

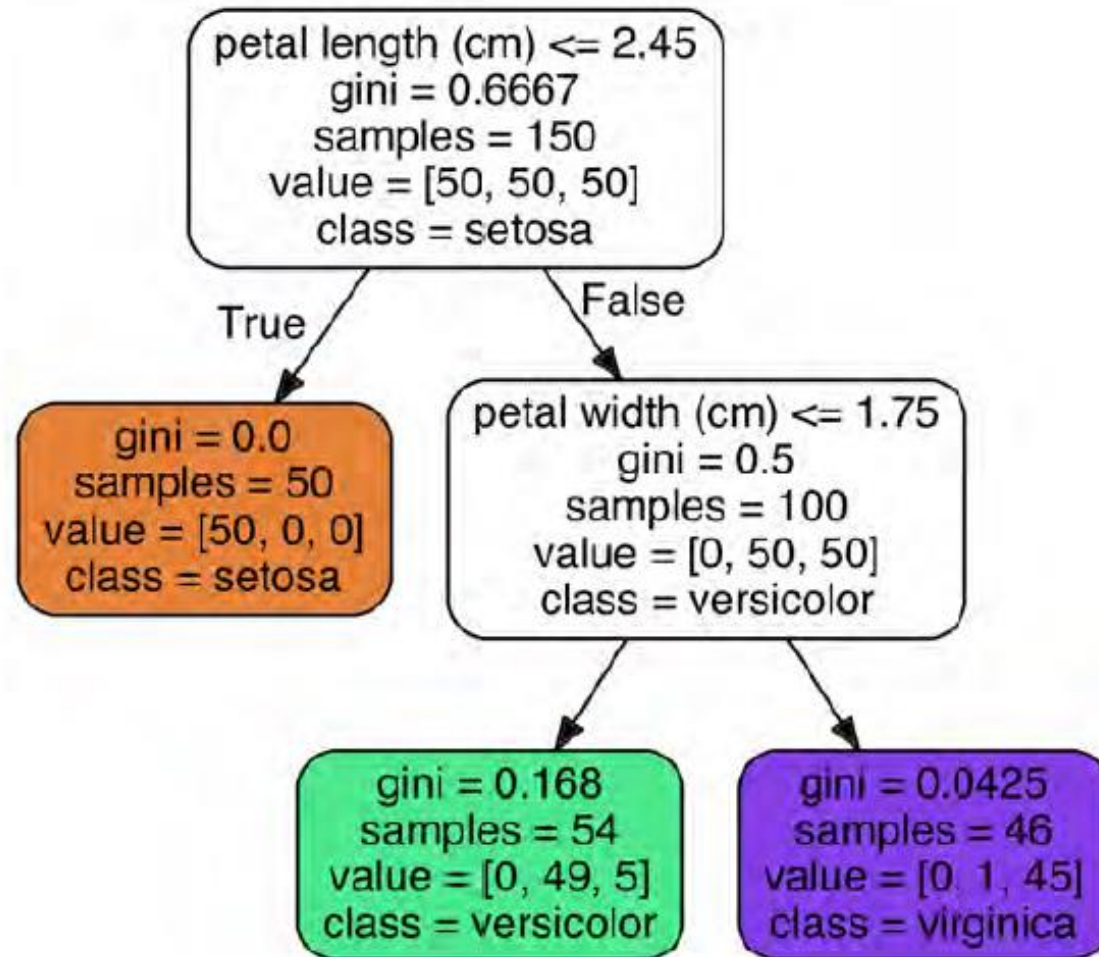
Stabla odlučivanja

Iris data set



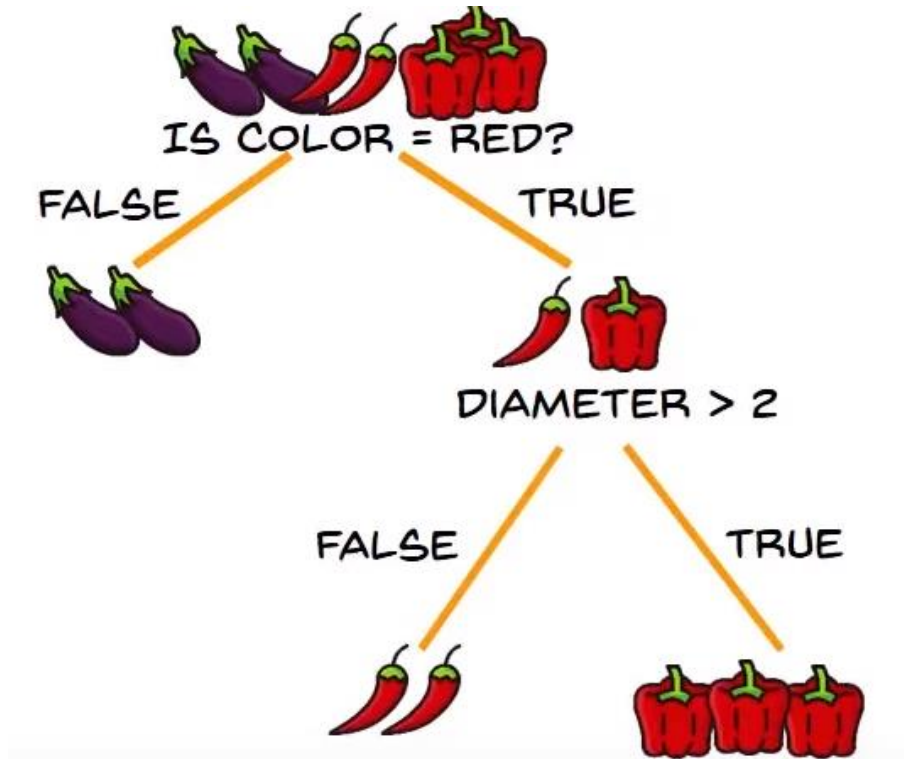
- Skup uključuje tri različite vrste irisa
- *Iris setosa*
- *Iris versicolor*
- *Iris virginica*
- U skupu je dato po 50 primera od svake vrste irisa
- Iri su opisani korišćenjem 4 svojstva (dužine i širine čašičnih listića (sepal) i latica (petal))

Stablo odlučivanja za Iris skup podataka



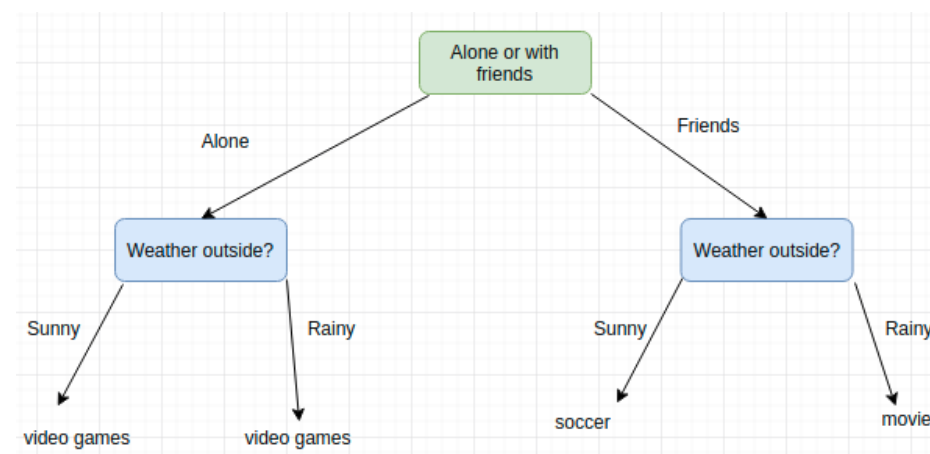
Šta je stablo odlučivanja?

- Stablo odlučivanja je dijagram oblika stabla koji se koristi za utvrđivanje toka akcija u procesu odlučivanja.
- Svaka grana predstavlja moguće odluke, pojave ili reakcije.



Zašto stablo odlučivanja?

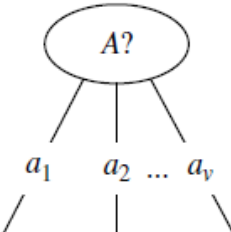
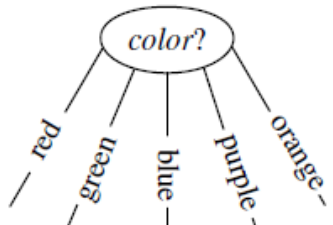
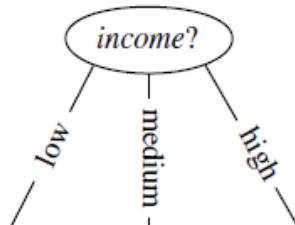
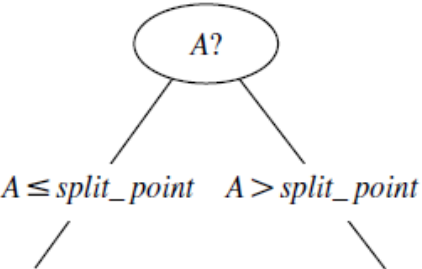
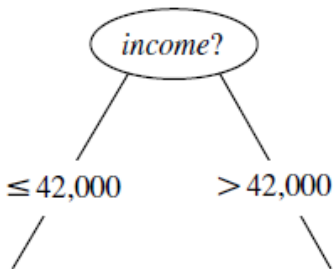
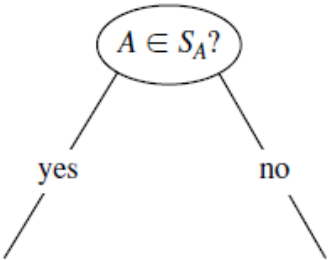
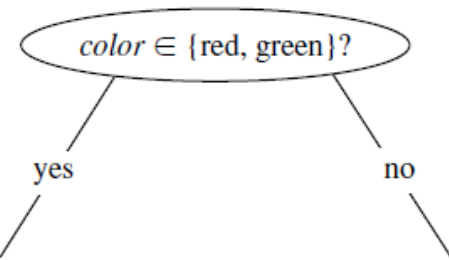
- Stabla odlučivanja su najjednostavniji mehanizam za predikciju.
- Na osnovu stabla odlučivanja mogu se generisati pravila, koja ljudi mogu da razumeju i koja mogu biti upotrebljena za formiranje baze znanja.



Stablo odlučivanja

- U svakom čvoru odluke (Decision node) testira po jedan atribut.
- Čvor list (Leaf node): određuje vrednost ciljne funkcije (rezultat klasifikacije).
- Svaka grana (branch) koja polazi iz nekog čvora odgovara jednoj od mogućih vrednosti atributa koji je testiran u tom čvoru.
- Putanja odlučivanja: niz čvorova i grana od korenog čvora do jednog od listova koji predstavlja konačnu odluku.

