Stabla odluke

- Upotreba u klasifikaciji ili regresiji
- Obuka
  - Izbor obeležja
  - Izbor skupa pitanja
  - Kombinovanje obeležja
  - Izbor kriterijuma za podelu čvora
  - Izbor kriterijuma za određivanje veličine stabla
- Računarska kompleksnost
- Prednosti i mane
- Ansambalsko učenje na primeru stabala odluke

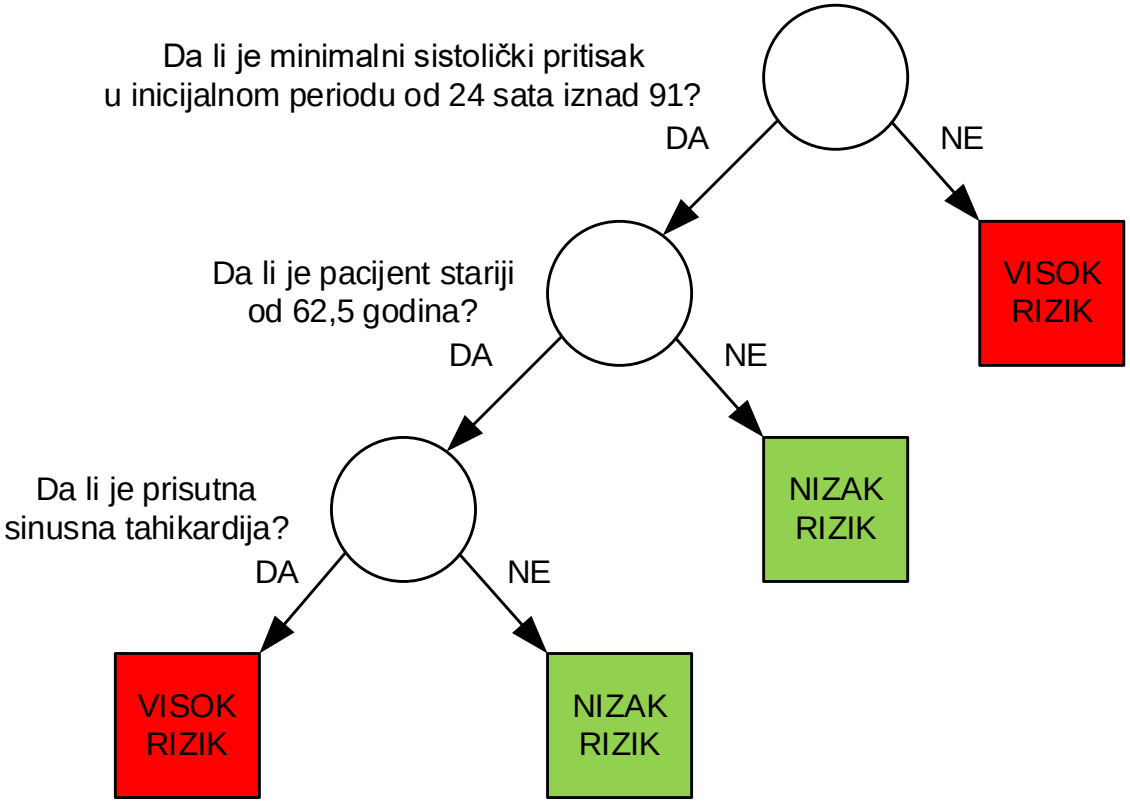
Stabla odluke

- Metoda ranog nadgledanog učenja koja se može koristiti i za **klasifikaciju** i za **regresiju**

Primer formiranog klasifikacionog stabla za predviđanje da li je pacijent visokog rizika od srčanog udara

