STABLA ODLUČIVANJA

Jelena Jovanovic

jelena.jovanovic@fon.bg.ac.rs

Zahvalnica

Ovi slajdovi su bazirani na materijalima 8. poglavlja knjige “An Introduction to Statistical Learning” (https://www.statlearning.com/)

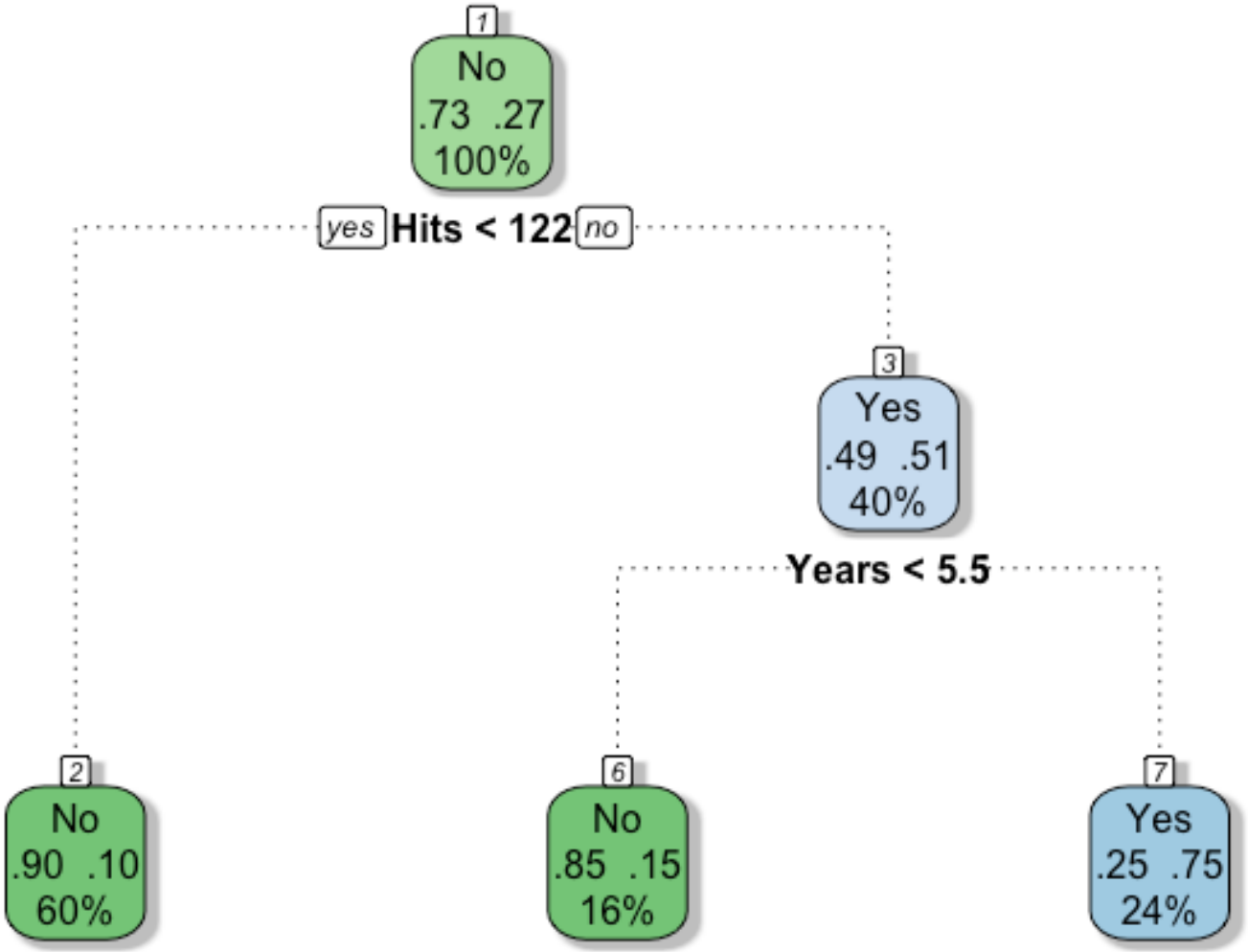
PRIMER: KLASIFIKACIJA IGRAČA BEJZBOLA

Potrebno je kreirati prediktivni model koji će za igrače bejzbola predvideti da li će biti jako dobro plaćeni ili ne (WellPaid), na osnovu