Različiti tipovi atributa

	Partitioning scenarios	Examples
diskretni		 
kontinualni		
binarni		

Obučavanje stabla odlučivanja

- Cilj: Formirati najmanje moguće stablo odlučivanja koje je u saglasnosti sa trening primerima.
- Ideja: (rekurzivno) odabrati “najznačajniji” atribut kao koren (pod)stabla

Koji atribut je najbolji klasifikator?

- Treba izabrati najvažniji atribut - onaj koji je najkorisniji za klasifikaciju primera
- Najkorisniji atribut je onaj koji vrši podelu primera na čiste podskupove u kojima su samo pozitivni ili samo negativni primeri.
- Odabirom najkorisnijeg atributa se praktično kreira koreni čvor stabla.
- Primere dalje smeštamo u podstabla na osnovu njihovih vrednosti najvažnijeg atributa.
- Novi čvorovi stabla dalje se generišu rekurzivno na osnovu sledećeg „najboljeg atributa“.
- Stabla koja se formiraju na ovaj način nazivaju se Top Down Induction of Decision Trees (TDIDT).
Konkretni algoritmi za formiranje stabala: ID3, CART, Assistant, C4.5, C5, ...

Primer – otkrij ko je vampir

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

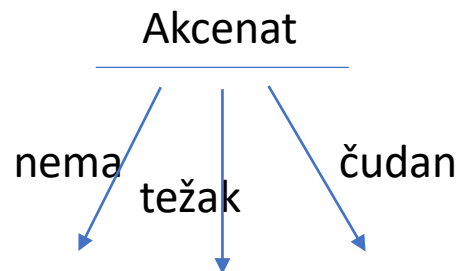
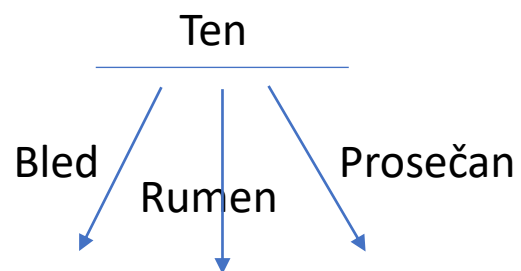
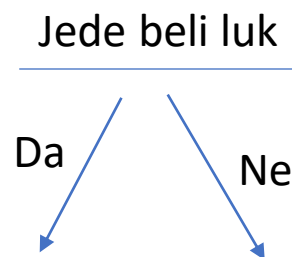
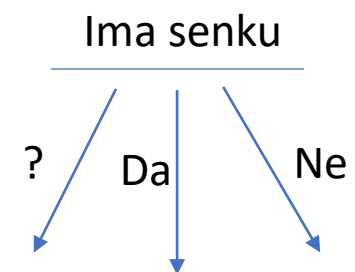
Na osnovu podataka iz tabele treba da napravimo mehanizam koji nam omogućava da zaključimo da li je neko vampir ili ne.

Primer – otkrij ko je vampir

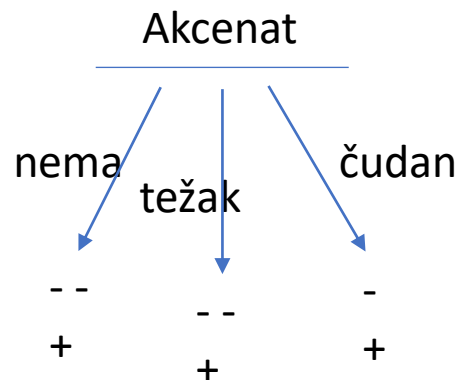
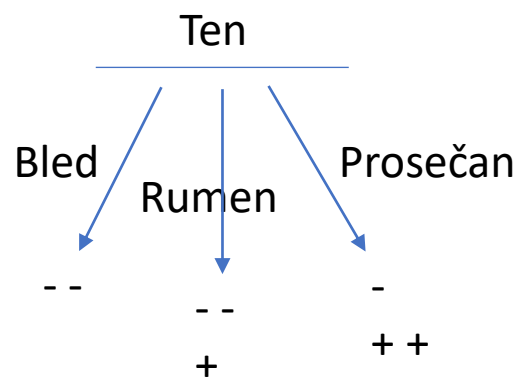
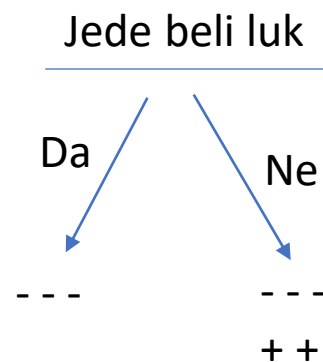
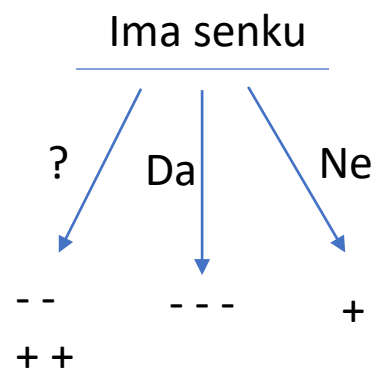
	Ima senku	Jede beli luk	Ten	Akcent	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

Da možete postaviti samo jedno pitanje, koje bi to pitanje bilo?

Odgovor na koje pitanje bi za većinu primera iz tabele odmah uputio na to da li su vampiri ili nisu?

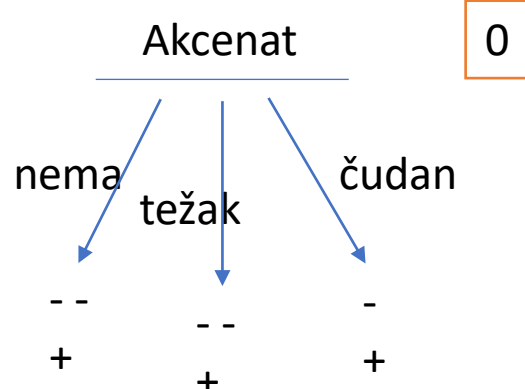
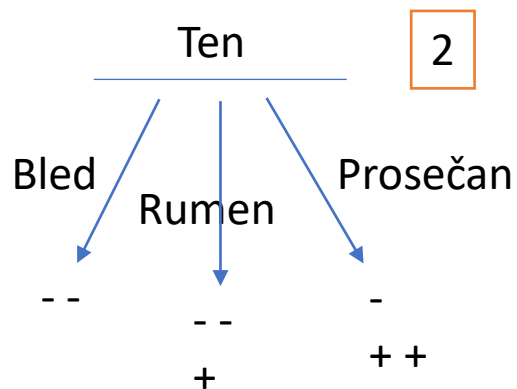
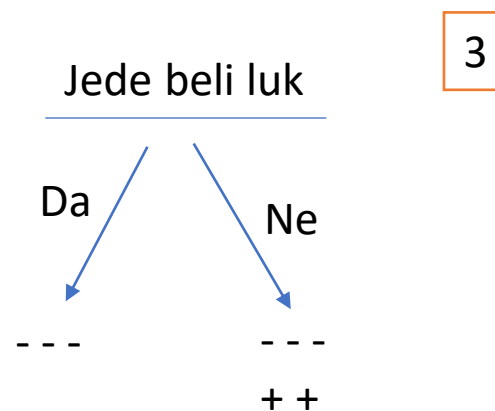
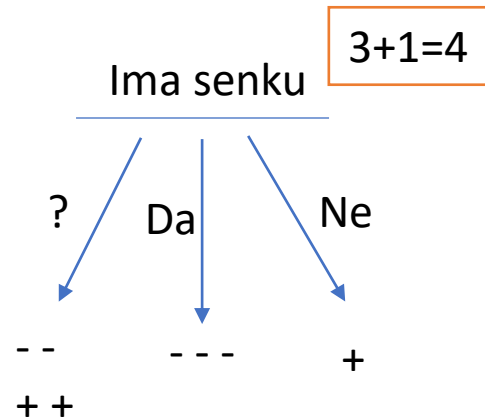


	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

Test je bolji ako se njime što više primera svrstava u homogene grupe

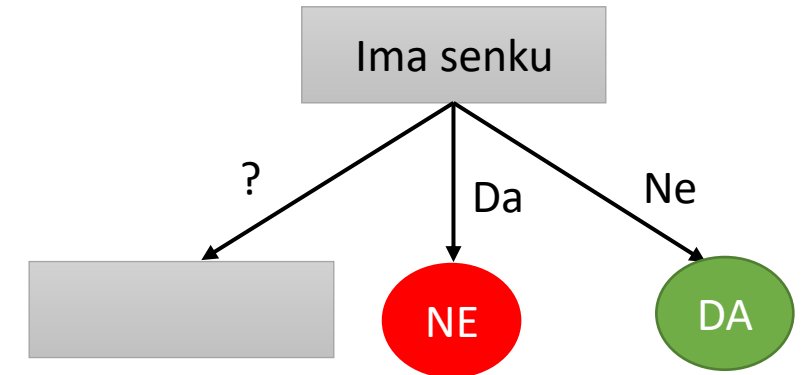


	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

- Test je bolji ako se njime što više primera svrstava u homogene grupe.
- Testiranjem atributa ima senku 4 od osam primera iz skupa za učenje bivaju razvrstani u odgovarajuću klasu.
- Ovaj test ćemo staviti u koren stabla.

