VEŠTAČKA INTELIGENCIJA

STABLA ODLUKE
ID3 ALGORITAM

Sadržaj

- Osnovne postavke
- Pregled algoritama
- ID3

Prof. dr Leonid Stoimenov

Učenje

- Akvizicija novog deklarativnog znanja
- Ravoj motornih veština
- Razvoj kognitivnih veština
- Otkrivanje novih činjenica/zakona na osnovu eksperimenata
- Reorganizacija znanja
- Učenje jezika
- …

Inteligentni agent koji uči može da:

- unapređuje svoje performanse
- na osnovu iskustva E koje stiče
- obavljajući neki zadatak T
- u svom okruženju,
- u odnosu na meru performansi P.

Učenje

- Unapredjivanje performansi u vremenu
- izvršavajući stare zadatke efektivnije ili efikasnije
- ili obavljajući zadatke koji se ranije nisu mogli izvršavati
- **bez ponovnog programiranja**.
- Poboljšanje performansi sistema ostvaruje se:
  - adaptivnim promenama u memorisanju znanja,
  - precišćavanjem postojećeg i sticanjem novog znanja,
  - promenom memorijske reprezentacije,
  - promenom modela itd.

Pristupi mašinskom učenju

- **Induktivno učenje**
- Analitičko učenje (analogija sa logikom)
- Case-based Learning (analogija sa ljudskim pamćenjem)
- Neuronske mreže (analogija sa neurobiologijom)
- Genetski algoritmi (analogija sa evolucijom)
- Hibridni modeli

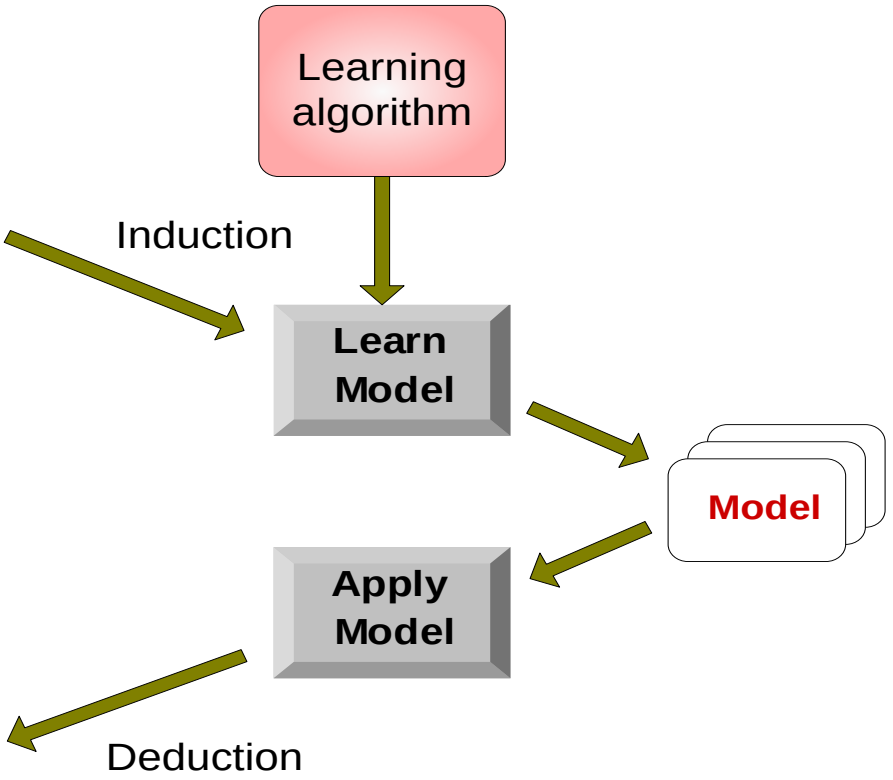
Problem klasifikacije

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

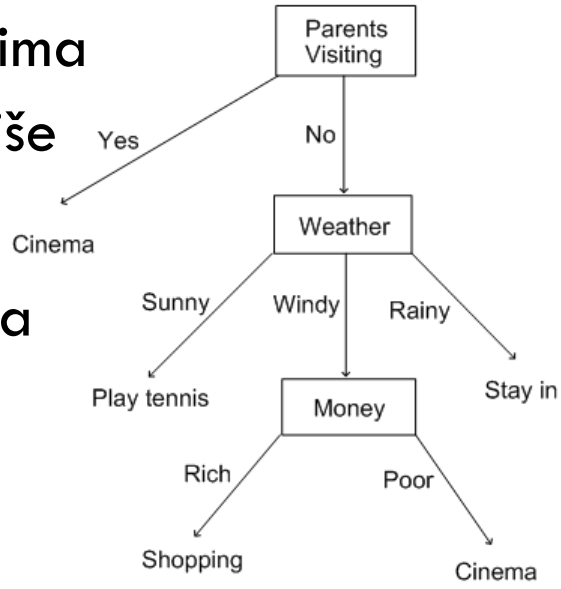


Algoritam učenja stablom odluke

- (Decision tree learning algorithm) - uspešno se primenjuje kod ekspertnih sistema za prikupljanje znanja.
- Osnovni zadatak kod takvih sistema:
  - korišćenjem **induktivnih metoda,**
    za zadate vrednosti atributa nepoznatog objekta, odrediti odgovarajuću **klasifikaciju**
    na osnovu pravila **Stabla odluke**.
- Šta je Stablo odluke?

Stablo odluke

- Stablo odluke predstavlja struktura tipa stabla gde:
  - **Unutrašnji čvorovi** odgovaraju atributima uzoraka i predstavljaju izbor između više alternativa,
  - (**grane** u stablu odgovaraju vrednostima određenog atributa)
  - **Listovi** predstavljaju odluke, odnosno klase kojima pripadaju uzorci sa vrednostima atributa definisanim putem do korena

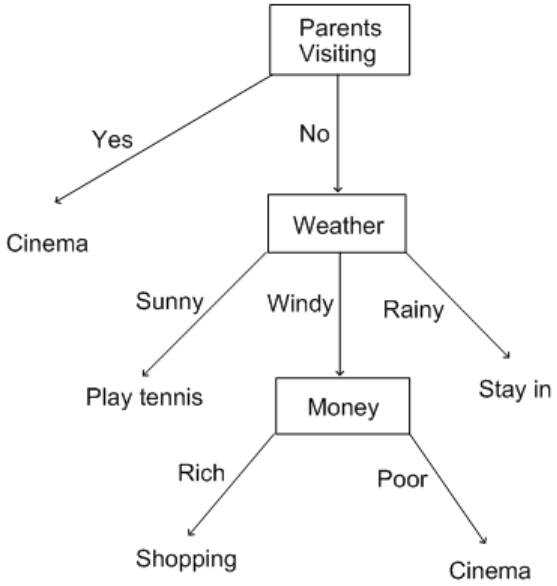


Izdvajanje klasifikacionih pravila iz stabla

- Stablo se može predstaviti u formi AKO-ONDA pravila (implikacija)
- Po jedno pravilo se kreira za svaku putanju od korena do lista
- Svi atribut-vrednost parovi na putanji formiraju konjunkciju
- Pravila su jednostavnija za razumevanje i primenu

