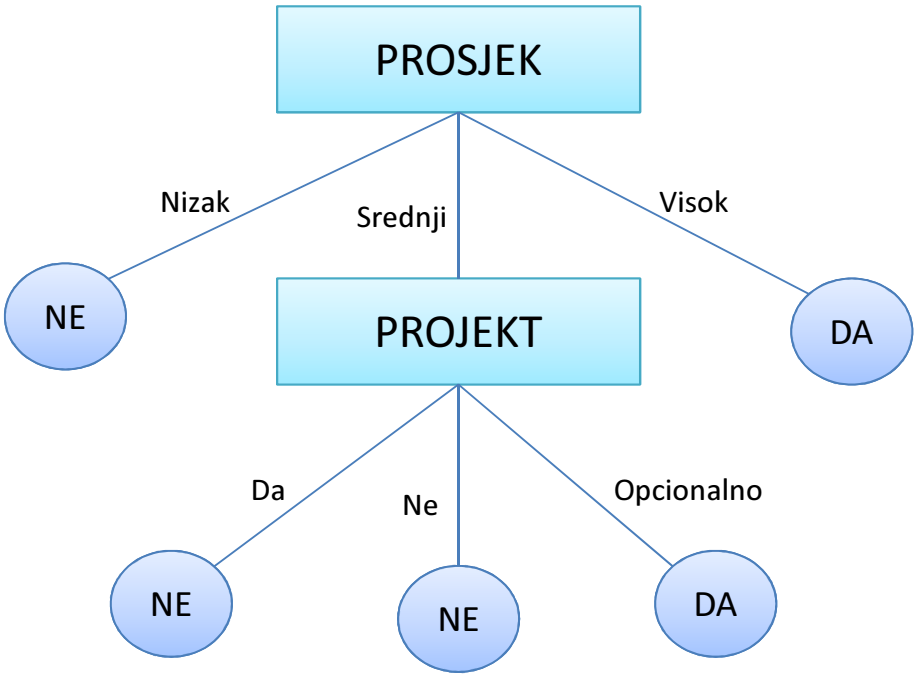
Zadatak 1: Stabla odluke

(a) Uporabom algoritma ID3 izgradite stablo odluke za klasifikaciju primjera u klasu "Popularan izborni predmet". Primjeri za učenje su sljedeći:

i	Ispit	Projekt	Predavanja	Labosi	Prosjek	$h(\mathbf{x}^{(i)})$
1	pismeni	da	obavezna	ne	visok	1
2	usmeni	da	neobavezna	ne	srednji	0
3	oboje	opcionalno	obavezna	da	nizak	0
4	oboje	ne	neobavezna	ne	visok	1
5	pismeni	ne	obavezna	da	srednji	0
6	usmeni	opcionalno	neobavezna	ne	srednji	1
7	pismeni	ne	obavezna	da	nizak	0
8	pismeni	ne	neobavezna	ne	visok	1

Korištenjem ID3 algoritma dobio sam sljedeće stablo:



$$Informacijska\_dobit(D,A) \equiv Entropija(D) - \sum_{v \in Vrijednost(A)} \frac{|Dv|}{|D|} Entropija(Dv)$$

Entropija izvornog skupa D

Očekivana vrijednost entropije nakon podjele D na temelju atributa A

Vrijednost(A) - skup svih mogućih vrijednosti atributa A
Dv - podskup od D za koji atribut A ima vrijednost v, tj.
Dv = {x ∈ D| A(x)= v}

Ako želite rezultate: (informacijske dobiti – ID)
[u 1° D je skup svih primjera, u 2° D je skup primjera kojima je prosjek = srednji]
1° ID(D,ispit) = 0, ID(D,projekt) = 0, ID(D,predavanja) = 0.189, ID(D,labosi) = 0.548, ID(D,prosjek) = 0.656
2° ID(D,ispit) = 0.252, ID(D,projekt) = 0.918, ID(D,predavanja) = 0.252, ID(D,labosi) = 0.252

(b) Pretpostavite da kod 6. primjera nedostaje vrijednost značajke *Prosjek*. Kako biste riješili taj problem?

Postoje tri pristupa tom problemu. Ja osobno bih koristio pristup u kojem prilikom računanja entropija koristimo relativne frekvencije poznatih primjera. Dakle, informacijsku dobit bih računao:

Entropija([+4,-4]) – 3/7 * Entropija([+3,0]) – 2/7 * Entropija([0,+2]) – 2/7 * Entropija([0,+2])

Druga dva pristupa su dodjeljivanje najčešće pojavljivane vrijednosti atributa.

(c) Može li se dogoditi da jedan se jedan te isti primjer $\mathbf{x}^{(i)}$ u skupu za učenje pojavi dva puta i to s različitom oznakom $y^{(i)}$? Zbog čega bi se to moglo dogoditi? Kako se taj problem rješava kod stabla odluke? Ilustrirajte na gornjem skupu podataka.