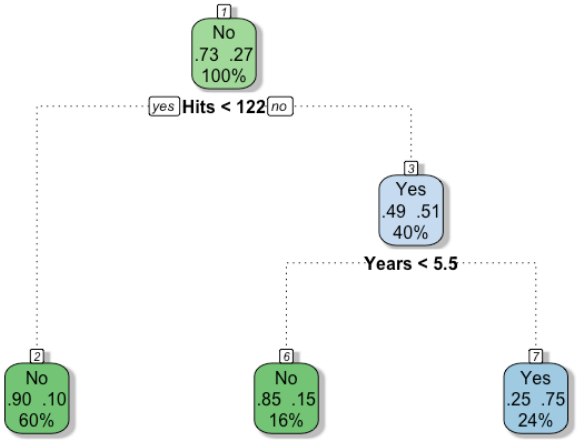
- broja ostvarenih poena u prethodnoj godini (Hits) i
- broja godina koje je igrač proveo u glavnoj ligi (Years)

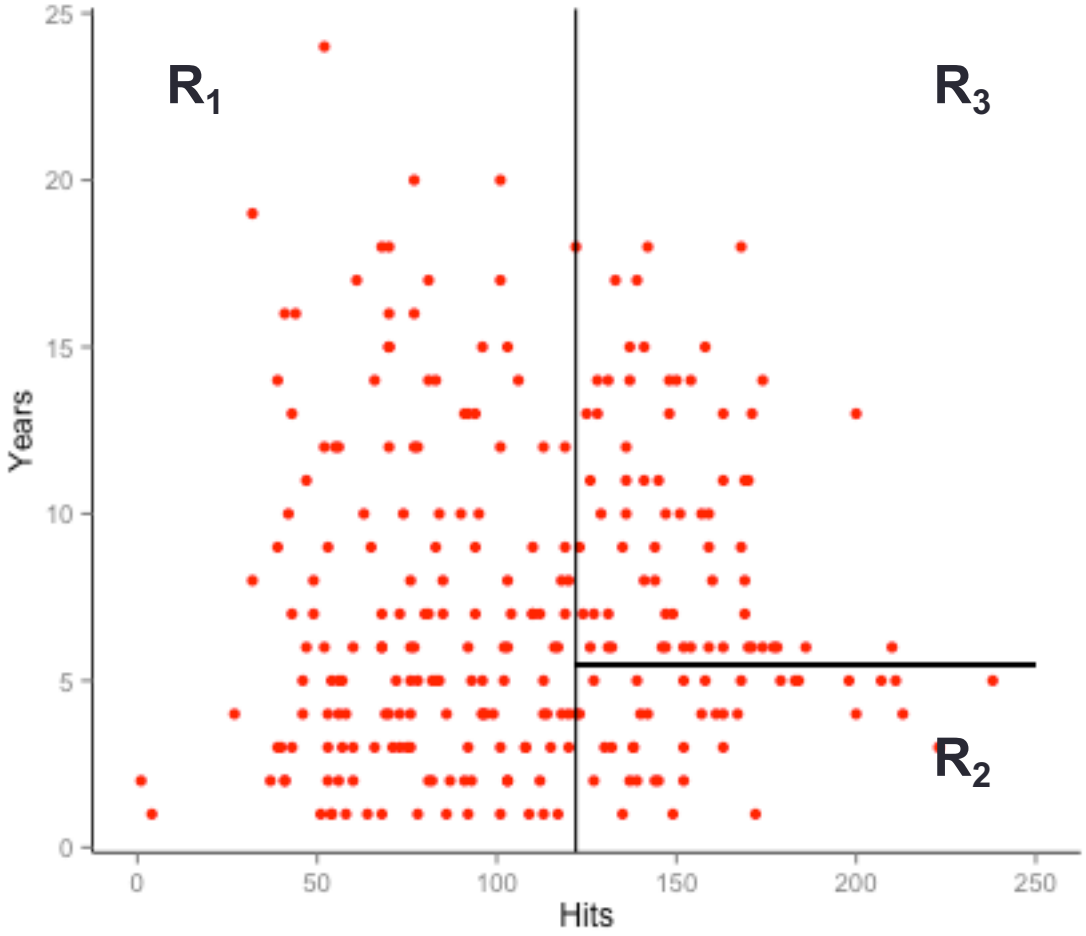
```
> str(hitters.subset)
'data.frame':   263 obs. of  3 variables:
 $ Hits    : int  81 130 141 87 169 37 73 81 92 159 ...
 $ Years   : int  14 3 11 2 11 2 3 2 13 10 ...
 $ WellPaid: Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 2 1 ...
> head(hitters.subset)
                Hits Years WellPaid
-Alan Ashby       81    14       No
-Alvin Davis     130     3       No
-Andre Dawson    141    11       No
-Andres Galarraga  87     2       No
-Alfredo Griffin  169    11      Yes
-Al Newman        37     2       No
>
```