Stablo odluke kao skup pravila

- (PARENTS_VISITING = YES) => CINEMA
- (PARENTS_VISITING = NO ∧ WEATHER = SUNNY) => PLAY_TENNIS
- (PARENTS_VISITING = NO ∧ WEATHER = WINDY ∧ MONEY = RICH) => SHOPPING

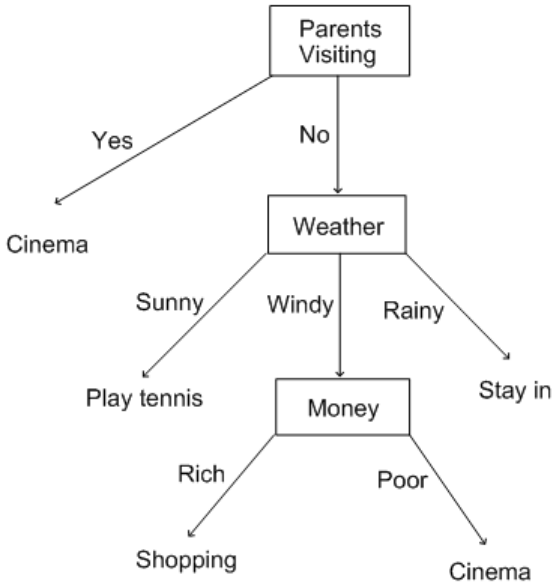


Kada su pogodna stabla odluke

- 1. Instance su predstavljene kao parovi atribut-vrednost.
  - Primer: atribut „Temperature“ i vrednosti 'hot', 'mild', 'cool'.
- 2. Ciljna funkcija ima diskretne izlazne vrednosti.
  - Jednostavna varijanta je da se radi sa instancama koje su povezane sa boolean odlukom, kao što su 'true' i 'false', ili 'p(positive)' i 'n(negative)'.
- 3. Trening podaci mogu da sadrže greške.

Kreiranje stabla odluke

- Dve faze:
  - Faza izgradnje stabla (*top-down*)
    - Na početku, svi primeri su u korenu stabla
    - Rekurzivno se vrši particionisanje primera odabirom po jednog atributa
  - Faza odsecanja stabla (*bottom-up*)
    - Uklanjanje podstabala ili grana u cilju unapređenja tačnosti modela



Faza izgradnje stabla

- “Pohlepna” strategija
  - Deli primere na osnovu testa na neki atribut, tako da podela bude optimalna po određenom kriterijumu
  - Algoritam bira najbolji atribut u datom trenutku i nikada se ne vraća da ponovo razmotri napravljene izbore
- Cilj je da stablo bude što manje, tj da se što pre dođe do odluke

Faza izgradnje stabla – pitanja

- Kako podeliti slogove?
  - Kako odrediti najbolji test na atribut?
  - Kako odrediti najbolji atribut za podelu?

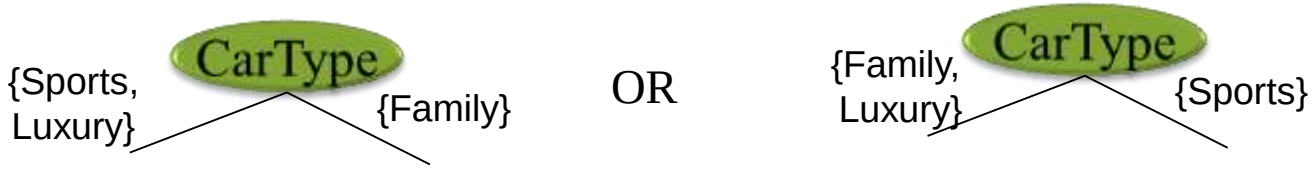
Kako odrediti atribut koji je na redu?

- Kriterijumi za podelu zavise od:
  - Tipa atributa
    - Diskretni (kategorički) - kriterijum u formi $A \in S'$, gde je S' podskup svih mogućih vrednosti atributa A
      - Nominalni – redosled nije bitan (plavo, crveno, žuto)
      - Ordinalni – redosled je bitan (nisko, srednje, visoko)
    - Kontinualni (kriterijum u formi $A \leq v$, gde je v vrednost koja pripada rangu iz kog atribut A može uzimati vrednosti)
  - Načina podele
    - Binarna podela
    - Višestruka podela

Podela na osnovu diskretnih atributa

- Višestruka podela: broj grana jednak broju različitih vrednosti atributa

- Binarna podela: podela vrednosti u dva podskupa (potrebno naći optimalnu podelu)



Podela na osnovu kontinualnih atributa

- Diskretizacija: formiranje diskretnih atributa
  - Statička – jedna diskretizacija na početku
  - Dinamička – određivanje rangova u toku izgradnje stabla

- Binarna odluka: (A < v) ili (A ≥ v)
  - uzima u obzir sve moguće podele i bira najbolju
  - iti zahtevnije u pogledu izrač ja

(i) Binary split

(ii) Multi-way split

Problem: prenaučenost

- Stablo raste dok se svi podaci pravilno ne klasificikuju.
- Problem:
  - Postoji šum u podacima
  - Skup za treniranje je premali

=> može doći do prenaučenosti (*eng. overfitting*) stabla odluke

- *Definicija*

"Neka je dat prostor hipoteza H. **Hipoteza h iz H je prenaučena** ako postoji hipoteza h' iz H takva da h ima manju grešku nego h' na primerima za učenje, ali h' ima manju grešku nego h na celom prostoru primera"

Kako se rešava problem prenaučenosti? (1)