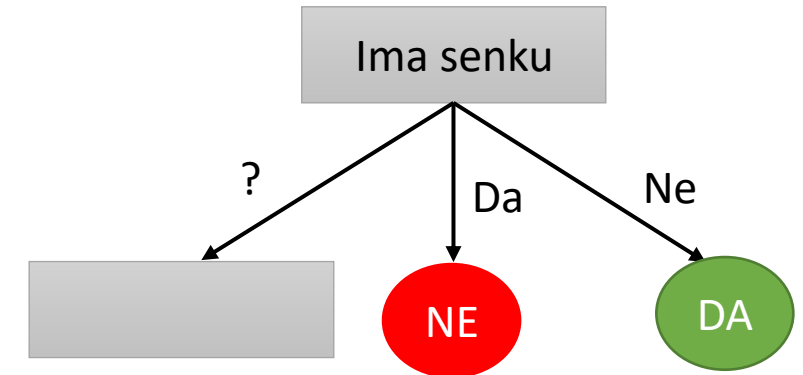
Primer – otkrij ko je vampir

	Ima senku	Jede beli luk	Ten	Akcent	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



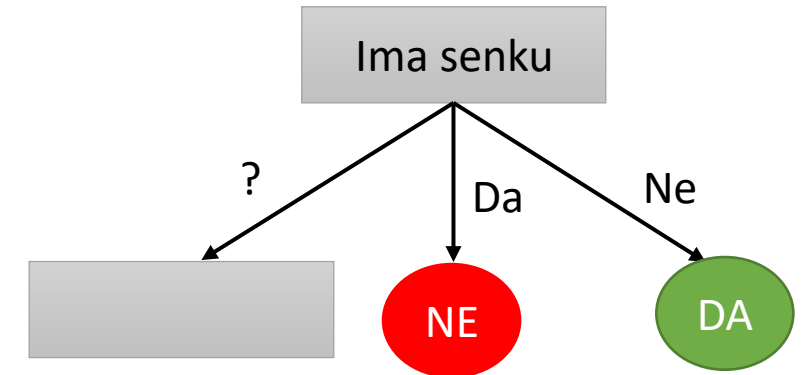
Primer – otkrij ko je vampir

	Ima senku	Jede beli luk	Ten	Akcent	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

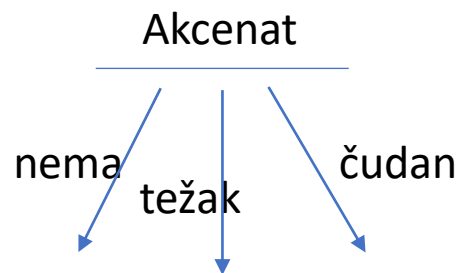
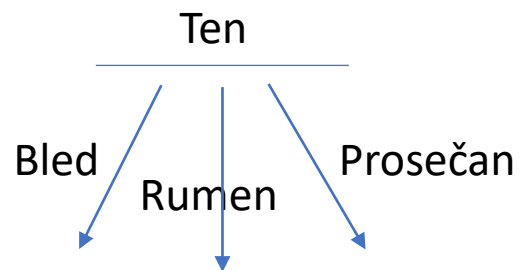
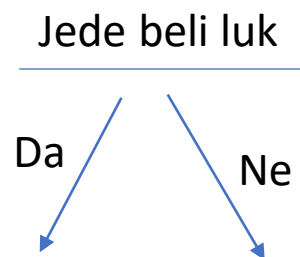


Primer – otkrij ko je vampir

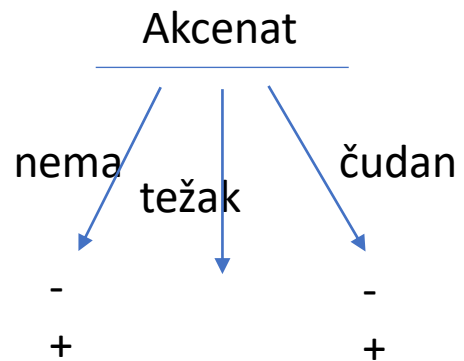
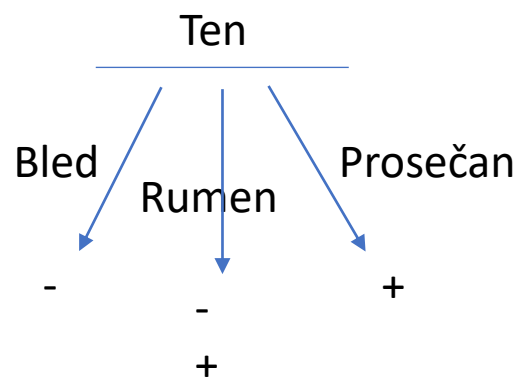
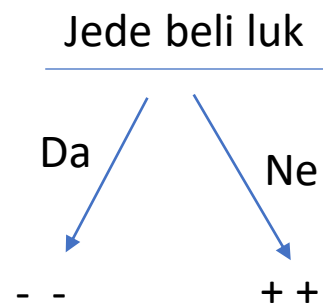
	Ima senku	Jede beli luk	Ten	Akcentat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



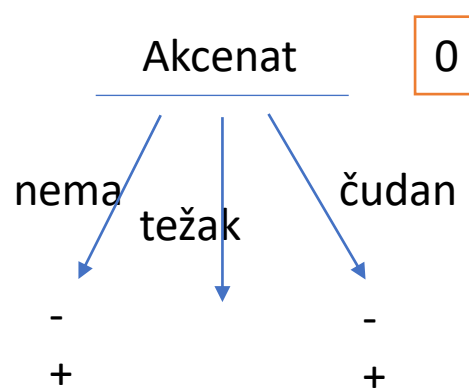
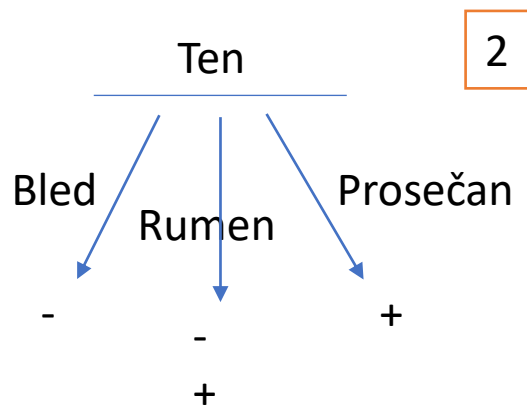
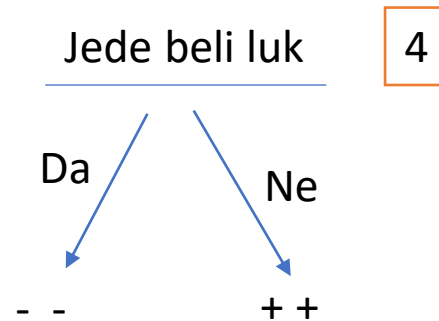
Koje bi pitanje trebalo postaviti sada?



	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

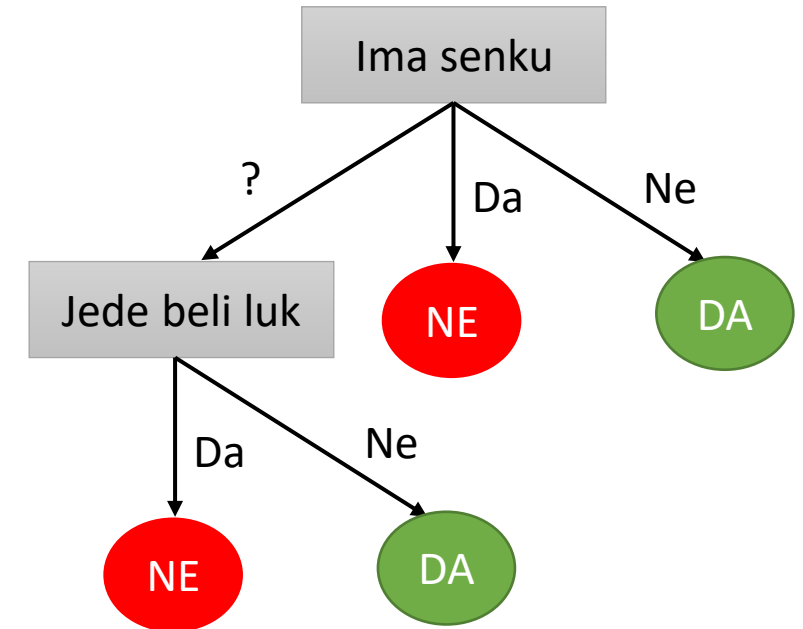


	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

- Sledeći najbolji atribut za klasifikaciju preostalih primera je „Jede beli luk“.
- Njega stavljamo u koren podstabla.

Primer – otkrij ko je vampir

	Ima senku	Jede beli luk	Ten	Akcent	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

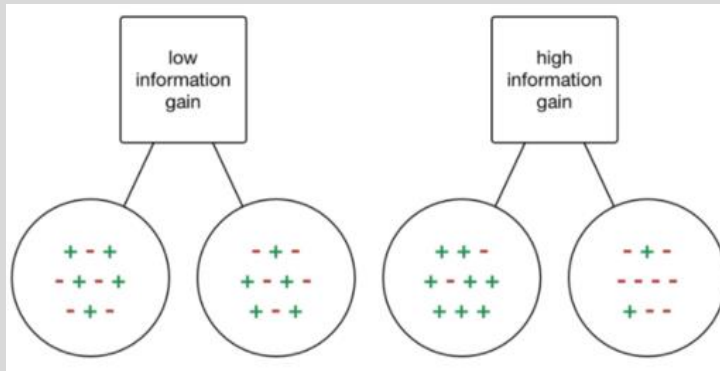


Mere za odabir najboljeg atributa

- Različiti algoritmi za formiranje stabala za odlučivanje koriste različite mere za odabir najboljeg atributa.
- **Information gain** (algoritmi ID3, C4.5 i C5.0)
- **GINI impurity** (algoritam CART (Classification and Regression Trees))

Koji atribut je najbolji klasifikator?

- Selekcija trenutno najboljeg atributa u svakom koraku konstruisanja stabla odluke može se vršiti korišćenjem mere koja se naziva *Informaciono poboljšanje* (*Information gain*)



- Meri koliko dobro dati atribut razdvaja trening primere prema ciljnoj klasifikaciji.
- Ova mera se koristi kako bismo u svakom koraku izgradnje stabla, između različitih atributa odabrali onaj koji će biti naredni čvor odluke.

Entropija

- Da bismo precizno definisali information gain, moramo definisati *entropiju*.
- Entropijom merimo homogenost (uređenost) skupa podataka.
- Ako je S skup pozitivnih (onih za koje je odluka True) i negativnih primera (onih za koje je odluka false), entropija skupa S , u odnosu na ovu binarnu klasifikaciju primera, je:

$$Entropy(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

gde je $p_{\oplus}$ udeo pozitivnih primera u S a $p_{\ominus}$, je udeo negativnih primera u S

Karakteristike entropije

- Entropija je 0 ako svi primeri iz S pripadaju istoj klasi.
- Na primer, ako su svi pozitivni ($p_{\oplus} = 1$), tada je $p_{\ominus} = 0$, i $Entropy(S) = -\log_2 1 = 0$
- Entropija je 1 ako ne posedujemo nikakvo znanje o sistemu (ili ako je svaki ishod podjednako moguć).
- Entropija je 1 ako kolekcija primera sadrži jednak broj pozitivnih i negativnih primera.
- Ako broj pozitivnih i negativnih primera nije jednak vrednost entropije je između 0 i 1.