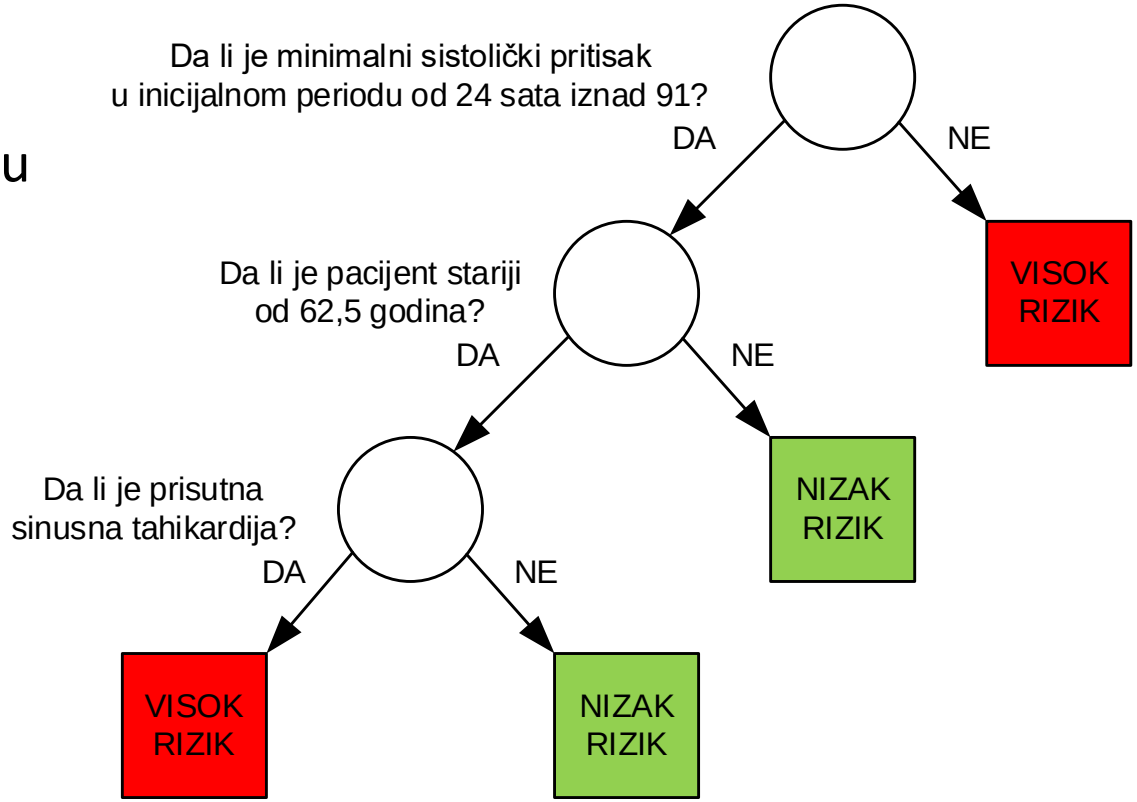
- formiranje stabla predstavlja **obuku**
- obučava se na bazi pacijenata za koje je poznato da li su u narednih godinu dana dana imali srčani udar

U opštem slučaju ova klasifikacija ne mora biti binarna



Stabla odluke

- Obeležja u stablu odluke mogu biti numerička, ali i **kategorička** (da li uzorak pripada određenoj klasi – tih klasa može biti i više od 2)
- Proces obuke zasniva se na izboru pitanja koja u datom momentu na „najlogičniji“ način dele skup uzoraka koji se posmatra
  - Svako pitanje deli skup posmatranih uzoraka na dva podskupa, zavisno od odgovora (DA/NE)



Obuka stabla odluke

- Za obuku je potrebno prvo ustanoviti **standardan skup pitanja**:
  - Za svako **numeričko** obeležje definisati relevantan skup pitanja na osnovu određenih pragova vrednosti – npr. ako pacijenti u bazi za obuku imaju vrednosti godišta između 18 i 90, standardan skup pitanja za ovo obeležje bio bi (npr.):
    - Da li je pacijent stariji od 18.5 godina?
    - Da li je pacijent stariji od 19.5 godina?