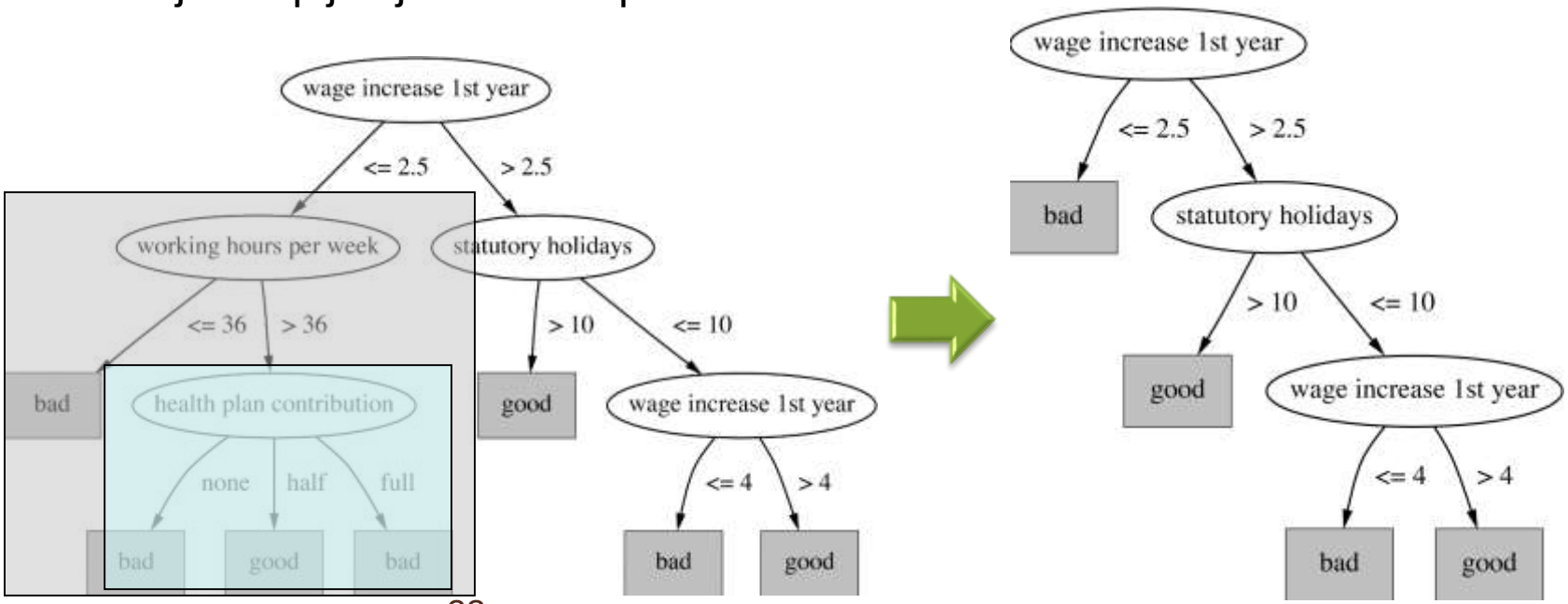
- Pre-odsecanje (pravilo zaustavljanja)
  - Algoritam se zaustavlja pre nego što formira kompletno stablo
  - Uslovi zaustavljanja u čvoru:
    - svi primeri pripadaju istoj klasi
    - vrednosti svih atributa su iste
  - Restriktivniji uslovi:
    - broj primera je manji od unapred definisane granice
    - distribucija klasa među primerima je nezavisna od dostupnih atributa (korišćenjem χ^2 testa)
    - najbolji kriterijum podele nije veći od zadate granice
    - dostignuta je maksimalna dubina stabla

Kako se rešava problem prenaučenosti? (2)

- Post-odsecanje
  - Izgraditi kompletno stablo
  - Odsecati čvorove stabla u *bottom-up* pristupu
  - Dve operacije odsecanja:
    1. Zamena podstabla – podstablo se zamenjuje listom
    2. Uzdizanje podstabla – podstablo se zamenjuje svojim podstablom (pod-podstablom)

ZAMENA PODSTABLA - primer

- Podstablo se zamenjuje listom čija je vrednost klasa koja je najzastupljenija u datom podstablu



Kriterijum za selekciju atributa

- **Informacijska dobit**
  - Pretpostavka da su svi atributi kategorički
  - Može se modifikovati za kontinualne atribute
- Mera dobitka
  - Normalizuje informacijsku dobit
  - Rešava problem što informacijska dobit favorizuje atribute sa mnogo vrednosti
- *Gini* indeks
  - Pretpostavka da su svi atributi kontinualni
  - Pretpostavka da za svaki atribut postoji nekoliko mogućih vrednosti za podelu; može zahtevati dodatne alate (npr. klasterovanje) za određivanje tih vrednosti
  - Može se modifikovati za kategoričke atribute

Entropija

- Mera neizvesnosti proizvoljne promenljive.

$$I = -\sum_{c} p(c) \log_2 p(c)$$

*p(c) – verovatnoća da proizvoljno izabran primer pripada klasi c

- Entropija čvora:

$$I_{res} = -\sum_{v} p(v) \sum_{c} p(c|v) \log_2 p(c|v)$$

*p(v) – verovatnoća da proizvoljno izabrani primer ima v kao vrednost odabranog atributa

*p(c|v) – verovatnoća da primer koji ima v kao vrednost odabranog atributa pripada klasi c

  - Meri homogenost čvora
    - Maksimum – kada su svi primeri jednako distribuirani među svim klasama – nosi najmanje informacija
    - Minimum – kada svi primeri pripadaju jednoj klasi – nosi najviše informacija

Informacijska dobit

- Meri smanjenje entropije postignuto podelom – bira se podela sa najvećom dobiti (najvećim smanjenjem entropije) $Gain(A) = I - I_{res}(A)$

podela po atributu A

- Nedostatak: daje prednost podelama koje kao rezultat imaju veliki broj particija – malih, ali "čistih"

GINI INDEKS

$$Gini = 1 - \sum_{c} p(c)^2$$

*p(c) – frekvencija klase c u nekom čvoru

- Maksimum – kada su svi primeri jednako distribuirani među svim klasama – nosi najmanje informacija
- Minimum – kada svi primeri pripadaju jednoj klasi – nosi najviše informacija

Gini podela

- *Gini* vrednost za podelu po atributu A:

$$Gini(A) = \sum_{v} p(v)\,(1 - \sum_{c} p(c|v)^2)$$

*p(v) – verovatnoća da proizvoljno izabrani primer ima v kao vrednost odabranog atributa

*p(c|v) – verovatnoća da primer koji ima v kao vrednost odabranog atributa pripada klasi c

- Atribut sa najmanjom vrednošću Gini(A) bira se za podelu.

$$GiniGain(A) = Gini - Gini(A)$$

- Ukoliko se koristi vrednost GiniGain(A), bira se atribut sa najvećom vrednošću.

ID3 algoritam (Quinlan, 1986.)

- Informacijska dobit
- Zaustavljanje:
  - Svi primeri pripadaju istoj klasi
  - Najbolja informacijska dobit nije veća od 0
- Ne primenjuje odsecanje

C4.5 algoritam (Quinlan, 1993.)

- Evolucija ID3 algoritma
- Kriterijum podele – mera dobiti
- Zaustavljanje: broj primera za podelu je manji od zadate granice
- Odsecanje : Error-Based

CART algoritam (Brieman et al.,1984.)