- Entropija je 0 ako svi primeri iz S pripadaju istoj klasi.
- Na primer, ako su svi pozitivni ($p_{\oplus} = 1$), tada je $p_{\ominus} = 0$, i $Entropy(S) = -\log_2 1 = 0$
- Entropija je 1 ako ne posedujemo nikakvo znanje o sistemu (ili ako je svaki ishod podjednako moguć).
- Entropija je 1 ako kolekcija primera sadrži jednak broj pozitivnih i negativnih primera.
- Ako broj pozitivnih i negativnih primera nije jednak vrednost entropije je između 0 i 1.

Entropy of a 2-class problem with regard to the portion of one of the two groups

Primer

- Entropija skupa podataka iz prethodnog primera
- Skup ima 8 primera od kojih:
 - 3 pozitivna primera (oni koji jesu vampiri)
 - 5 negativnih primera (oni koji nisu vampiri)

$$p_{\oplus} = \frac{3}{8} \quad p_{\ominus} = \frac{5}{8}$$

$$Entropy(S) = -\frac{3}{8} \log_2 \frac{3}{8} - \frac{5}{8} \log_2 \frac{5}{8} = -\frac{3}{8} \cdot (-1.415) - \frac{5}{8} \cdot (-0.678) =$$

$$= -0.375 \cdot (-1.415) - 0.625 \cdot (-0.678) = 0.531 + 0.424 = 0.955$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

-
- Ako ciljna funkcija može uzeti neku od c različitih vrednosti (npr. ako klasifikujemo voće na jabuke, kruške i pomorandže, tada je $c=3$), onda je entropija od S

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

Gde je p_i udeo primera iz S koji pripadaju klasi i .

Šta ako
klasifikacija
nije binarna?

- Maksimalna vrednost entropije je sada $\log_2 c$

Informaciono poboljšanje (Information Gain)

- Entropija je mera neizvesnosti kolekcije trening primera, koja nam omogućava da definišemo meru efikasnosti nekog atributa u klasifikovanju trening podataka. Ta mera, **information gain**, predstavlja očekivano umanjenje entropije nakon podele trening primera korišćenjem nekog atributa.
- Preciznije, informaciono poboljšanje, $Gain(S, A)$ koje se dobija raspodelom primera iz S na osnovu njihovih vrednosti za atribut A je:

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

Gde je $Values(A)$ skup svih mogućih vrednosti atributa A , a S_v , je podskup primera iz S za koje atribut A ima vrednost v , tj. $S_v = \{s \in S | A(s) = v\}$.

Informaciono poboljšanje

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

- Prvi deo prethodne jednakosti je entropija originalnog skupa S , a drugi izraz u jednakosti je očekivana vrednost entropije nakon podele primera iz S prema vrednosti atributa A .

Izračunavanje $Gain(S, A_1)$ za atribut A_1 „Ima senku“

$$v \in Values(A_1) = \{?, Da, Ne\}$$

Vrednost atributa A_1 koja je jednaka ? ($v = ?$) ima 4 primera iz skupa $\Rightarrow |S_{v=?}|=4$.
 Vrednost atributa A_1 koja je jednaka Da ($v = Da$) ima 3 primera iz skupa $\Rightarrow |S_{v=Da}|=3$.
 Vrednost atributa A_1 koja je jednaka Ne ($v = Ne$) ima 1 primer iz skupa $\Rightarrow |S_{v=Ne}|=1$.

$$\frac{|S_{v=?}|}{|S|} = \frac{4}{8}, \quad \frac{|S_{v=da}|}{|S|} = \frac{3}{8}, \quad \frac{|S_{v=Ne}|}{|S|} = \frac{1}{8}$$

$$Entropy(S_v) = -p_{v\oplus} \log_2 p_{v\oplus} - p_{v\ominus} \log_2 p_{v\ominus}$$

$Entropy(S_{v=?}) = 1$ podjednak udeo pozitivnih i negativnih primera

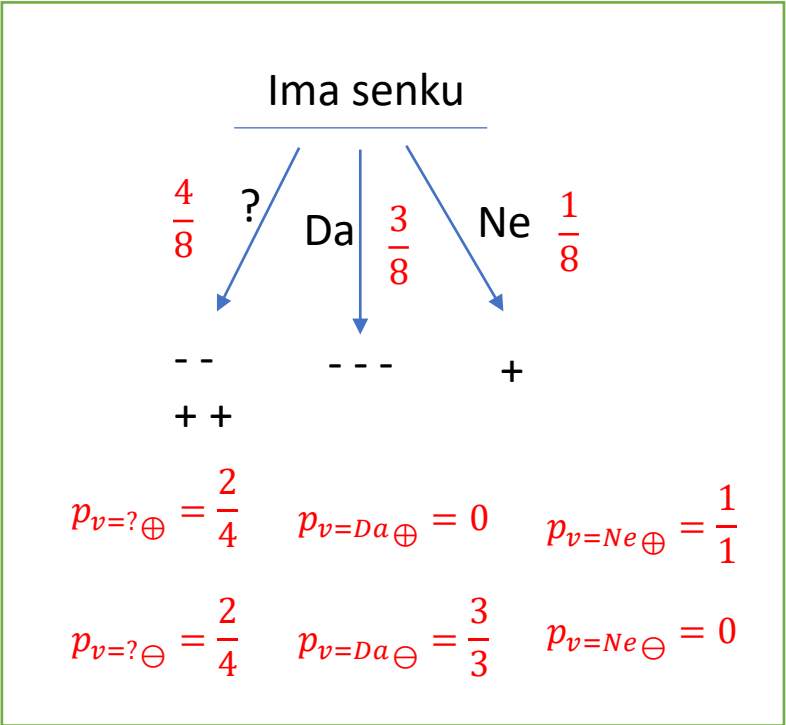
$Entropy(S_{v=Da}) = 0$ svi primeri pripadaju istoj klasi

$Entropy(S_{v=Ne}) = 0$ svi primeri pripadaju istoj klasi

$$Gain(S, A_1) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) =$$

$$0.955 - \left(\frac{4}{8} \cdot 1 + \frac{3}{8} \cdot 0 + \frac{1}{8} \cdot 0 \right) = \boxed{0.455}$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



Izračunavanje $Gain(S, A_2)$ za atribut A_2 „Jede beli luk“

$$v \in Values(A_2) = \{Da, Ne\}$$

Vrednost atributa A_2 koja je jednaka Da ($v = Da$) ima 3 primera iz skupa $\Rightarrow |S_{v=Da}|=3$.

Vrednost atributa A_2 koja je jednaka Ne ($v = Ne$) ima 5 primera iz skupa $\Rightarrow |S_{v=Ne}|=5$.

$$\frac{|S_{v=Da}|}{|S|} = \frac{3}{8}, \quad \frac{|S_{v=Ne}|}{|S|} = \frac{5}{8}$$

$$Entropy(S_v) = -p_{v\oplus} \log_2 p_{v\oplus} - p_{v\ominus} \log_2 p_{v\ominus}$$

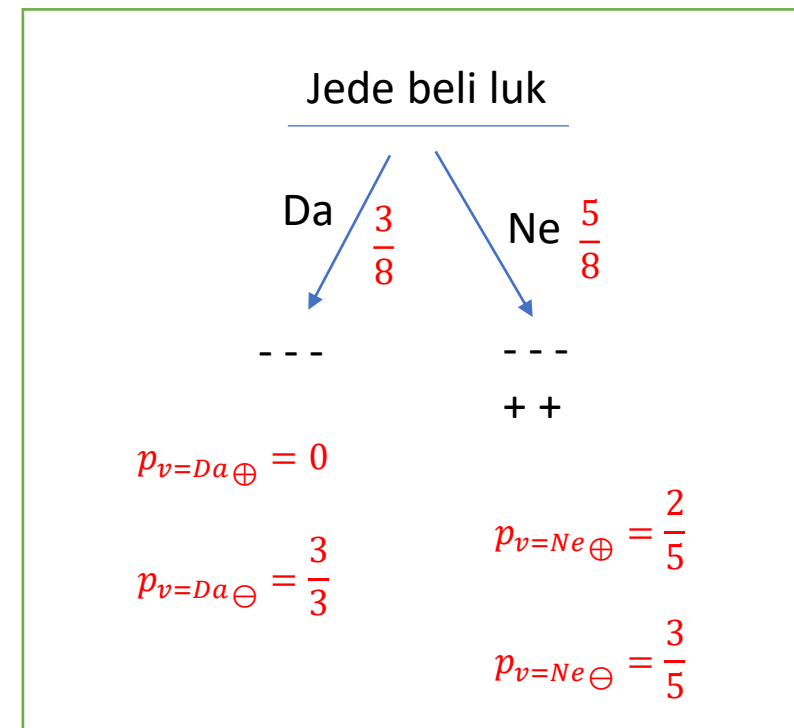
$Entropy(S_{v=Da}) = 0$ svi primeri pripadaju istoj klasi

$$Entropy(S_{v=Ne}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$Gain(S, A_2) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) =$$

$$0.955 - \left(\frac{3}{8} \cdot 0 + \frac{5}{8} \cdot 0.97 \right) = \boxed{0.35}$$

	Ima senku	Jede beli luk	Ten	Akcent	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



Izračunavanje $Gain(S, A_3)$ za atribut A_3 „Ten“

$$v \in Values(A_3) = \{Bled, Rumen, Prosečan\}$$

Vrednost atributa A_3 koja je jednaka ? ($v = Bled$) ima 4 primera iz skupa $\Rightarrow |S_{v=B}|=2$.
Vrednost atributa A_3 koja je jednaka Da ($v = Rumen$) ima 3 primera iz skupa $\Rightarrow |S_{v=R}|=3$.
Vrednost atributa A_3 koja je jednaka Ne ($v = Prosečan$) ima 3 primera iz skupa $\Rightarrow |S_{v=P}|=3$.

$$\frac{|S_{v=B}|}{|S|} = \frac{2}{8}, \quad \frac{|S_{v=R}|}{|S|} = \frac{3}{8}, \quad \frac{|S_{v=P}|}{|S|} = \frac{3}{8}$$

$$Entropy(S_v) = -p_{v\oplus} \log_2 p_{v\oplus} - p_{v\ominus} \log_2 p_{v\ominus}$$

