je 39.2.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 10.87.

Kombinovano standardno odstupanje za atribute „trajanje igre“ i „vreme“ je:

$$S(\text{trajanje igre, vreme}) = 4/14 * 3.49 + 5/14 * 7.78 + 5/14 * 10.87 = 7.66$$

Redukcija standardnog odstupanja za ovakvu podelu je:

$$SDR(\text{trajanje igre, vreme}) = S(\text{trajanje igre}) - S(\text{trajanje igre, vreme})$$

$$SDR(\text{trajanje igre, vreme}) = 9.32 - 7.66 = 1.66$$

Vreme	Trajanje igre
Oblačno	46
Oblačno	43
Oblačno	52
Oblačno	44

Vreme	Trajanje igre
Kiša	25
Kiša	30
Kiša	35
Kiša	38
Kiša	48

Vreme	Trajanje igre
Sunčano	45
Sunčano	52
Sunčano	23
Sunčano	46
Sunčano	30

Zadatak 3 – Rešenje

Atribut „temperatura“ ima tri vrednosti: Hladno, Vruće, Blago.

Podskup skupa S u kojem je vrednost atributa „temperatura“ Hladno je veličine 4.
Srednja vrednost izlazne vrednosti takvog podskupa je 39.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 10.51.

Podskup skupa S u kojem je vrednost atributa „temperatura“ Vruće je veličine 4.
Srednja vrednost izlazne vrednosti takvog podskupa je 36.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 8.95.

Podskup skupa S u kojem je vrednost atributa „temperatura“ Blago je veličine 6.
Srednja vrednost izlazne vrednosti takvog podskupa je 42.67.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 7.65.

Kombinovano standardno odstupanje za atribute „trajanje igre“ i „temperatura“ je:

$$S(\text{trajanje igre, temperatura}) = 4/14 * 10.51 + 4/14 * 8.95 + 6/14 * 7.65 = 8.83$$

Redukcija standardnog odstupanja za ovakvu podelu je:

$$SDR(\text{trajanje igre, temperatura}) = S(\text{trajanje igre}) - S(\text{trajanje igre, temperatura})$$

$$SDR(\text{trajanje igre, temperatura}) = 9.32 - 8.83 = 0.49$$

Temperatura	Trajanje igre
Hladno	52
Hladno	23
Hladno	43
Hladno	38

Temperatura	Trajanje igre
Vruće	25
Vruće	30
Vruće	46
Vruće	44

Temperatura	Trajanje igre
Blago	45
Blago	35
Blago	46
Blago	48
Blago	52
Blago	30

Zadatak 3 – Rešenje

Atribut „vlažnost“ ima dve vrednosti: Visoka, Normalna.

Podskup skupa S u kojem je vrednost atributa „vlažnost“ Visoka je veličine 7.
Srednja vrednost izlazne vrednosti takvog podskupa je 37.57.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 9.36.

Podskup skupa S u kojem je vrednost atributa „vlažnost“ Normalna je veličine 7.
Srednja vrednost izlazne vrednosti takvog podskupa je 42.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 8.73.

Kombinovano standardno odstupanje za atribute „trajanje igre“ i „vlažnost“ je:

S(trajanje igre, vlažnost) = 7/14 * 9.36 + 7/14 * 8.73 = 9.05

Redukcija standardnog odstupanja za ovakvu podelu je:

SDR(trajanje igre, vlažnost) = S(trajanje igre) – S(trajanje igre, vlažnost)

SDR(trajanje igre, vlažnost) = 9.32 – 9.04 = 0.28

Vlažnost	Trajanje igre
Visoka	25
Visoka	30
Visoka	46
Visoka	45
Visoka	35
Visoka	52
Visoka	30

Vlažnost	Trajanje igre
Normalna	52
Normalna	23
Normalna	43
Normalna	38
Normalna	46
Normalna	48
Normalna	44

Zadatak 3 – Rešenje

Atribut „vetar“ ima dve vrednosti: Da, Ne.