PRIMER: KLASIFIKACIJA IGRAČA BEJZBOLA



DRUGAČIJI POGLED NA STABLA ODLUČIVANJA…

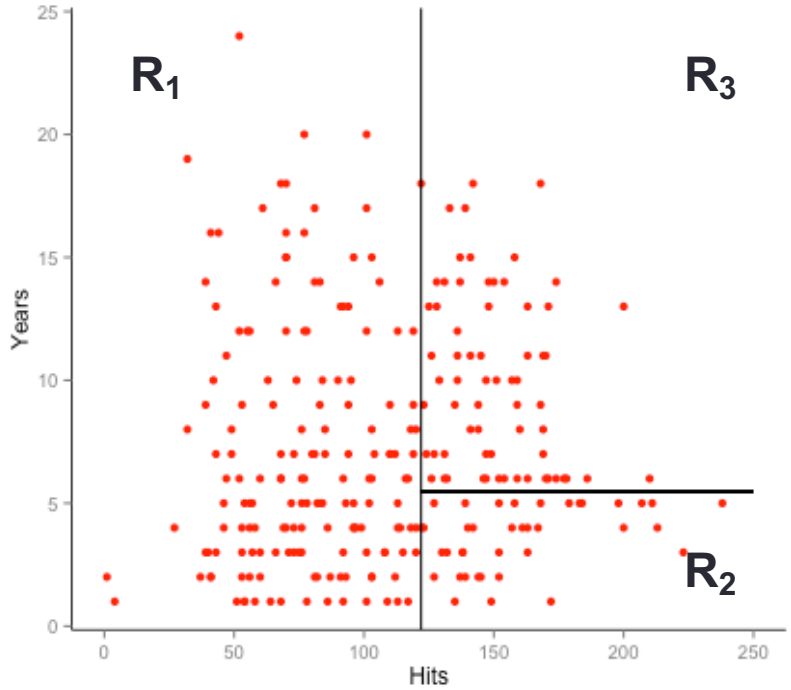




OSNOVNA IDEJA KLASIFIKACIONIH STABALA

Postaviti *prostor atributa* kao *p*-dimenzionalni prostor koga čine moguće vrednosti *p* atributa (x_1,x_2,…,x_p) kojima su instance opisane