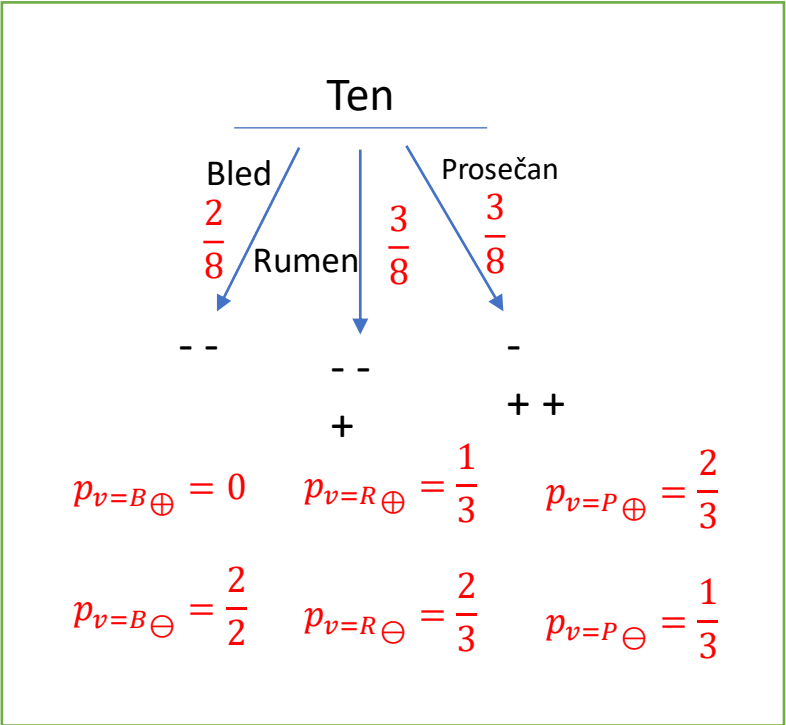
$Entropy(S_{v=B}) = 0$ svi primeri pripadaju istoj klasi

$$Entropy(S_{v=R}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9$$

$$Entropy(S_{v=P}) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9$$

$$Gain(S, A_3) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) =$$
$$0.955 - \left(\frac{2}{8} \cdot 0 + \frac{3}{8} \cdot 0.9 + \frac{3}{8} \cdot 0.9 \right) = \boxed{0.28}$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



Izračunavanje $Gain(S, A_4)$ za atribut A_4 „Akcenat“

$$v \in Values(A_4) = \{Nema, Te\breve{z}ak, \breve{C}udan\}$$

Vrednost atributa A_4 koja je jednaka ? ($v = Nema$) ima 3 primera iz skupa $\Rightarrow |S_{v=N}|=3$.
 Vrednost atributa A_4 koja je jednaka Da ($v = Te\breve{z}ak$) ima 3 primera iz skupa $\Rightarrow |S_{v=T}|=3$.
 Vrednost atributa A_4 koja je jednaka Ne ($v = \breve{C}udan$) ima 2 primer iz skupa $\Rightarrow |S_{v=\breve{C}}|=2$.

$$\frac{|S_{v=N}|}{|S|} = \frac{3}{8}, \quad \frac{|S_{v=T}|}{|S|} = \frac{3}{8}, \quad \frac{|S_{v=\breve{C}}|}{|S|} = \frac{2}{8}$$

$$Entropy(S_v) = -p_{v\oplus} \log_2 p_{v\oplus} - p_{v\ominus} \log_2 p_{v\ominus}$$

$$Entropy(S_{v=N}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9$$

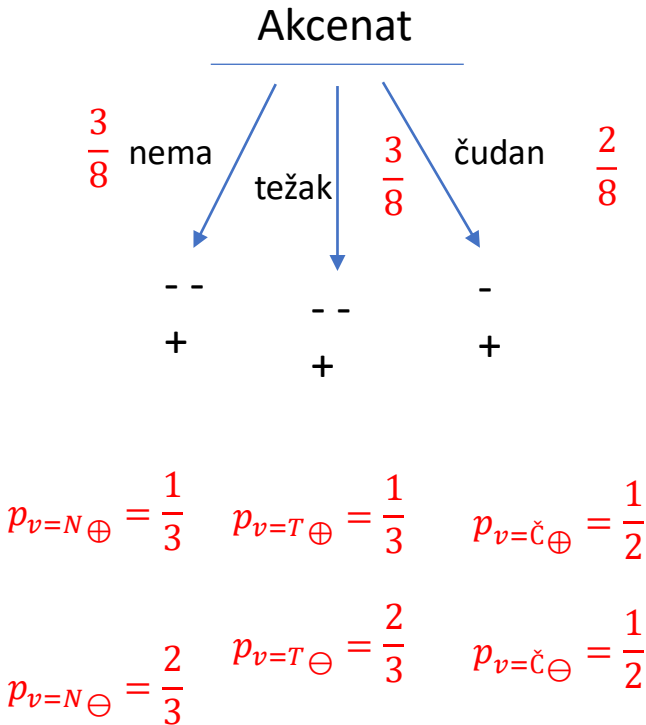
$$Entropy(S_{v=T}) = -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} = 0.9$$

$$Entropy(S_{v=\breve{C}}) = 1 \text{ podjednak udeo pozitivnih i negativnih primera}$$

$$Gain(S, A_4) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) =$$

$$0.955 - \left(\frac{3}{8} \cdot 0.9 + \frac{3}{8} \cdot 0.9 + \frac{2}{8} \cdot 1 \right) = \boxed{0.03}$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



$$\text{Gain}(S, A_1) = 0.455$$

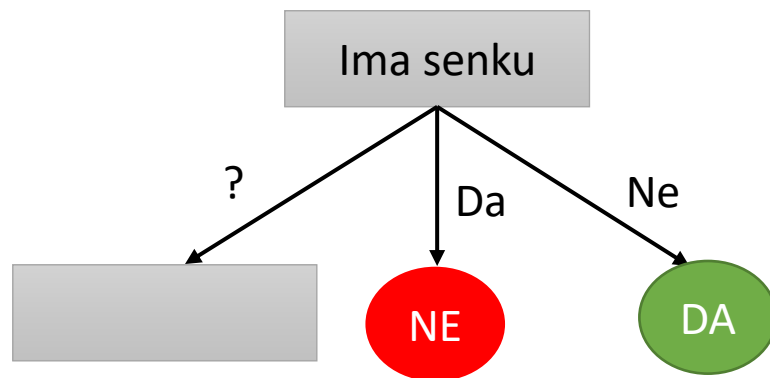
$$\text{Gain}(S, A_2) = 0.35$$

$$\text{Gain}(S, A_3) = 0.28$$

$$\text{Gain}(S, A_4) = 0.03$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

Atribut A_1 stavljamo u koren stabla, jer ima najveći *Information gain*



Sada treba odrediti koji od preostala 3 atributa ima najveći Information gain nad primerima koji nisu klasifikovani korišćenjem atributa „Ima senku“, i on postaje koren podstabla.

- Prvo računamo Entropiju skupa preostalih podataka
- Skup ima 4 primera od kojih:
 - 2 pozitivna primera (oni koji jesu vampiri)
 - 2 negativnih primera (oni koji nisu vampiri)

$$p_{\oplus} = \frac{2}{4} = \frac{1}{2} \quad p_{\ominus} = \frac{2}{4} = \frac{1}{2}$$

$$Entropy(S) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = -2 \cdot \frac{1}{2} \log_2 \frac{1}{2} = (-1) \cdot (-1) = 1$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

Izračunavanje $Gain(S, A_2)$ za atribut A_2 „Jede beli luk“

$$v \in Values(A_2) = \{Da, Ne\}$$

Vrednost atributa A_2 koja je jednaka Da ($v = Da$) ima 2 primera iz skupa $\Rightarrow |S_{v=Da}|=2$.

Vrednost atributa A_2 koja je jednaka Ne ($v = Ne$) ima 2 primera iz skupa $\Rightarrow |S_{v=Ne}|=2$.

$$\frac{|S_{v=da}|}{|S|} = \frac{2}{4}, \quad \frac{|S_{v=Ne}|}{|S|} = \frac{2}{4}$$

$$Entropy(S_v) = -p_{v\oplus} \log_2 p_{v\oplus} - p_{v\ominus} \log_2 p_{v\ominus}$$

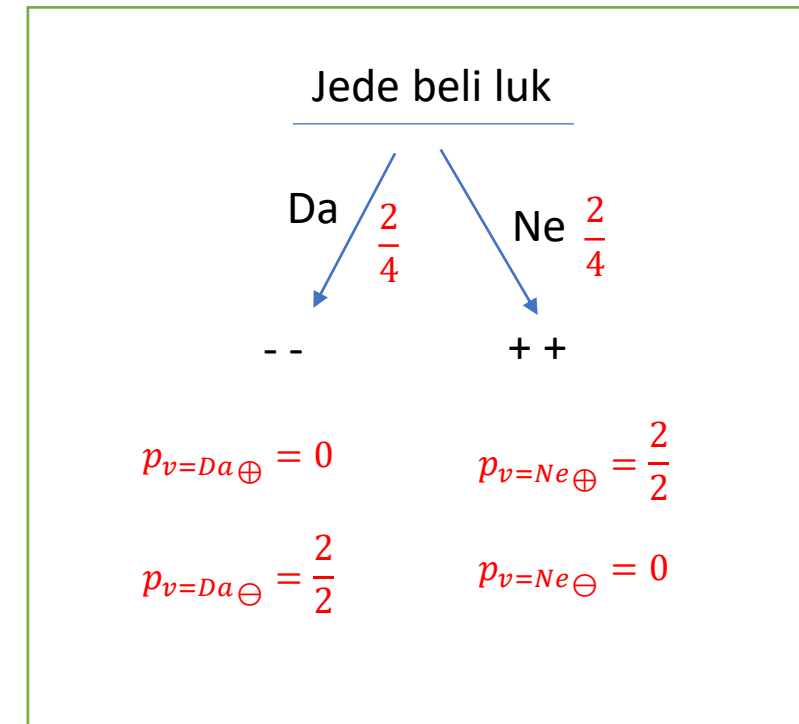
$Entropy(S_{v=Da}) = 0$ svi primeri pripadaju istoj klasi

$Entropy(S_{v=Ne}) = 0$ svi primeri pripadaju istoj klasi

$$Gain(S, A_2) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) =$$

$$1 - \left(\frac{2}{4} \cdot 0 + \frac{2}{4} \cdot 0 \right) = 1$$

	Ima senku	Jede beli luk	Ten	Akcentat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