Podskup skupa S u kojem je vrednost atributa „vetar“ Da je veličine 6.
Srednja vrednost izlazne vrednosti takvog podskupa je 37.67.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 10.59.

Podskup skupa S u kojem je vrednost atributa „vetar“ Ne je veličine 8.
Srednja vrednost izlazne vrednosti takvog podskupa je 41.375.
Standardno odstupanje izlazne vrednosti u takvom podskupu je 7.87.

Kombinovano standardno odstupanje za atribute „trajanje igre“ i „vetar“ je:

S(trajanje igre, vetar) = 6/14 * 10.59 + 8/14 * 7.87 = 9.04

Redukcija standardnog odstupanja za ovakvu podelu je:

SDR(trajanje igre, vetar) = S(trajanje igre) – S(trajanje igre, vetar)

SDR(trajanje igre, vetar) = 9.32 – 9.04 = 0.28

SDR(trajanje igre, vlažnost) = 0.28

SDR(trajanje igre, temperatura) = 0.49

SDR(trajanje igre, vreme) = 1.66

Najveći SDR ima atribut „vreme“ pa će nad njim da se vrši podela.

Vetar	Trajanje igre
Da	30
Da	23
Da	43
Da	48
Da	52
Da	30

Vetar	Trajanje igre
Ne	25
Ne	46
Ne	45
Ne	52
Ne	35
Ne	38
Ne	46
Ne	44

Zadatak 3 – Rešenje

Vreme	Temperatura	Vlažnost	Vetar	Trajanje igre
Kiša	Vruće	Visoka	Ne	25
Kiša	Vruće	Visoka	Da	30
Kiša	Blago	Visoka	Ne	35
Kiša	Hladno	Normalna	Ne	38
Kiša	Blago	Normalna	Da	48

Veličina skupa $|\text{trajanje igre}| = 5$

Srednja vrednost izlaza $x_{sr} = 35.2$

Standardno odstupanje izlaza $S(\text{trajanje igre}) = 7.78$

Koeficijent varijacije izlaza $CF(\text{trajanje igre}) = 22.10\%$

I veličina skupa i koeficijent varijacije su veći od kriterijuma zaustavljanja te se algoritam nastavlja po ovom podskupu.

Zadatak 3 – Rešenje

Vreme	Temperatura	Vlažnost	Vetar	Trajanje igre
Oblačno	Vruće	Visoka	Ne	46
Oblačno	Hladno	Normalna	Da	43
Oblačno	Blago	Visoka	Da	52
Oblačno	Vruće	Normalna	Ne	44

Veličina skupa **|trajanje igre| = 4**

Srednja vrednost izlaza X_{sr} = 46.25

Standardno odstupanje izlaza **S(trajanje igre) = 3.49**

Koeficijent varijacije izlaza **CF(trajanje igre) = 7.54%**

Koeficijent varijacije je manji od kriterijuma zaustavljanja te se algoritam ne nastavlja po ovom podskupu. Vrednost lista odgovara srednjoj vrednosti izlaza 46.25.

Zadatak 3 – Rešenje

Vreme	Temperatura	Vlažnost	Vetar	Trajanje igre
Sunčano	Blago	Visoka	Ne	45
Sunčano	Hladno	Normalna	Ne	52
Sunčano	Hladno	Normalna	Da	23
Sunčano	Blago	Normalna	Ne	46
Sunčano	Blago	Visoka	Da	30

Veličina skupa $|\text{trajanje igre}| = 5$

Srednja vrednost izlaza $x_{sr} = 39.2$

Standardno odstupanje izlaza $S(\text{trajanje igre}) = 10.87$

Koeficijent varijacije izlaza $CF(\text{trajanje igre}) = 27.73\%$

I veličina skupa i koeficijent varijacije su veći od kriterijuma zaustavljanja te se algoritam nastavlja po ovom podskupu.