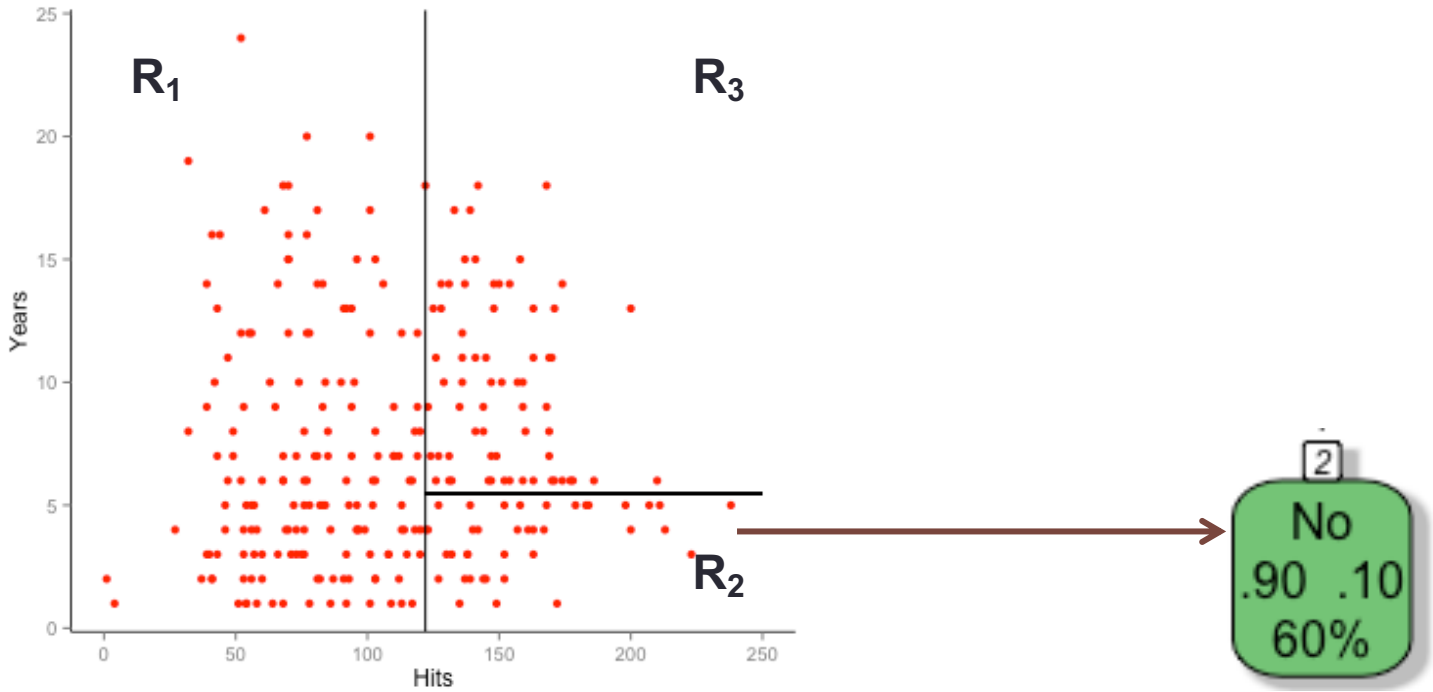
Podeliti *prostor atributa* u više međusobno neprekloplјenih regiona R_1, R_2, …, R_n na način koji optimizuje vrednost izabrane mere kvaliteta klasifikacije



OSNOVNA IDEJA KLASIFIKACIONIH STABALA

Za novu instancu X, određuje se pripadnost jednom od regiona $R_1 \ldots R_n$ na osnovu vrednosti atributa $(x_1, x_2, \ldots, x_p)$ kojima je X opisana

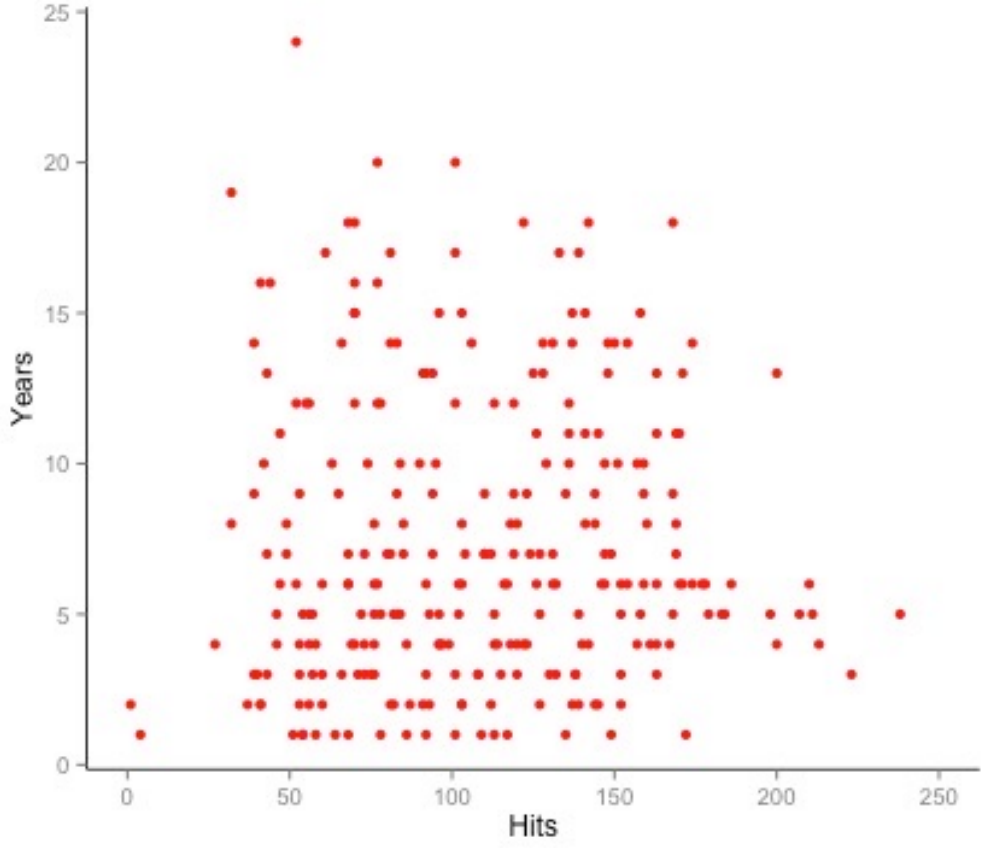
Klasa nove instance će biti ona klasa koja dominira (*majority class*) u regionu R_j u koji je X svrstana