Izračunavanje $Gain(S, A_3)$ za atribut A_3 „Ten“

$$v \in Values(A_3) = \{Bled, Rumen, Prosečan\}$$

Vrednost atributa A_3 koja je jednaka ? ($v = Bled$) ima 1 primera iz skupa $\Rightarrow |S_{v=B}|=1$.
 Vrednost atributa A_3 koja je jednaka Da ($v = Rumen$) ima 2 primera iz skupa $\Rightarrow |S_{v=R}|=2$.
 Vrednost atributa A_3 koja je jednaka Ne ($v = Prosečan$) ima 1 primer iz skupa $\Rightarrow |S_{v=P}|=1$

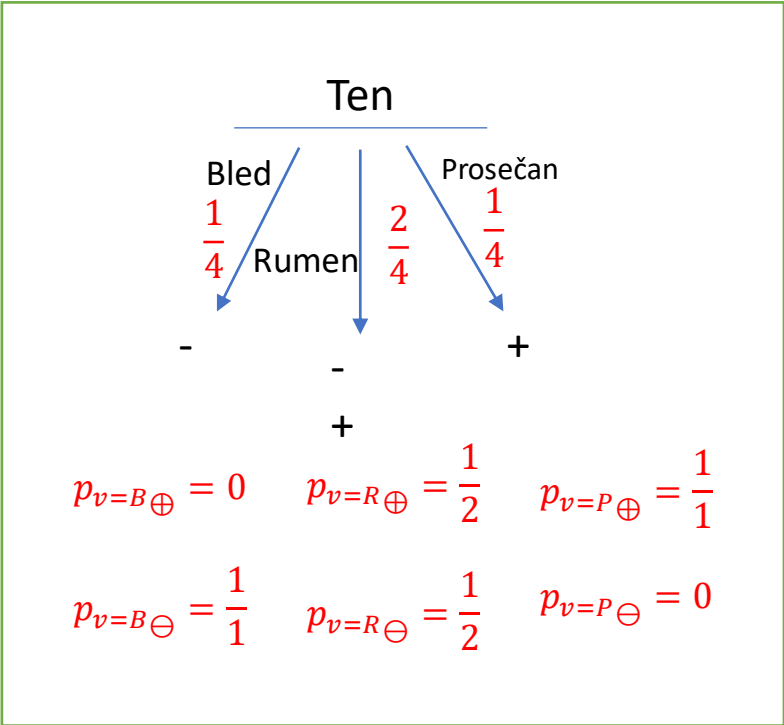
$$\frac{|S_{v=B}|}{|S|} = \frac{1}{4}, \quad \frac{|S_{v=R}|}{|S|} = \frac{2}{4}, \quad \frac{|S_{v=P}|}{|S|} = \frac{1}{4}$$

$$Entropy(S_v) = -p_{v\oplus} \log_2 p_{v\oplus} - p_{v\ominus} \log_2 p_{v\ominus}$$

$Entropy(S_{v=B}) = 0$ svi primeri pripadaju istoj klasi
 $Entropy(S_{v=R}) = 1$ podjednak udeo pozitivnih i negativnih primera
 $Entropy(S_{v=P}) = 0$ svi primeri pripadaju istoj klasi

$$Gain(S, A_3) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) = 1 - \left(\frac{1}{4} \cdot 0 + \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 0\right) = 0.5$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne



Izračunavanje $Gain(S, A_4)$ za atribut A_4 „Akcenat“

$$v \in Values(A_4) = \{Nema, Te\breve{z}ak, \breve{C}udan\}$$

Vrednost atributa A_4 koja je jednaka ? ($v = Nema$) ima 2 primera iz skupa $\implies |S_{v=N}|=2$.
 Vrednost atributa A_4 koja je jednaka Da ($v = Te\breve{z}ak$) ima 0 primera iz skupa $\implies |S_{v=T}|=0$.
 Vrednost atributa A_4 koja je jednaka Ne ($v = \breve{C}udan$) ima 2 primer iz skupa $\implies |S_{v=\breve{C}}|=2$.

$$\frac{|S_{v=N}|}{|S|} = \frac{2}{4}, \quad \frac{|S_{v=T}|}{|S|} = 0, \quad \frac{|S_{v=\breve{C}}|}{|S|} = \frac{2}{4}$$

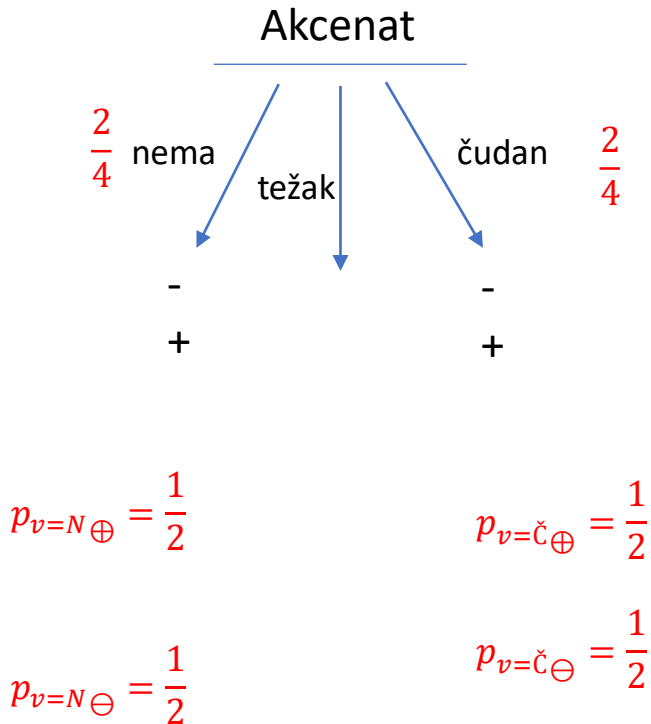
$$Entropy(S_v) = -p_{v\oplus} \log_2 p_{v\oplus} - p_{v\ominus} \log_2 p_{v\ominus}$$

$Entropy(S_{v=N}) = 1$ podjednak udeo pozitivnih i negativnih primera

$Entropy(S_{v=\breve{C}}) = 1$ podjednak udeo pozitivnih i negativnih primera

$$Gain(S, A_4) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) = 1 - \left(\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1\right) = 0$$

	Ima senku	Jede beli luk	Ten	Akcenat	Vampir
1	?	Da	Bled	Nema	Ne
2	Da	Da	Rumen	Nema	Ne
3	?	Ne	Rumen	Nema	Da
4	Ne	Ne	Prosečan	Težak	Da
5	?	Ne	Prosečan	Čudan	Da
6	Da	Ne	Bled	Težak	Ne
7	Da	Ne	Prosečan	Težak	Ne
8	?	da	Rumen	Čudan	Ne

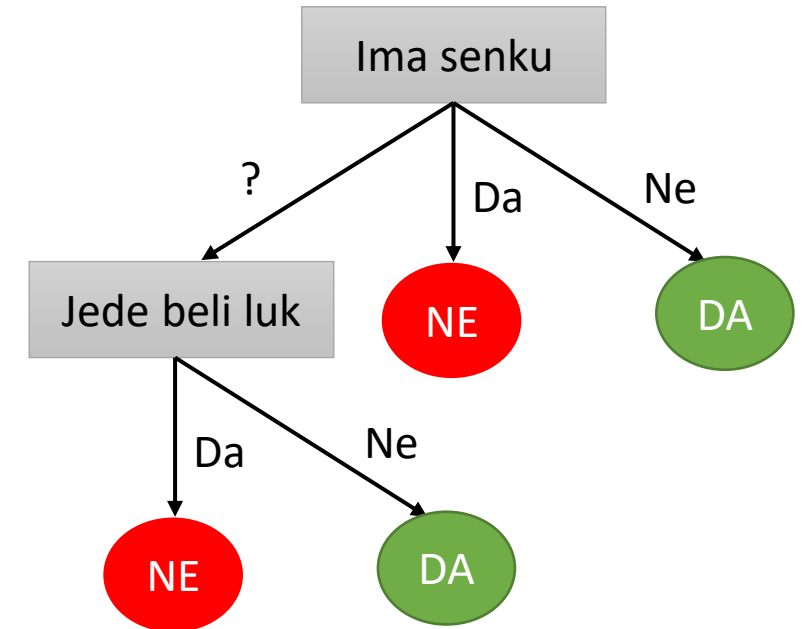


$$\text{Gain}(S, A_2) = 1$$

$$\text{Gain}(S, A_3) = 0.5$$

$$\text{Gain}(S, A_4) = 0$$

Atribut A_2 stavljamo u koren podstabla, jer ima najveći *Information gain*



Primer

Klasifikacija primera na pozitivne (Yes) ili negativne (No)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Oznake primera

Ne koriste se za obučavanje

Primer

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- 14 primera za učenje
- 9 pozitivnih
- 5 negativnih