- CART = Classification and Regression Trees
- Binarna stabla
- Kriterijum – TWOING
- Odsecanje – Cost-Complexity
- Regresiona stabla – stabla kod kojih listovi predviđaju realan broj, a ne klasu

CHAID algoritam

- Originalno – samo za nominalne atribute (kategorički, redosled nije bitan)
- Za svaki atribut traži parove vrednosti i njihovu razliku, izraženu vrednošću p, dobijenom statističkim testom
- Ukoliko je p manje od zadate granice, vrednosti se spajaju, i traži se novi potencijalni par za spajanje; proces se nastavlja dok trenutka kad nema više značajnih parova
- Nakon toga, bira se najbolji atribut za podelu
- Podela se završava u sledećim sluačajevima:
  - Dostignuta je maksimalna dubina stabla
  - Dostignut je minimalan broj primera u čvoru koji mogu biti roditelji
  - Dostignut je minimalan broj primera koji mogu da budu potomci
- Nema odsecanja

Prednosti

- Stabla odluke “objašnjavaju sama sebe”, ukoliko su kompaktna, lako se prate (ukoliko imaju razuman broj listova, mogu ih razumeti i laici). Takođe, mogu se predstaviti preko skupa pravila, pa se smatraju jako razumljivim.
- Mogu koristiti i nominalne(kategoričke) i numeričke(kontinualne) atribute.
  Mogu raditi sa skupovima podataka koji poseduju greške
- Mogu raditi sa skupovima podataka koji imaju nedostajuće vrednosti
- Smatraju se neparametarskim metodom – stabla odluke nemaju pretpostavke o distribuciji prostora ili strukturi klasifikatora

Mane

- Mnogi algoritmi (poput ID3 i C4.5) zahtevaju altribute sa diskretnim vrednostima
- Pošto koriste “zavadi pa vladaj” metod, pokazuju se dobro kada postoji mali broj visoko relevantnih atributa, ali gore kada postoji mnogo kompleksnih interakcija
- Pohlepna karakteristika vodi do “preosetljivosti” skupa za treniranje na irelevantne atribute i šum.

ID3: Indukovanje stabla odluke

- ID3 je algoritam koji se koristi za indukovanje stabla odluke na osnovu primera tipa:
  - (v_*atribut1*, v_*atribut2*, ..., v_*atributN*, *klasa*)
- Dobijeno stablo odluke se kasnije koristi za klasifikaciju novih uzoraka.
  - (v_*atribut1*, v_*atribut2*, ..., v_*atributN*)
  - *klasa* = ?
- Za nalaženje optimalnog puta za klasifikaciju skupa primera za učenje, potrebno je mimimizovati uslove/pitanja (minimizacija dubine stabla)
- Neophodna je funkcija koja će biti mera za izbor atributa/pitanja koje vrši najbolju (najbolje balansiranu) podelu.
- Takva funkcija je metrika ***Informacijska dobit***.

Entropija – merenje homogenosti skupa učenja

- Da bi definisali pojam informacijska dobit, neophodna nam je definicija entropije.
- Pretpostavka: stablo odluke koje treba da dobijemo klasifikuje instance u dve kateorije: P(positive) and N(negative)
- Za zadati skup S, koji sadrži takve pozitivne i negativne klase, entropija za S u odnosu na takvu Bulovu klasifikaciju je:
- **H(S)= - P(positive)log2P(positive) - P(negative)log2P(negative)**
- **P(positive):** broj pozitivnih primera u S
- **P(negative):** broj negativnih primera u S

Značenje entropije

- Primer:
  - Ako je S (0.5+, 0.5-) tada je H(S) = 1,
  - Ako je S (0.67+, 0.33-) tada je H(S) = 0.92,
  - Ako je P (1+, 0 -) tada je H(S) = 0.
- Vrednost entropije se kreće u intervalu [**0**, **1**].
- Vrednost entropije je **0** kada svi elementi skupa pripadaju jednoj istoj klasi.
- Vrednost entropije je **1** kada svakoj klasi pripada podjednak broj elemenata.
- Treba uočiti da što je uniformnija distribucija verovatnoće, veća je informacija koju nosi.
- Praktično, entropija meri „nečistoću“ u kolekciji skupa za učenje.

Entropija - generalizacija

- Generalno, Entropija skupa **S** se računa kao:
- $$H(S) = \Sigma_{i=1..k}(- p(C_i) * \log_k(p(C_i)))$$
- gde su:
- $p(C_i)$ – procenat elemenata skupa **S** koji pripadaju klasi

  C_i (i = 1,...,k)
- $\log_k$ – je logaritam osnove k

Primer računanja entropije

- Ako je **S** skup koji sadrži **14** primera od kojih su **9** svrstani u klasu **C1**, a ostalih **5** u klasu **C2**, entropija skupa **S** je:
  - $H(S) = - (9/14) * \log_2(9/14) - (5/14) * \log_2(5/14)$

    $= - (9/14) * (\ln(9/14) / \ln2) - (5/14) * (\ln(5/14) / \ln2)$

    $= \mathbf{0,94}$

Dobitak u informacijama – merenje očekivane redukcije u Entropiji

- Da bi minimizirali dubinu stabla, kada obilazimo stablo, treba da izaberemo optimalni atribut za podelu – najbolji izbor je atribut koji ima najveću redukciju entropije.
- Dobitak u informacijama (information gain) definišemo kao očekivanu redukciju entropije koja se odnosi na specifičan atribut.
- Dobitak u informacijama, G(S,A) za atribut A, se računa kao:
  - $\mathbf{G(S, A) = H(S) - \Sigma_{i=1..m}(|S_{Ai}|/|S| * H(S_{Ai}))}$
- gde je:
  - **H(X)** – entropija skupa X
  - **m** – broj različitih vrednosti atributa A
  - $\mathbf{S_{Ai}}$ – podskup skupa S gde atribut A ima vrednost A_i
  - **|X|** - broj elemenata skupa X

ID3 algoritam

1. Ako svi primeri pripadaju istoj klasi:

   Tada: Kreiraj list sa vrednošću koja odgovara toj klasi.

2. U suprotnom:
   a) Nađi atribut sa najvećom dobiti.
   b) Dodaj granu za svaku vrednost tog atributa.
   c) Rasporedi primere u odgovarajuće podskupove.
   d) Za svaki podskup ponovi algoritam.

Primer: „Da li je vreme pogodno za igranje tenisa?”

- Vreme je opisano sledećim atributima:
  - **Izgled vremena**, sa vrednostima:
    - **sunčano, oblačno** i **kiša**.
  - **Temperatura**, sa vrednostima:
    - **toplo, prijatno** i **hladno**.
  - **Vlažnost**, sa vrednostima:
    - **normalna** i **visoka**.
  - **Vetar**, sa vrednostima:
    - **slab** i **jak**.
- Primeri su klasifikovani u dve klase:
  - **pogodno** i **nepogodno** vreme za tenis.