PODELA PROSTORA ATRIBUTA

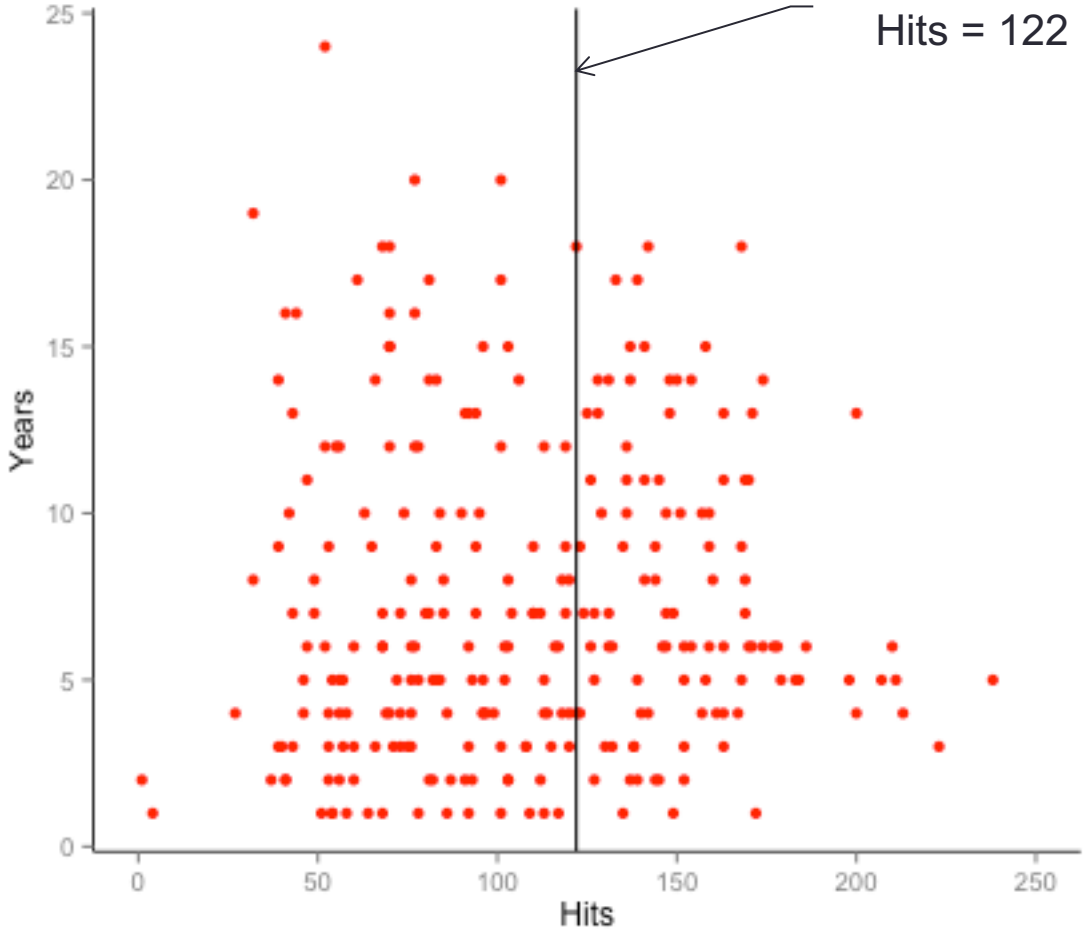
Podela prostora atributa na regione R_j je iterativni proces koji se sastoji od:

- izbora atributa x_i koji će biti osnova za podelu
- izbora vrednosti atributa x_i koja će poslužiti kao ‘granična’ vrednost



PODELA PROSTORA ATRIBUTA

Za prvu podelu, u datom primeru, izabran je atribut Hits, i vrednost 122

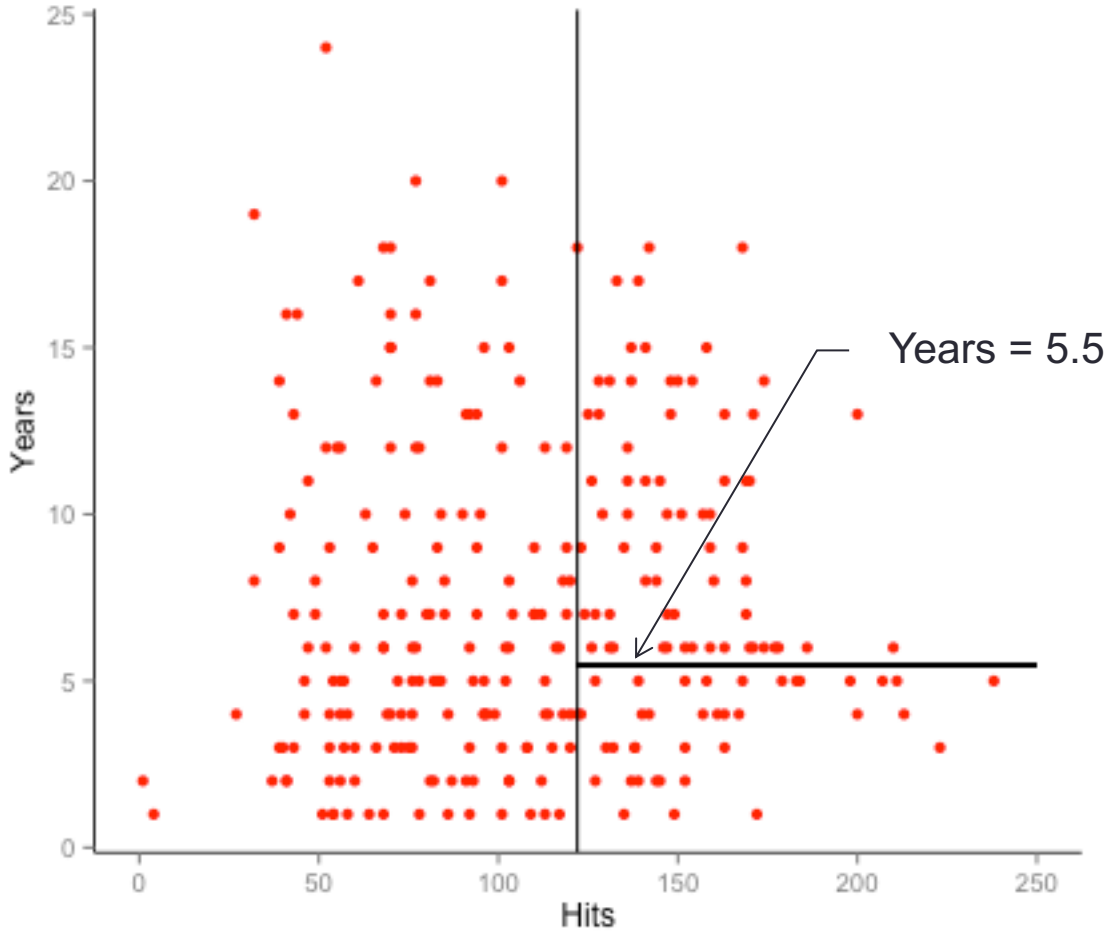


PODELA PROSTORA ATRIBUTA

Prva podela: Hits = 122

Ukoliko je Hits > 122, sledeća podela je na atributu Years:

Years= 5.5



PODELA PROSTORA ATRIBUTA

Pitanja koja se prirodno nameću:

Kako i gde izvršiti podelu? Kako kreiramo regione R_1, R_2,…,R_n?

- Kako biramo atribute na osnovu kojih se vrši podela?
- Kako određujemo vrednosti atributa za formiranje uslova podele?
- Kad / kako se proces podele zaustavlja?

KAKO I GDE IZVRŠITI PODELU?