$$Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

Informaciono poboljšanje za atribut Wind

$$\text{Values}(\text{Wind}) = \text{Weak}, \text{Strong}$$

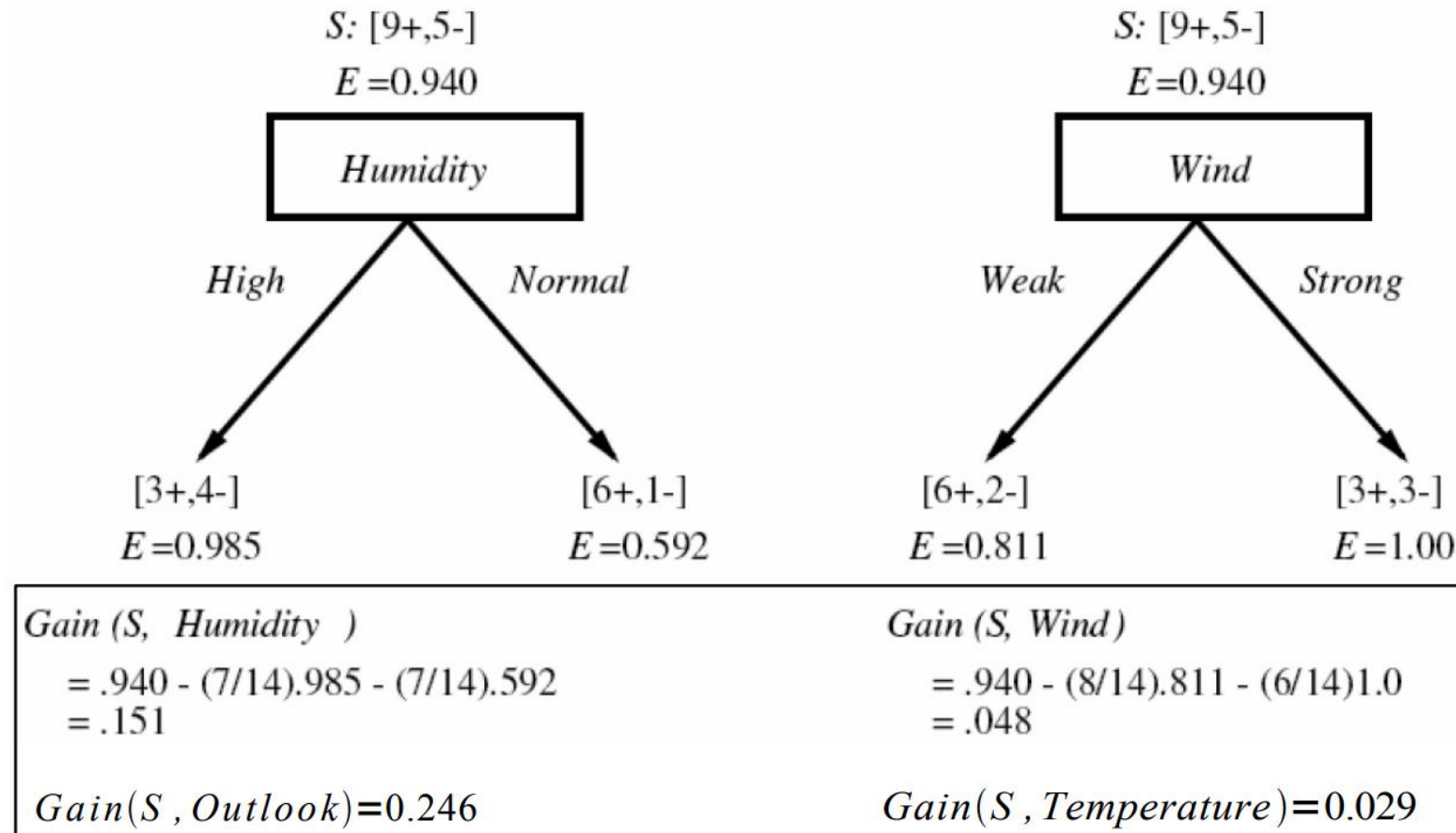
$$S = [9+, 5-]$$

$$S_{\text{Weak}} \leftarrow [6+, 2-]$$

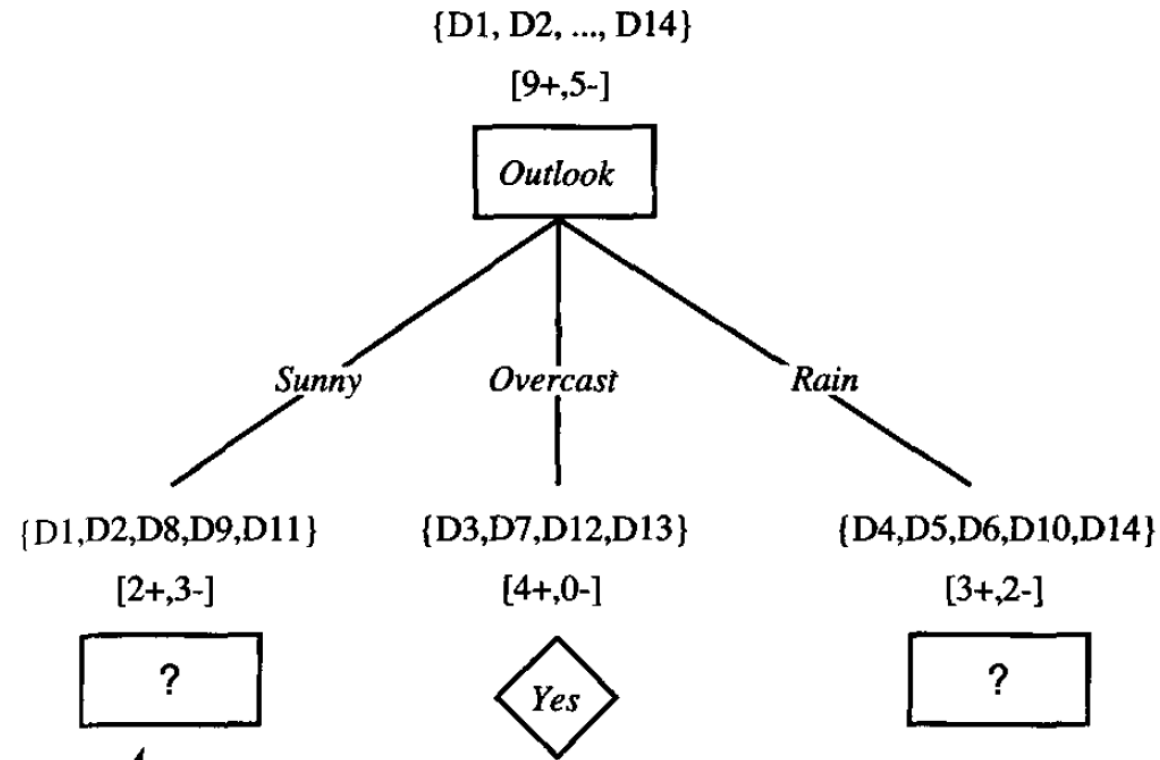
$$S_{\text{Strong}} \leftarrow [3+, 3-]$$

$$\begin{aligned}\text{Gain}(S, \text{Wind}) &= \text{Entropy}(S) - \sum_{v \in \{\text{Weak}, \text{Strong}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \\ &= \text{Entropy}(S) - (8/14) \text{Entropy}(S_{\text{Weak}}) \\ &\quad - (6/14) \text{Entropy}(S_{\text{Strong}}) \\ &= 0.940 - (8/14)0.811 - (6/14)1.00 \\ &= 0.048\end{aligned}$$

Informaciono poboljšanje za svaki od atributa



Nakon što atribut Outlook postavimo u koren stabla:



Koji atribut treba testirati u ovom čvoru?

Sledeći najbolji atribut je Humidity

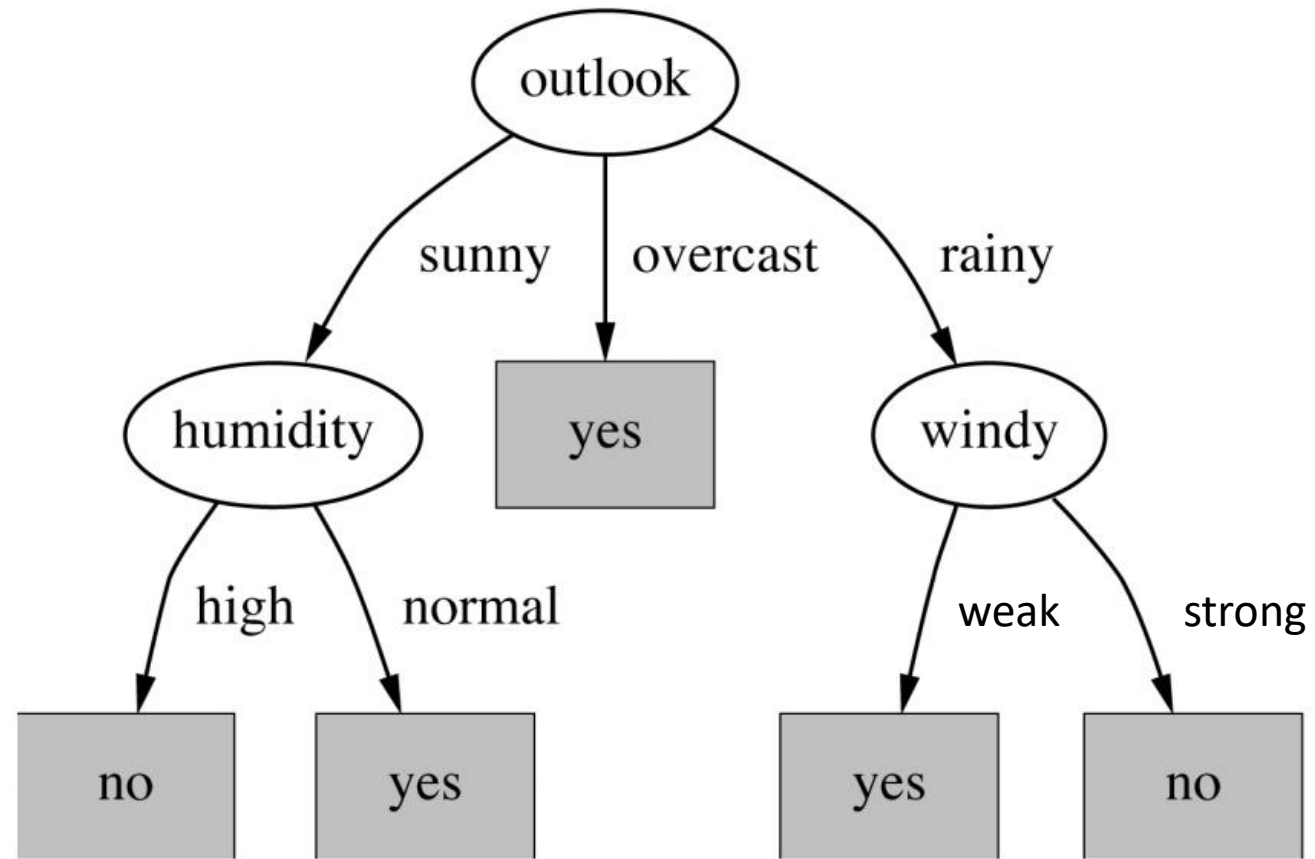
$$S_{\text{sunny}} = \{D1, D2, D8, D9, D11\}$$

$$\text{Gain}(S_{\text{sunny}}, \text{Humidity}) = .970 - (3/5) 0.0 - (2/5) 0.0 = .970$$

$$\text{Gain}(S_{\text{sunny}}, \text{Temperature}) = .970 - (2/5) 0.0 - (2/5) 1.0 - (1/5) 0.0 = .570$$

$$\text{Gain}(S_{\text{sunny}}, \text{Wind}) = .970 - (2/5) 1.0 - (3/5) .918 = .019$$