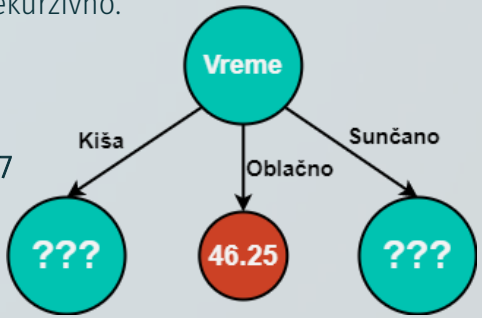
Zadatak 3 – Rešenje

Algoritam se za preostala dva podskupa nastavlja rekurzivno.

Levi podskup (Vreme == Sunčano):

Veličina skupa |trajanje igre| = 5

Standardno odstupanje izlaza S(trajanje igre) = 10.87



SDR(trajanje igre, temperatura) = 0.07

SDR(trajanje igre, vlažnost) = 0.37

SDR(trajanje igre, vetar) = 7.62

Za podelu se dalje bira podela po vetru. Svi podskupovi imaju veličinu manju od 3 te je podela u tom delu završena. Listovi dobijaju vrednosti koje odgovaraju srednjim vrednostima odgovarajućih podskupova.

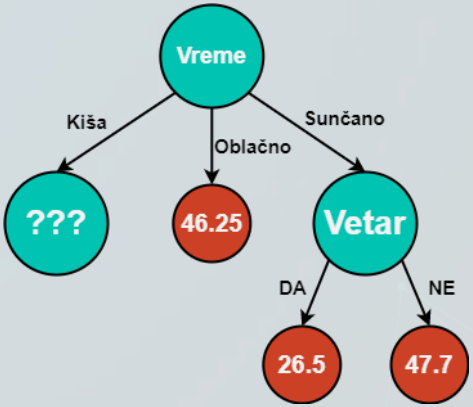
Zadatak 3 – Rešenje

Algoritam se za preostala dva podskupa nastavlja rekurzivno.

Levi podskup (Vreme == Kiša):

Veličina skupa **|trajanje igre| = 5**

Standardno odstupanje izlaza **S(trajanje igre) = 7.78**



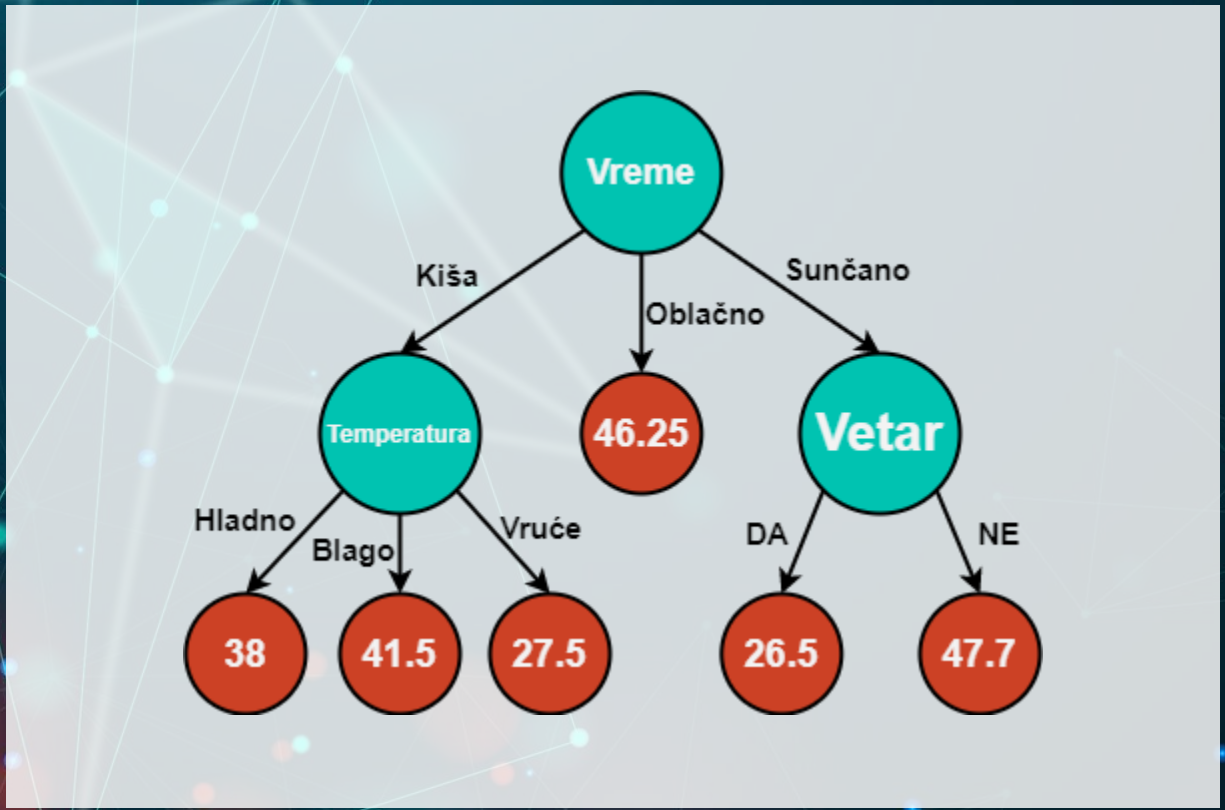
SDR(trajanje igre, temperatura) = **4.18**