Skup primera za učenje

	Izgled vremena	Temperatura	Vlažnost	Vetar	KLASA
1.	sunčano	toplo	visoka	slab	nepogodno
2.	sunčano	toplo	visoka	jak	pogodno
3.	oblačno	toplo	visoka	slab	pogodno
4.	kiša	prijatno	visoka	slab	nepogodno
5.	kiša	hladno	normalna	slab	nepogodno
6.	kiša	hladno	normalna	jak	nepogodno
7.	oblačno	hladno	normalna	jak	nepogodno
8.	sunčano	prijatno	visoka	slab	pogodno
9.	sunčano	hladno	normalna	slab	pogodno
10.	kiša	prijatno	normalna	slab	nepogodno
11.	sunčano	prijatno	normalna	jak	pogodno
12.	oblačno	prijatno	visoka	jak	pogodno
13.	oblačno	toplo	normalna	slab	pogodno
14.	kiša	prijatno	visoka	jak	nepogodno

Rešenje: računanje entropije

- Računamo ukupnu entropiju skupa primera:
  - Imamo dve klase: pogodno i nepogodno.
  - Imamo ukupno 14 primera.
  - Klasi pogodno pripada 7 primera i klasi nepogodno pripada 7 primera.
- $H = - (7/14) * \log_2(7/14) - (7/14) * \log_2(7/14)$

  $= - \ln(0,5) / \ln(2) = \mathbf{1}$

Za koren stabla odluke bira se atribut koji nosi najviše informacija za ceo skup primera.

Rešenje: računanje inf. dobiti

- Računamo dobit za atribut **izgled vremena**:
  - Broj primera po vrednostima je:
    - **sunčano** -> 5, **oblačno** -> 4, **kiša** -> 5
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **sunčano** -> 4:1, **oblačno** -> 3:1, **kiša** -> 0:5
  - Entropija podskupa za vrednost **sunčano** je:
    - $H = -(4/5) * \log_2(4/5) - (1/5) * \log_2(1/5) =$ **0,72**
  - Entropija podskupa za vrednost **oblačno** je:
    - $H = -(3/4) * \log_2(3/4) - (1/4) * \log_2(1/4) =$ **0,81**
  - Entropija podskupa za vrednost **kiša** je:
    - $H =$ **0**
- $G = 1 - 5/14 * 0,72 - 4/14 * 0,81 - 5/14 * 0 =$ **0,51**

Rešenje: računanje inf. dobiti

- Računamo dobit za atribut **temperatura:**
  - Broj primera po vrednostima je:
    - **toplo** -> 4, **prijatno** -> 6, **hladno** -> 4
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **toplo** -> 3:1, **prijatno** -> 3:3, **hladno** -> 1:3
  - Entropija podskupa za vrednost **toplo** je:
    - $H = -(3/4) * \log_2(3/4) - (1/4) * \log_2(1/4) =$ **0,81**
  - Entropija podskupa za vrednost **prijatno** je:
    - $H =$ **1**
  - Entropija podskupa za vrednost **hladno** je:
    - $H = -(1/4) * \log_2(1/4) - (3/4) * \log_2(3/4) =$ **0,81**
- $G = 1 - 4/14*0{,}81 - 6/14*1 - 4/14*0{,}81 =$ **0,11**

Rešenje: računanje inf. dobiti

- Računamo dobit za atribut **vlažnost:**
  - Broj primera po vrednostima je:
    - **normalna** -> 7, **visoka** -> 7
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **normalna** -> 3:4, **visoka** -> 4:3
  - Entropija podskupa za vrednost **normalna** je:
    - $H = -(3/7) * \log_2(3/7) - (4/7) * \log_2(4/7) =$ **0,98**
  - Entropija podskupa za vrednost **visoka** je:
    - $H = -(4/7) * \log_2(4/7) - (3/7) * \log_2(3/7) =$ **0,98**
- $G = 1 - 7/14 * 0{,}98 - 7/14 * 0{,}98 =$ **0,02**

Rešenje: računanje inf. dobiti

- Računamo dobit za atribut **vetar:**
  - Broj primera po vrednostima je:
    - **slab** -> 8, **jak** -> 6
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **slab** -> 4:4, **jak** -> 3:3
  - Entropija podskupa za vrednost **slab** je:
    - H = **1**
  - Entropija podskupa za vrednost **jak** je:
    - H = **1**
- G = 1 - 8/14 * 1 - 6/14 * 1 = **0**

Rešenje: izbor atributa

- Izračunate dobiti za pojedine atribute su:
  - Izgled vremena -> 0,51
  - Temperatura -> 0,11
  - Vlažnost -> 0,02
  - Vetar -> 0
- Najbolji atribut za klasifikaciju je:
  - **Izgled vremena** jer ima najveću dobit.
  - Pravimo tri nova podskupa za vrednosti atributa **sunčano, oblačno** i **kiša.**

Rešenje: podskup “sunčano”

	Temperatura	Vlažnost	Vetar	KLASA
1.	toplo	visoka	slab	nepogodno
2.	toplo	visoka	jak	pogodno
8.	prijatno	visoka	slab	pogodno
9.	hladno	normalna	slab	pogodno
11.	prijatno	normalna	jak	pogodno

$$H = - (4/5) * \log_2(4/5) – (1/5) * \log_2(1/5) = \mathbf{0,72}$$

Rešenje: podskup “oblačno”

	Temperatura	Vlažnost	Vetar	KLASA
3.	toplo	visoka	slab	pogodno
7.	hladno	normalna	jak	nepogodno
12.	prijatno	visoka	jak	pogodno
13.	toplo	normalna	slab	pogodno

$$H = - (3/4) * \log_2(3/4) – (1/4) * \log_2(1/4) = \mathbf{0,81}$$

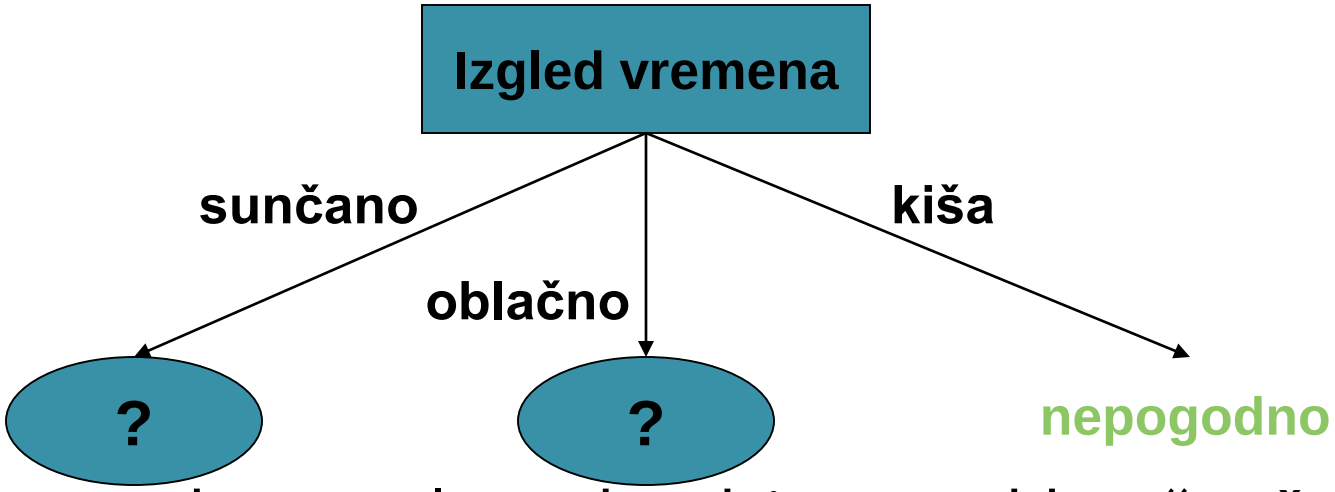
Rešenje: podskup “kiša”