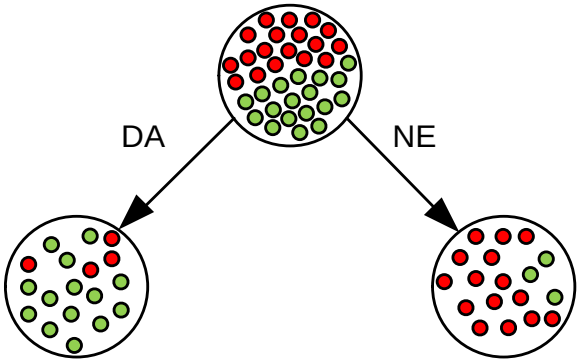
      ...
    - Da li je pacijent stariji od 89.5 godina?
  - Za svako **kategoričko** obeležje definisati skup pitanja koji se zasniva na mogućim participama u skupu kategorija – npr, ukoliko postoje kategorije A, B, C i D, standardan skup pitanja za ovo obeležje mogao bi biti:
    - Da li je kategorija A?
    - Da li je kategorija B?
    - Da li je kategorija C?
    - Da li je kategorija D?
    - Da li je kategorija A ili B?
    - Da li je kategorija A ili C?
    - Da li je kategorija A ili D?

    (Pitanje „da li je kategorija C ili D?“ nije potrebno jer postoji pitanje „da li je A ili B?“)
- U osnovnoj verziji standardnog skupa nema kombinovanja obeležja

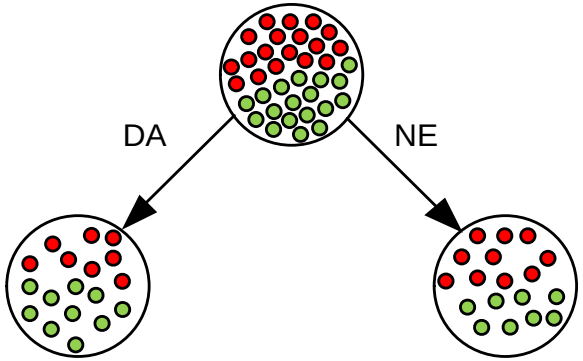
Obuka stabla odluke

- Prvi korak obuke je odrediti pitanje koje će rastaviti početni skup uzoraka (korenski čvor stabla) na dva što „čistija“ podskupa (čvora-potomka)

Da li je minimalni sistolički pritisak u inicijalnom periodu od 24 sata iznad 91?



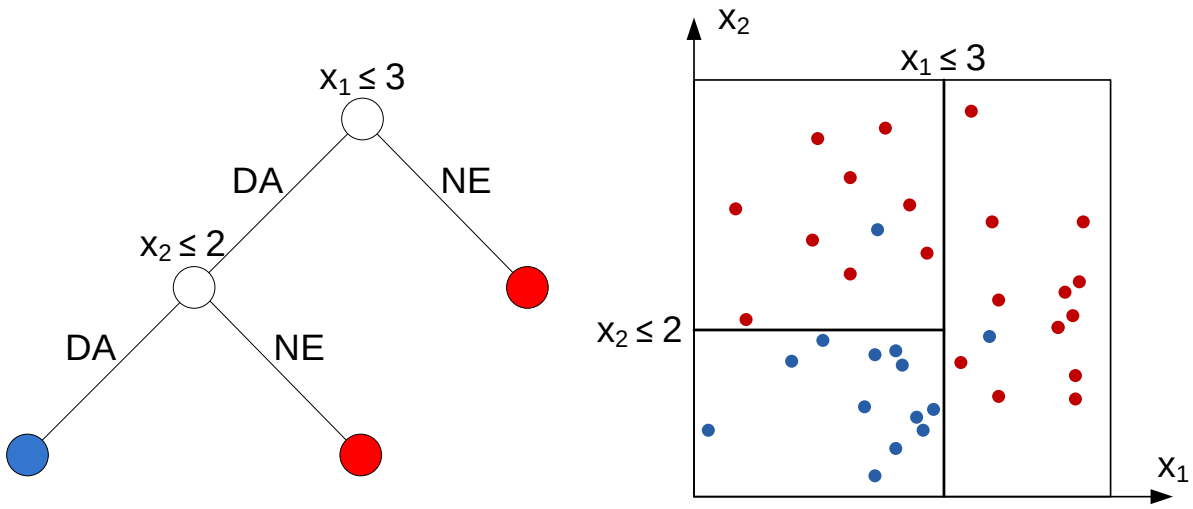
Da li je pacijent ženskog pola?