SDR(trajanje igre, vlažnost) = 3.32

SDR(trajanje igre, vetar) = 0.82

Za podelu se dalje bira podela po temperaturi. Svi podskupovi imaju veličinu manju od 3 te je podela u tom delu završena. Listovi dobijaju vrednosti koje odgovaraju srednjim vrednostima odgovarajućih podskupova.

Zadatak 3 – Rešenje



Zadatak za samostalnu vežbu – Nekretnine

★★★★☆

Konstruisati stablo odlučivanja za određivanje cene stanova u Beogradu na osnovu sledećeg skupa podataka i njihovih karakteristika. Kriterijumi zaustavljanja su 10% za koeficijent varijacije i 2 za veličinu skupa.

Kvadratura	Broj soba	Opština	Cena
110	3	Palilula	210.000
73	3	Novi Beograd	129.000
65	2	Novi Beograd	171.000
74	3.5	Palilula	159.000
70	2.5	Novi Beograd	139.500
30	1	Palilula	35.000
43	1.5	Palilula	105.000
110	4	Stari grad	219.000
52	2	Stari grad	160.000

NASUMIČNA ŠUMA (RANDOM FOREST)

Iako su brza za komputaciju, Stabla odlučivanja daju generalno lošije rezultate i podložna su problemu preobučavanja (engl. *overfitting* - nepoželjno svojstvo koje se javlja kada model mašinskog učenja daje tačna predviđanja za trening podatke, ali ne i za nove podatke).

Nasumična šuma je algoritam mašinskog učenja koji kombinuje rezultate nekoliko stabala odlučivanja kako bi donela konačan rezultat, povećavajući vreme komputacije, ali smanjujući šanse za preobučavanjem.

Primer: Marko nije siguran koji fakultet da upiše pa se savetuje sa roditeljima, bratom, sestrom i prijateljima. Marko im je izložio prednosti i mane oba fakulteta koja su mu u užem izboru i na osnovu toga svako mu je dao savet šta da upiše. Na osnovu svih saveta, Marko donosi svoju odluku. Ovakav način donošenja odluke daje bolje rezultate od donošenja odluke samo na osnovu jednog saveta (npr. savet brata).