	Temperatura	Vlažnost	Vetar	KLASA
4.	prijatno	visoka	slab	nepogodno
5.	hladno	normalna	slab	nepogodno
6.	hladno	normalna	jak	nepogodno
10.	prijatno	normalna	slab	nepogodno
14.	prijatno	visoka	jak	nepogodno

H = **0**, pošto svi primeri ovog podskupa pripadaju klasi **nepogodno**, pravi se list stabla sa ovom vrednošću.

Rešenje:
stablo odluke, koren stabla

- Nakon prvog prolaza formirano je sledeće stablo odluke:



- U narendnom prolazu obrađuje se podskup “sunčano”, nakon toga podskup „oblačno“ na isti način kao u prvom prolazu.

Rešenje: podskup “sunčano”

	Temperatura	Vlažnost	Vetar	KLASA
1.	toplo	visoka	slab	nepogodno
2.	toplo	visoka	jak	pogodno
8.	prijatno	visoka	slab	pogodno
9.	hladno	normalna	slab	pogodno
11.	prijatno	normalna	jak	pogodno

$$H = - (4/5) * \log_2(4/5) – (1/5) * \log_2(1/5) = \mathbf{0,72}$$

Rešenje:
nastavak kreiranja stabla

- Računamo dobit za atribut **temperatura:**
  - Broj primera po vrednostima je:
    - **toplo** -> 2, **prijatno** -> 2, **hladno** -> 1
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **toplo** -> 1:1, **prijatno** -> 2:0, **hladno** -> 1:0
  - Entropija podskupa za vrednost **toplo** je:
    - H = **1**
  - Entropija podskupa za vrednost **prijatno** je:
    - H = **0**
  - Entropija podskupa za vrednost **hladno** je:
    - H = **0**
- G = 0,72 - 2/5 * 1 = **0,32**

Rešenje: nastavak, inf. dobit

- Računamo dobit za atribut **vlažnost:**
  - Broj primera po vrednostima je:
    - **normalna** -> 2, **visoka** -> 3
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **normalna** -> 2:0, **visoka** -> 2:1
  - Entropija podskupa za vrednost **normalna** je:
    - H = **0**
  - Entropija podskupa za vrednost **visoka** je:
    - $H = -(2/3) * \log_2(2/3) - (1/3) * \log_2(1/3) =$ **0,92**
- G = 0,72 - 3/5 * 0,92 = **0,168**

Rešenje: nastavak, inf. dobit

- Računamo dobit za atribut **vetar:**
  - Broj primera po vrednostima je:
    - **slab** -> 3, **jak** -> 2
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **slab** -> 2:1, **jak** -> 2:0
  - Entropija podskupa za vrednost **slab** je:
    - $H = -(2/3) * \log_2(2/3) - (1/3) * \log_2(1/3) =$ **0,92**
  - Entropija podskupa za vrednost **jak** je:
    - $H =$ **0**
- $G = 0{,}72 - 3/5 * 0{,}92 =$ **0,168**

Rešenje:
izbor atributa u drugom prolazu

- Izračunate dobiti za pojedine atribute su:
  - Temperatura -> 0,32
  - Vlažnost -> 0,168
  - Vetar -> 0,168
- Najbolji atribut za dalju klasifikaciju je:
  - **Temperatura** jer ima najveću dobit.
- Dobijamo tri nova podskupa za vrednosti atributa **toplo**, **prijatno** i **hladno**.