  - Od ova dva pitanja mnogo korisnije je prvo jer deli uzorke tako da su u čvorovima-potomcima zastupljeni u velikoj meri pripadnici iste klase
- U svakom koraku identifikuje se pitanje koje prema određenoj objektivnoj meri uspešnosti podele najuspešnije deli čvor na dva potomka i formiraju se odgovarajući čvorovi potomci, a zatim se postupak rekurzivno nastavlja

Geometrijska interpretacija obuke

- Obuka zapravo predstavlja podelu prostora na višedimenzionalne pravougaonike (ako su sve promenljive numeričke)



- U naprednijim varijantama, gde je dozvoljeno kombinovanje pojedinih obeležja, linije razgraničenja ne moraju biti paralelne osama x_i
  - Bolje se uočavaju veze između obeležja, dobija se kompaktnije stablo
  - Međutim, u tom slučaju skup pitanja koja treba razmotriti drastično raste, što usporava ionako računarski zahtevnu obuku

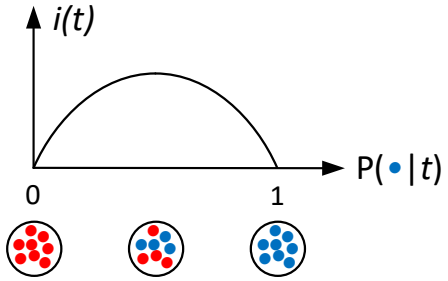
Pitanja na koja treba dati odgovor

- Kako definisati objektivnu meru uspešnosti podele?
  - Usvajaju se mere koje se zasnivaju na proceni smanjenja ukupne nečistoće čvorova potomaka u odnosu na nečistoću čvora čijom su podelom nastali
- Kada prestati s podelama na manje čvorove?
  - Ako se ne prestane na vreme doći će do natprilagođenja
  - Treba prestati kada broj pripadnika čvora padne ispod određenog praga, ili (još bolje), razgranati stablo do samog kraja, a zatim izvršiti „potkresivanje“ (eng. *pruning*) spajanjem najsličnijih čvorova
  - Optimalna veličina stabla može se odrediti procenom na nezavisnom skupu, ali se u praksi češće radi unakrsna validacija
- Kako dodeliti klase (ili izlazne vrednosti) terminalnim čvorovima
  - Prema većinskom principu (a u slučaju regresionih stabala svakom terminalnom čvoru dodeljuje se izlazna vrednost koja je jednaka proseku ili medijanu izlaznih vrednosti uzoraka koji se u tom čvoru našli)

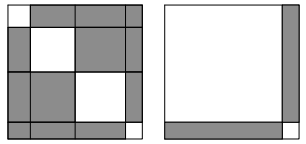
Podela čvora u slučaju klasifikacije

- Mera nečistoće čvora je izmešanost pripadnika raznih klasa u istom čvoru
- Ako ima *K* klasa, meru nečistoće čvora *t* treba uvesti kao funkciju koja zavisi od P(1|*t*), P(2|*t*),... P(*K*|*t*)) i pri tom:
  - ima minimum jednak 0 kada je jedna od njih jednaka 1
  - ima maksimum kad su sve te verovatnoće jednake 1/*K*

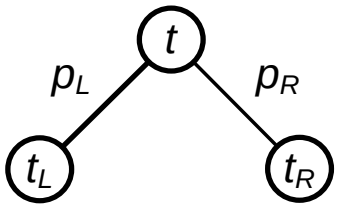
  gde je P(*k*|*t*) zastupljenost pripadnika klase *k* u čvoru *t*



- Konkretno, kao mera nečistoće često se koriste:
  - **Ginijev indeks diverziteta** $i(t) = \sum_{i \neq j} \mathrm{P}(i|t)\mathrm{P}(j|t) = 1 - \sum_j \mathrm{P}^2(j|t)$
  - **Entropija** $i(t) = -\sum_j \mathrm{P}(j|t)\log \mathrm{P}(j|t)$