NASUMIČNA ŠUMA (RANDOM FOREST)

Postoje dva načina agregiranja više odluka u jednu:

- *Bagging* – Iz početnog skupa podataka se izvlače podskupovi (koji smeju da imaju deljene primerke ili primerke koji se ponavljaju) na osnovu kojih se treniraju modeli paralelno. Svaki model se koristi za predikciju i daje svoj izlaz, a konačan izlaz se dobija kao najzastupljeniji izlaz/klasa (klasifikacija) ili kao prosečna vrednost svih izlaza (regresija).

- *Boosting* – Modeli se ne kreiraju paralelno već sekvencijalno, tako što se kreira prvi model (uglavnom nasumičan) nad kojim se vrši predikcija. Iz skupa podataka se izdvajaju svi primerci koji su pogrešno prediktovani prvim modelom i koriste se zajedno sa novim podskupom podataka da se istrenira drugi model, itd. Nakon određenog vremena, poslednji model će biti najnapredniji, jer je koristio informacije o greškama svih prethodnih modela.

Nasumična šuma koristi *bagging* tehniku pri čemu za treniranje svakog podstabla ne koristi nužno isti skup ulaznih atributa, već se i za njih bira nasumični podskup (Marko nije svima dao iste informacije o fakultetima).

Zadatak 4 – Životinje koje ležu jaja

★★★☆☆

Konstruisati nasumičnu šumu sa tri stabla za određivanje da li životinja leže jaja na osnovu sledećeg skupa podataka i njihovih karakteristika. Nasumičnim odabirom skup je podeljen na tri podskupa S1(noj, krokodil, gavran, delfin), S2(noj, albatros, delfin, koala) i S3(noj, gavran, albatros, delfin). Stabla koriste atribute A1(toplokrvna, pernata, ima krzno), A2(pernata, ima krzno, pliva) i A3(toplokrvna, ima krzno, pliva) i Gini podelu.

Nezavisni atributi					**Atribut odluke**
Životinja	Toplokrvna	Pernata	Ima krzno	Pliva	Leže jaja
Noj	Da	Da	Ne	Ne	Da
Krokodil	Ne	Ne	Ne	Da	Da
Gavran	Da	Da	Ne	Ne	Da
Albatros	Da	Da	Ne	Ne	Da
Delfin	Da	Ne	Ne	Da	Ne
Koala	Da	Ne	Da	Ne	Ne

Zadatak 4 – Rešenje

S1

Životinja	Toplokrvna	Pernata	Ima krzno	Leže jaja
Noj	Da	Da	Ne	Da
Krokodil	Ne	Ne	Ne	Da
Gavran	Da	Da	Ne	Da
Delfin	Da	Ne	Ne	Ne

$Gini(toplokrvna=DA) = 1 - (2/3)^2 - (1/3)^2 = 0.44$

$Gini(toplokrvna=NE) = 1 - (1/1)^2 - (0/1)^2 = 0$

$Gini(S1, toplokrvna) = 3/4 * 0.44 + 1/4 * 0 = 0.33$

$Gini(pernata=DA) = 1 - (2/2)^2 - (0/2)^2 = 0$

$Gini(pernata=NE) = 1 - (1/2)^2 - (1/2)^2 = 0.5$

$Gini(S1, pernata) = 2/4 * 0 + 2/4 * 0.5 = 0.25$

Po krznu nije moguće podeliti ovaj skup podataka.

Podela se vrši po atributu „pernata“, a potom po preostalom atributu „toplokrvna“.