Rešenje: podskupovi dobijeni nakon drugog prolaza

Podskup "sunčano-toplo"
H = 1

	Vlažnost	Vetar	KLASA
1.	visoka	slab	nepogodno
2.	visoka	jak	pogodno

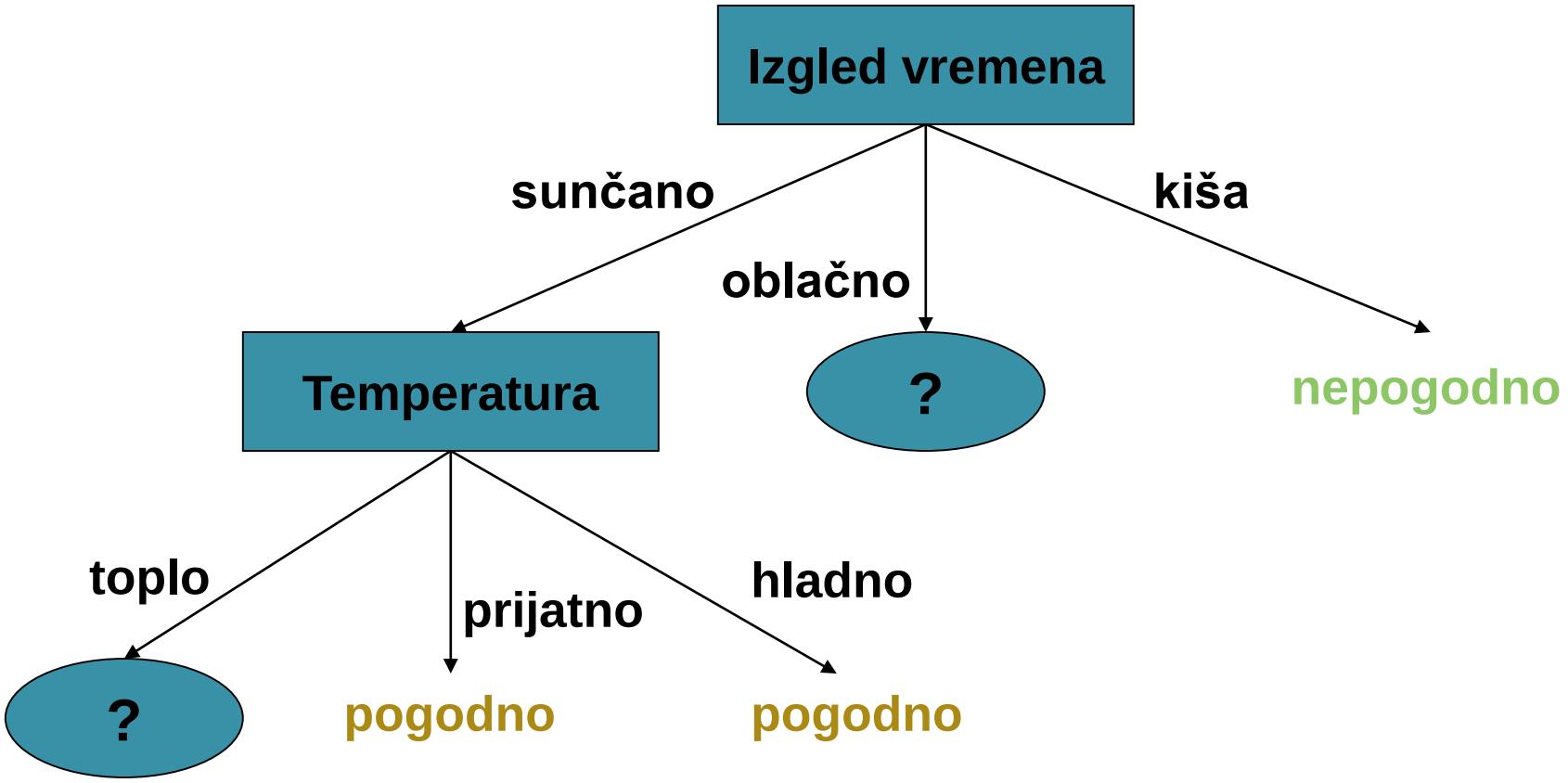
Podskup "sunčano-prijatno"
H = 0

	Vlažnost	Vetar	KLASA
8.	visoka	slab	pogodno
11.	normalna	jak	pogodno

Podskup "sunčano-hladno"
H = 0

	Vlažnost	Vetar	KLASA
9.	normalna	slab	pogodno

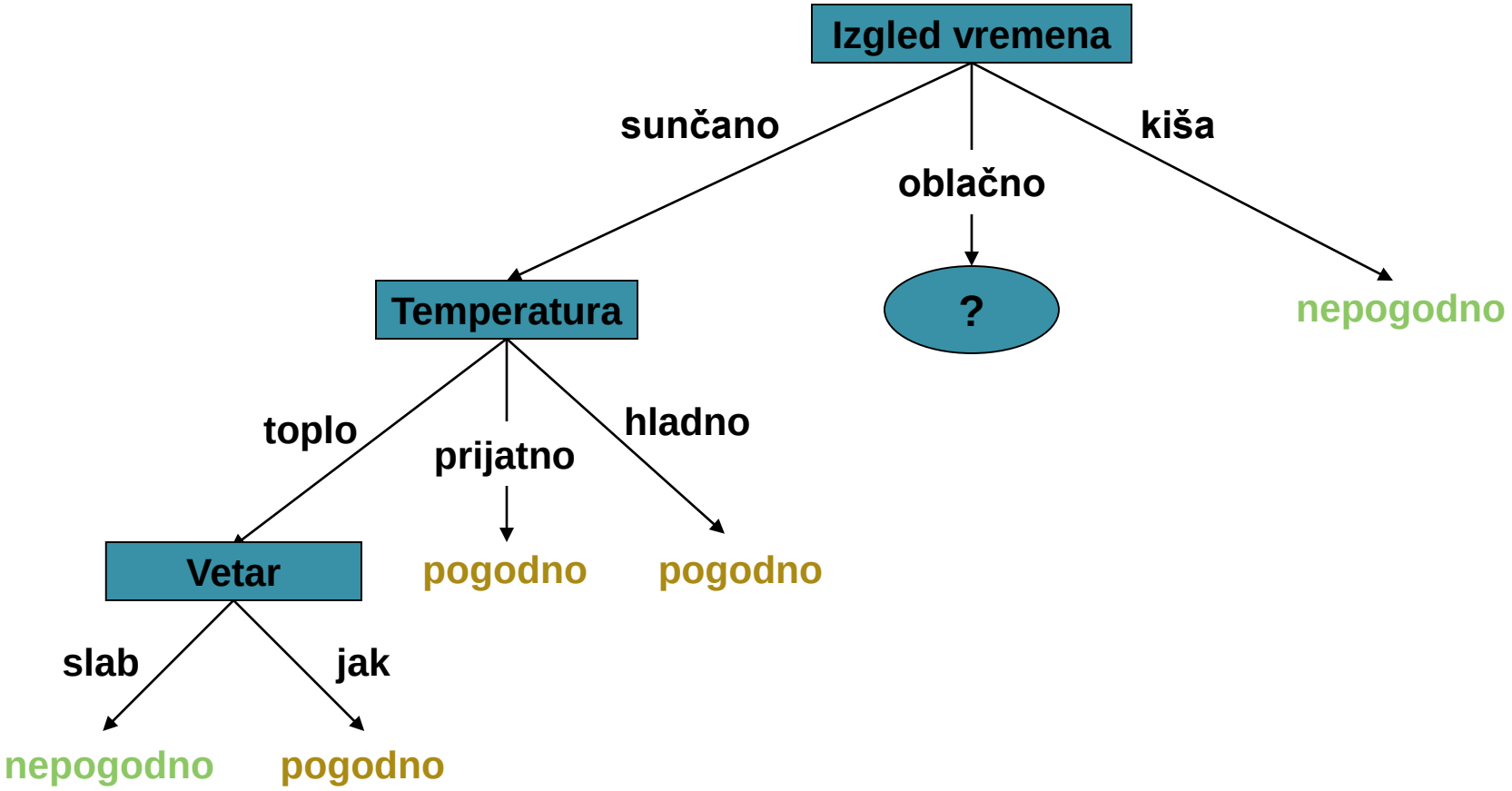
Rešenje: Stablo odluke nakon drugog prolaza („sunčano“)



Rešenje: „sunčano-toplo“

- Posmatramo podskup “sunčano-toplo”.
- Dobitak atributa **vlažnost** je:
  - G = 1 - 1 = **0**
- Dobitak atributa **vetar** je:
  - G = 1 - 0 = **1**
- Znači biramo atribut **vetar:**
  - za vrednost **slab** pravimo list **nepogodno**
  - za vrednost **jak** pravimo list **pogodno**

Rešenje: stablo nakon analize „sunčano-toplo“



Rešenje: podskup “oblačno”

- Razmatramo podskup “oblačno”:

	Temperatura	Vlažnost	Vetar	KLASA
3.	toplo	visoka	slab	pogodno
7.	hladno	normalna	jak	nepogodno
12.	prijatno	visoka	jak	pogodno
13.	toplo	normalna	slab	pogodno

$$H = -(3/4) * \log_2(3/4) - (1/4) * \log_2(1/4) = \mathbf{0,81}$$

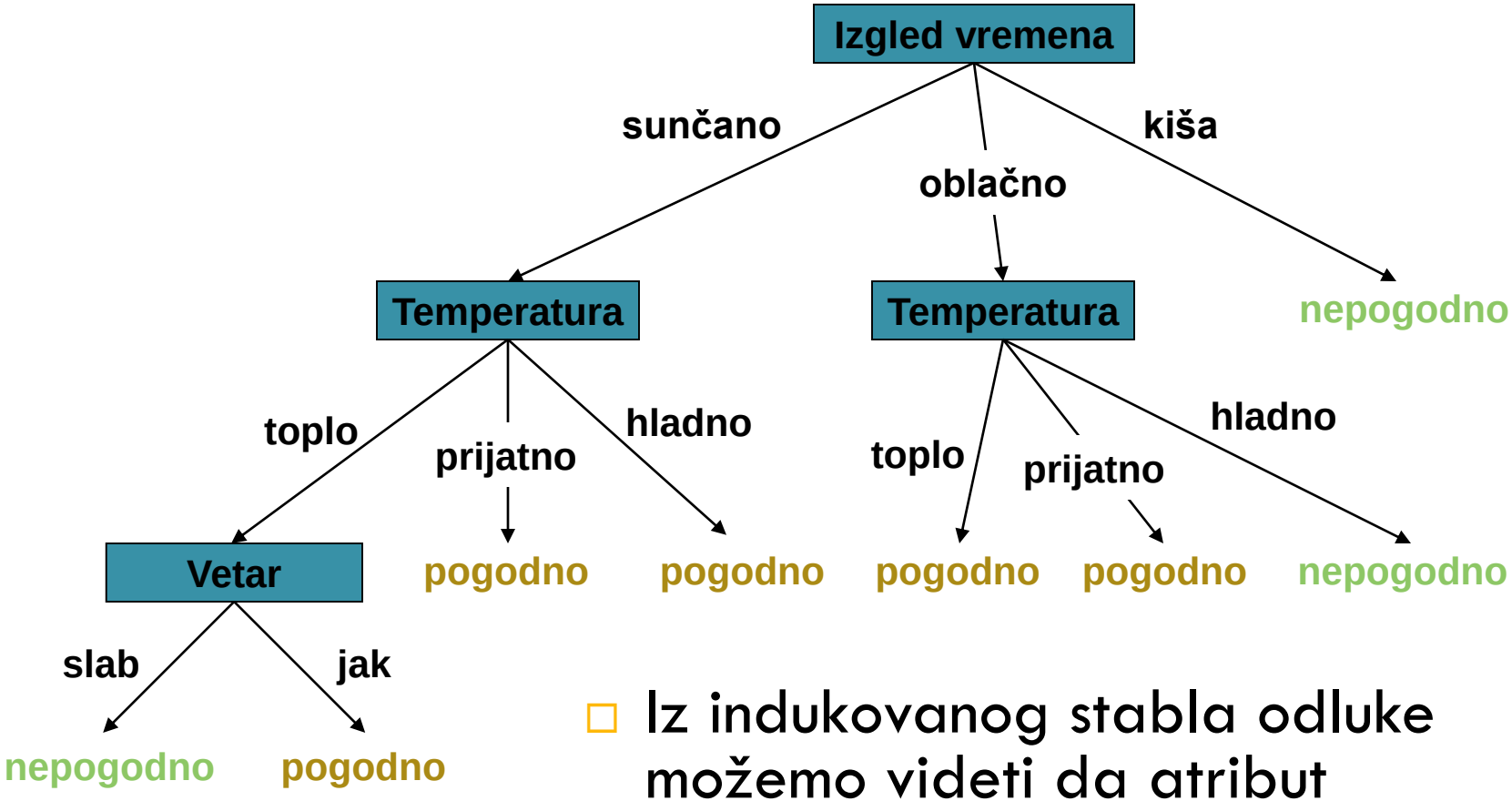
Rešenje: inf. dobit

- Računamo dobit za atribut **temperatura:**
  - Broj primera po vrednostima je:
    - **toplo** -> 2, **prijatno** -> 1, **hladno** -> 1
  - Broj primera po klasama (pogodno:nepogodno) je:
    - **toplo** -> 2:0, **prijatno** -> 1:0, **hladno** -> 0:1
  - Entropija podskupa za vrednost **toplo** je:
    - H = **0**
  - Entropija podskupa za vrednost **prijatno** je:
    - H = **0**
  - Entropija podskupa za vrednost **hladno** je:
    - H = **0**
- G = 0,81 - 0 = **0,81**
  - Ovo je sigurno najveća dobit pa dalje nećemo da računamo.

Rešenje: izbor atributa

- Za vrednost **toplo** pravimo list
  - **pogodno**
- Za vrednost **prijatno** pravimo list
  - **pogodno**
- Za vrednost **hladno** pravimo list
  - **nepogodno**
- **Nakon ovog koraka dobijamo konačno stablo odluke.**

Rešenje: konačno stablo odluke



- Iz indukovanog stabla odluke možemo videti da atribut **vlažnost** uopšte ne utiče na klasifikaciju.

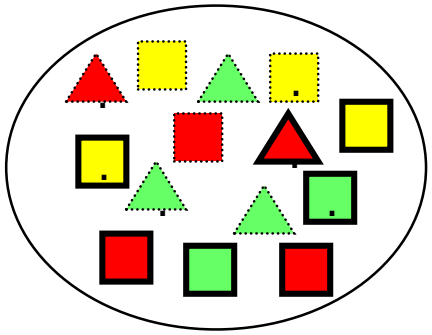
Indukcija pravila

- Stablo traženja se može predstaviti i u obliku pravila:

1. ***IF*** *izgled-vremena = sunčano* ***AND*** *temperatura = toplo* ***AND*** *vetar = slab* ***THEN*** *vreme-za-košarku = nepogodno*
2. ***IF*** *izgled-vremena = sunčano* ***AND*** *temperatura = prijatno* ***THEN*** *vreme-za-košarku = pogodno*
3. ***IF*** *izgled-vremena = sunčano* ***AND*** *temperatura = hladno* ***THEN*** *vreme-za-košarku = pogodno*
4. ***IF*** *izgled-vremena = oblačno* ***AND*** *temperatura = toplo* ***THEN*** *vreme-za-košarku = pogodno*
5. ***IF*** *izgled-vremena = oblačno* ***AND*** *temperatura = prijatno* ***THEN*** *vreme-za-košarku = pogodno*
6. ***IF*** *izgled-vremena = oblačno* ***AND*** *temperatura = hladno* ***THEN*** *vreme-za-košarku = nepogodno*
7. ***IF*** *izgled-vremena = kiša* ***THEN*** *vreme-za-košarku = nepogodno*

ID3 – primer za vežbu

- Atributi:
  - Color
  - Outline
  - Dot
- Klase
  - Square
  - Triangle



Početni elementi rešenja:

$$p(\square) = \frac{9}{14}$$

$$p(\triangle) = \frac{5}{14}$$

Entropija H $= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$

PITANJA?

Dileme?

Komentari?