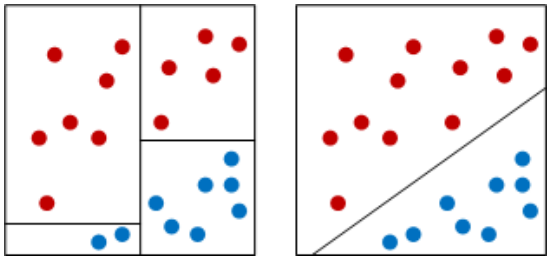


  ali izbor mere iznenađujuće malo utiče na tačnost formiranog stabla!
- Uspešnost podele ogleda se u smanjenju nečistoće
- Ako je pre podele čvor imao meru nečistoće *i*(*t*), a posle podele $p_L i(t_L) + p_R i(t_R)$, pri čemu $p_L + p_R = 1$, bira se pitanje koje maksimizuje smanjenje nečistoće



Računarska kompleksnost

- Obuka je generalno vrlo složena jer za svako moguće pitanje treba izračunati smanjenje nečistoće do kog bi odgovarajuća podela dovela
- Šta ako u čvoru ima veoma velik broj uzoraka?
  - Ne mora se optimalna podela čvora naći na osnovu svih uzoraka, već na osnovu slučajnog podskupa određene veličine
  - Kad se optimalno pitanje nađe, ceo čvor se podeli po tom kriterijumu
- Ako se dozvoli kombinovanje obeležja, kompleksnost obuke se dodatno povećava, ali čak i u tom slučaju postoje efikasni algoritmi za nalaženje potencijalno najkorisnijih kombinacija



  - Primera radi, ako su varijable kategoričke, bilo bi zanimljivo pitati i npr. “Da li A ili (B i ne C)”
  - Obeležja se mogu formirati ili kombinovati i ručno, na osnovu ekspertskog znanja
- Uprkos složenosti obuke, stabla odluke su **ekstremno brza** u fazi eksploatacije (kao klasifikatori ili regresori)

Prednosti i mane stabala odluke

- Jednostavnost – u osnovnoj verziji algoritma treba specificirati samo :
  - Obeležja (time je automatski definisan skup pitanja)
  - Kriterijum za podelu čvora
  - Kriterijum za izbor veličine stabla
- Dobijeni klasifikator/regresor je:
  - Nezavisan od tipa podataka
  - Kompaktan i brz
  - Vrlo lak za interpretaciju
  - Invarijantan na monotonu transformaciju bilo koje koordinate
  - Veoma robustan na besmislena obeležja ili besmislena pitanja
    - Besmisleno pitanje verovatno nikada neće biti identifikovano kao najuspešnije, pa je situacija ista kao i da ono ne postoji
  - Veoma robustan na uzorke koji značajno odstupaju od populacije
- Međutim, dolazi do fragmentacije podataka, što je značajan nedostatak
  - Podaci koji završe u nekom podstablu postaju nevidljivi i irelevantni za ostatak stabla

Slučajna šuma

- Skup stabala odluke kreiranih na podskupovima polaznog skupa za obuku (primer **ansambalskog učenja**, čija je ideja da odluku zajednički donese veći broj jednostavnijih klasifikatora ili regresora)
  - Konačna odluka se, kao i u slučaju pojedinačnog stabla odluke, donosi:
    - većinski, u slučaju klasifikacije
    - računanjem prosečne vrednosti (ili medijana) izlaza, u slučaju regresije
  - Ovo je naročito efikasno ako je skup za obuku veoma velik
  - Podskupovi za obuku kreiraju se pomoću *bootstrap* metode, i obično svaki podskup na kraju sadrži oko 2/3 jedinstvenih uzoraka iz polaznog skupa za obuku (sa ponavljanjem), a neviđeni uzorci čine preostalih oko 1/3
  - Nije dozvoljeno svim stablima da koriste sva obeležja već samo nasumično odabrani podskup polaznih obeležja (s ciljem da se stabla dekorelišu)
- Slučajna šuma predstavlja primer tzv. **bagging** (eng. *bootstrap aggregation*) pristupa, a postoje i **boosting** pristupi, kod kojih se stabla formiraju jedno za drugim, pri čemu svako stablo nastoji da ispravi greške prethodnih