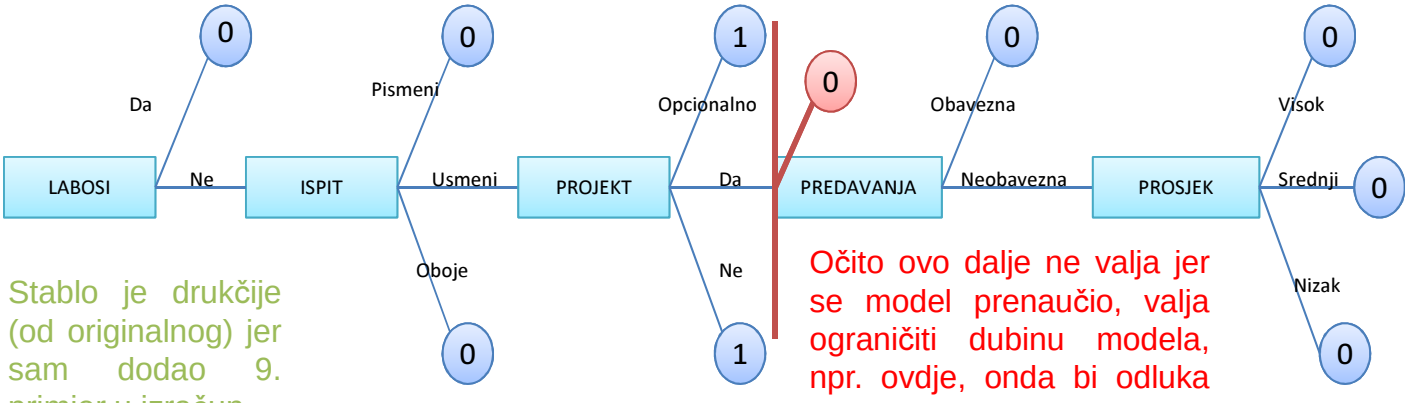
Može. Razlog može biti greška u označavanju (netko slučajno stavio primjer $x^{(i)}$ sa istim atributima a različitom oznakom $y^{(i)}$), ili postojanje latentnih varijabli (nismo dobro odabrali atribute, postoje atributi koji nisu u našem modelu, model ne obuhvaća sve atribute).

Kod stabala odluke taj problem se rješava tako da... Hm, dajte da prvo objasnim što se događa xD. Prvi problem je taj što će stablo odluke stvarati sve dublje stablo, dok ne potroši sve atribute (jer su svi atributi isti a oznake različite), a to je problem SLOŽENOSTI koji se rješava određivanjem dubine do koje se stvaraju čvorovi. E sad, kad se dođe do 'kraja', dakle, prošli smo sve atribute i dalje imamo primjere za različitim oznakama, algoritam ID3 će staviti čvor odluke sa oznakom koja je jednaka najčešće pojavljivanoj oznaci u skupu D prijašnjeg čvora.

U kodu:

Stvori korijen stabla ROOT
Ako su svi primjeri pozitivni, vrati stablo s jednim čvorom čija je oznaka = +
Ako su svi primjeri negativni, vrati stablo s jednim čvorom čija je oznaka = -
Ako je atribut prazan, vrati stablo s jednim čvorom ROOT, s oznakom = najčešća vrijednost *Ciljnog _atributa* u skupu
(Q: kada se ovo događa?)

U našem primjeru, recimo da imamo još i 9. primjer koji je jednak primjeru 2 ali oznaka = 1.



Stablo je drukčije (od originalnog) jer sam dodao 9. primjer u izračun.

Očito ovo dalje ne valja jer se model prenaučio, valja ograničiti dubinu modela, npr. ovdje, onda bi odluka za 'Da' bila 0.

(d) Stablo odluke je neparametarski model. Što to konkretno znači?

To znači da broj parametara modela ovisi o broju primjera. Konkretno: više primjera -> složeniji model.

(e) Razmatramo skupove primjera za učenje $\mathcal{D} \subset \mathbb{R}^2 \times \{0, 1\}$ za koje *ne* postoji hipoteza

$$h_1(x_1, x_2|\theta_0, \theta_1, \theta_2) = \mathbf{1}\{\theta_1 x_1 + \theta_2 x_2 + \theta_0 \geqslant 0\}$$

koja bi bila konzistentna s primjerima za učenje. Možemo li algoritmom ID3 za svaki takav skup $\mathcal{D}$ naučiti hipotezu h_2 koja jest konzistentna s primjerima za učenje? Obrazložite odgovor.

Možemo. ID3 ima svojstvo da pretražuje prostor svih hipoteza, tako da on može naučiti hipotezu koja je konzistentna s svim primjerima za učenje.

(f) Odgovorite na gornje pitanje uz pretpostavku da je stablo odluke ograničeno na dubinu 3. Ima li općenito smisla ograničiti stablo na neku dubinu? Kako biste odabrali na koju dubinu ga ograničiti? Kako biste to napravili ako na raspolaganju imate samo navedenih osam primjera za učenje?

Opet isti odgovor, jer i sa ograničenjem na dubinu tri čvora opet dobivamo rješenje unutar te granice. Općenito ograničavanje stabla na neku dubinu ima smisla, jer time ograničavamo složenost stabla. Kako odrediti dubinu? Kao i uvijek kad biramo neki model, koristili bismo unakrsnu provjeru, vidjeli koji nam model ima najbolju pogrešku generalizacije, i taj model koristili. Ne bih to radio na našem skupu, jer je skup malen te već samim time ne može postati pretjerano složen, jer je model neparametarski, a imamo malo primjera (a samim time i parametara).