Zadatak 4 – Rešenje

S2

Životinja	Pernata	Ima krzno	Pliva	Leže jaja
Noj	Da	Ne	Ne	Da
Albatros	Da	Ne	Ne	Da
Delfin	Ne	Ne	Da	Ne
Koala	Ne	Da	Ne	Ne

$Gini(pernata=DA) = 1 - (2/2)^2 - (0/2)^2 = 0$

$Gini(pernata=NE) = 1 - (0/2)^2 - (2/2)^2 = 0$

$Gini(S1, pernata) = 2/4 * 0 + 2/4 * 0 = 0$

$Gini(krzno=DA) = 1 - (0/1)^2 - (1/1)^2 = 0$

$Gini(krzno=NE) = 1 - (2/3)^2 - (1/3)^2 = 0.44$

$Gini(S1, krzno) = 1/4 * 0 + 3/4 * 0.44 = 0.33$

$Gini(pliva=DA) = 1 - (0/1)^2 - (1/1)^2 = 0$

$Gini(pliva=NE) = 1 - (2/3)^2 - (1/3)^2 = 0.44$

$Gini(S1, pliva) = 1/4 * 0 + 3/4 * 0.44 = 0.33$

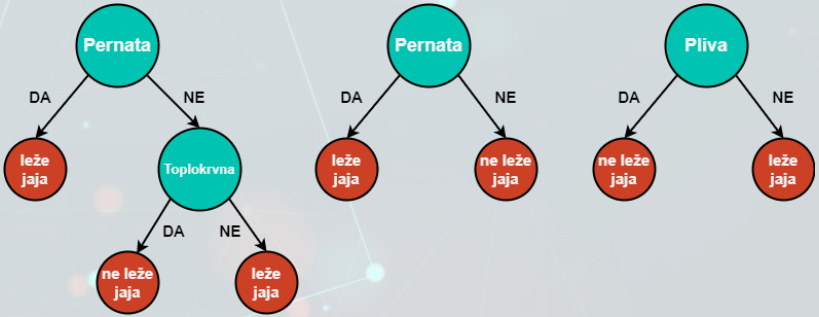
Podela se vrši po atributu „pernata“ čime dobijamo čiste skupove.

Zadatak 4 – Rešenje

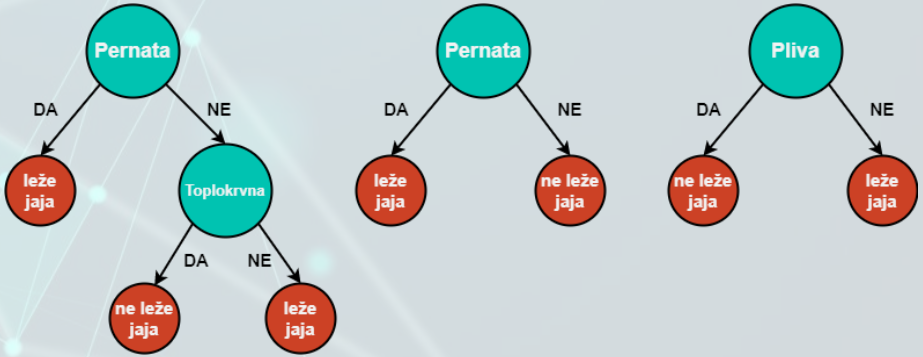
S3

Životinja	Toplokrvna	Ima krzno	Pliva	Leže jaja
Noj	Da	Ne	Ne	Da
Gavran	Da	Ne	Ne	Da
Albatros	Da	Ne	Ne	Da
Delfin	Da	Ne	Da	Ne

Podela se vrši po atributu „pliva“, jer je to jedini atribut koji može da napravi podelu. Takođe, ovakvom podelom dobijamo čiste skupove. Time dobijamo sledeću nasumičnu šumu.



Zadatak 4 – Rešenje



Konstruisana je sledeća nasumična šuma. Sada možemo da izvršimo predikciju:

Životinja	Toplokrvna	Pernata	Ima krzno	Pliva	Leže jaja		
					S1	S2	S3
Čekić ajkula	Ne	Ne	Ne	Da	Da	Ne	Ne
Gušter	Ne	Ne	Ne	Ne	Da	Ne	Da

Da li bismo iste rezultate dobili koristeći stablo iz Zadatka 1?

PITANJA?

http://ri4es.etf.rs/