Konačno stablo odluke

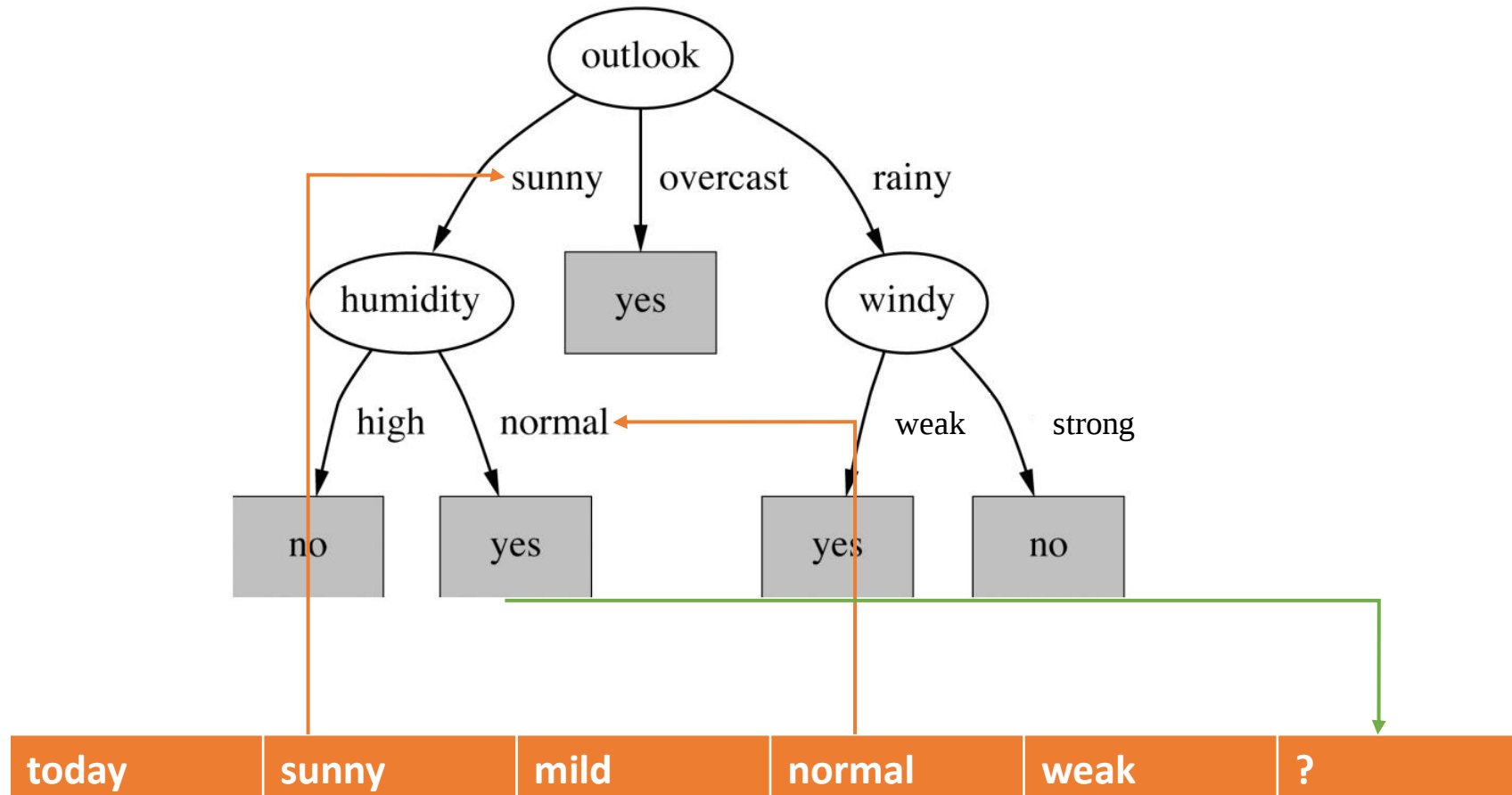


- Ako znamo da je danas sunčano, da je umereno toplo, da je normalna vlažnost vazduha i da duva slab vetar, da li ćemo igrati tenis?

today	sunny	mild	normal	weak	?
-------	-------	------	--------	------	---

- Na ovo pitanje treba da nam da odgovor stablo odlučivanja.
- Obučavanjem stabla odlučivanja dobija se mehanizam za klasifikaciju koji može naučena pravila da primeni na nove primere i donese odluku.

Predikcija pomoću formiranog stabla



Gini indeks

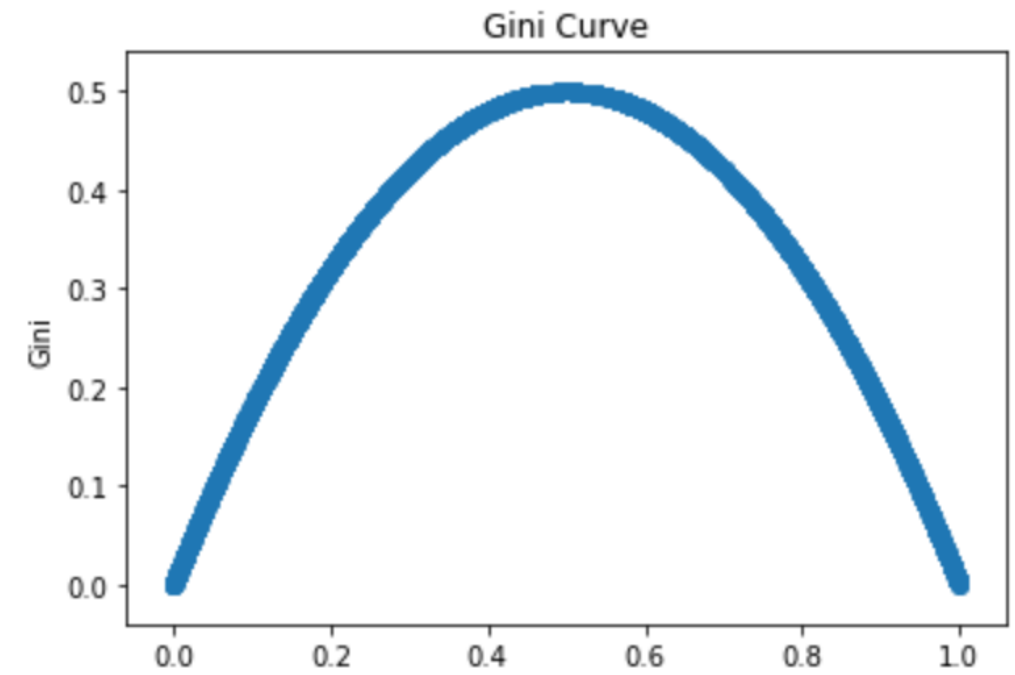
- Gini impurity meri nečistoću instanci razvrstanih u nekom čvoru – cilj je da sve instance koje „odu“ niz jednu granu drвета pripadaju istoj klasi.
- Za svaki od k atributa izračunavamo:

$$G_k = 1 - \sum_{i=1}^c P_i^2$$

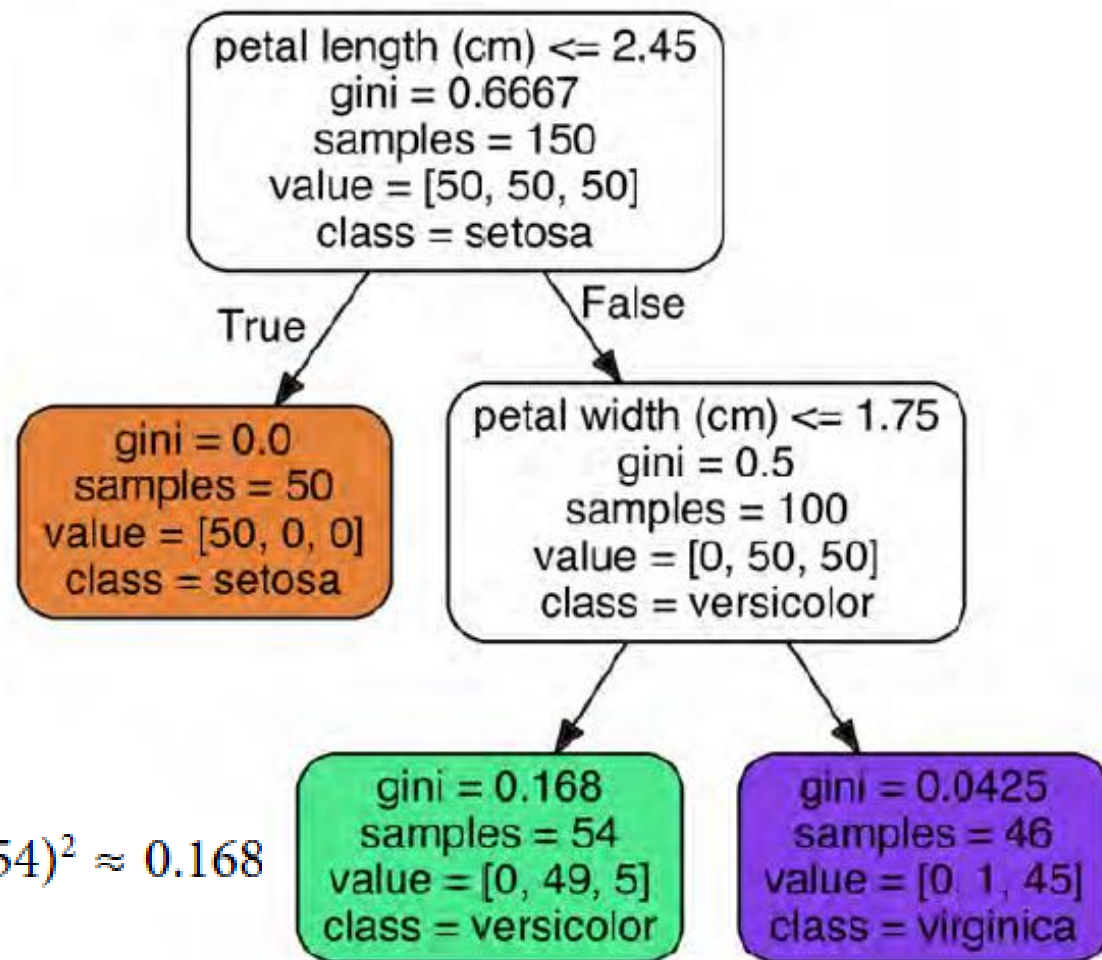
P_i je udeo instanci klase i među instancama u k -tom čvoru

Out of 14 instances ,
yes=9,no=5
 $1 - (9/14)^2 - (5/14)^2$

 $1 - 0.413 - 0.127 = 0.46$
Gini = 0.46



Stablo odlučivanja za Iris skup podataka



$$1 - (0/54)^2 - (49/54)^2 - (5/54)^2 \approx 0.168$$