Primenom ***rekurzivne, binarne podele*** (*recursive binary splitting*) prostora atributa

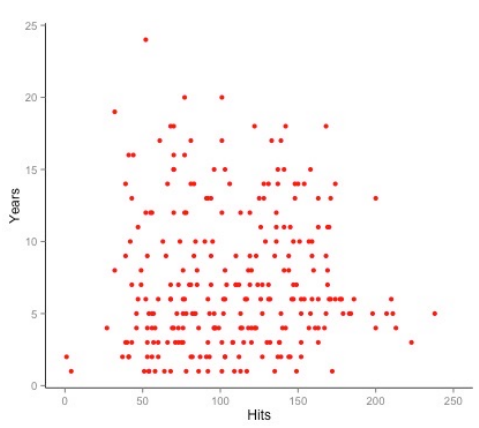
Osnovne karakteristike ovog pristupa:

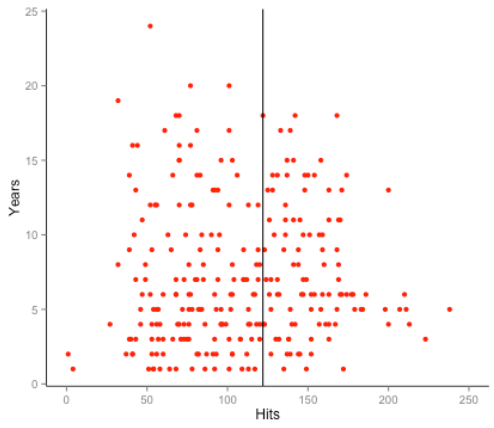
- *top-down* pristup
- *greedy* pristup

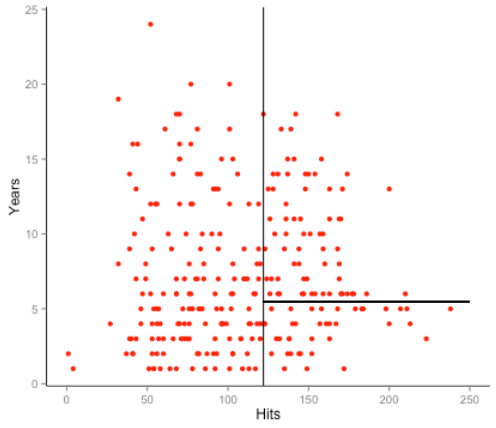
REKURZIVNA, BINARNA PODELA

Top-down pristup

Kreće od vrha stabla, gde sve (trening) instance pripadaju jednoj (zajedničkoj) regiji, a zatim sukcesivno deli prostor atributa na regione







Rekurzivna, binarna podela

Greedy pristup

Pri svakom koraku, najbolja podela se određuje *samo* na osnovu stanja u tom koraku, odnosno bira se podela koja optimizuje meru kvaliteta čvorova (tj regiona) nastalih iz datog čvora (regiona)

Ne uzima se u obzir šta će biti u narednim koracima, tj koja bi podela, u narednim koracima, mogla dovesti do sveukupno bolje klasifikacije

REKURZIVNA, BINARNA PODELA

Algoritam razmatra svaki atribut x_j (j=1,p) i

različite moguće vrednosti s_j tog atributa, i

bira onu [atribut (x_j) – vrednost (s_j)] kombinaciju koja će

podeliti prostor atributa, odnosno tekući region tog prostora,

u dva regiona koji optimizuju izabranu metriku kvaliteta klasifikacije

REKURZIVNA, BINARNA PODELA

Koje vrednosti atributa x_j se razmatraju za definisanje kriterijuma podele?

- Ako je atribut x_j nominalni, bilo koja od vrednosti atributa s_j može poslužiti za grananje oblika $\{X|x_j = s_j\}$ i $\{X|x_j \; ! = \; s_j\}$
- Ako je atribut x_j numerički, njegove vrednosti se prvo sortiraju, a zatim se za kandidate za grananje uzimaju proseci (s_j) svake dva susedne vrednosti; grananje je oblika $\{X|x_j \geq s_j\}$ i $\{X|x_j < s_j\}$

REKURZIVNA, BINARNA PODELA:

METRIKE KVALITETA KLASIFIKACIJE

- Stopa greške pri klasifikaciji (*Classification Error Rate*)
- Gini index
- Cross-entropy

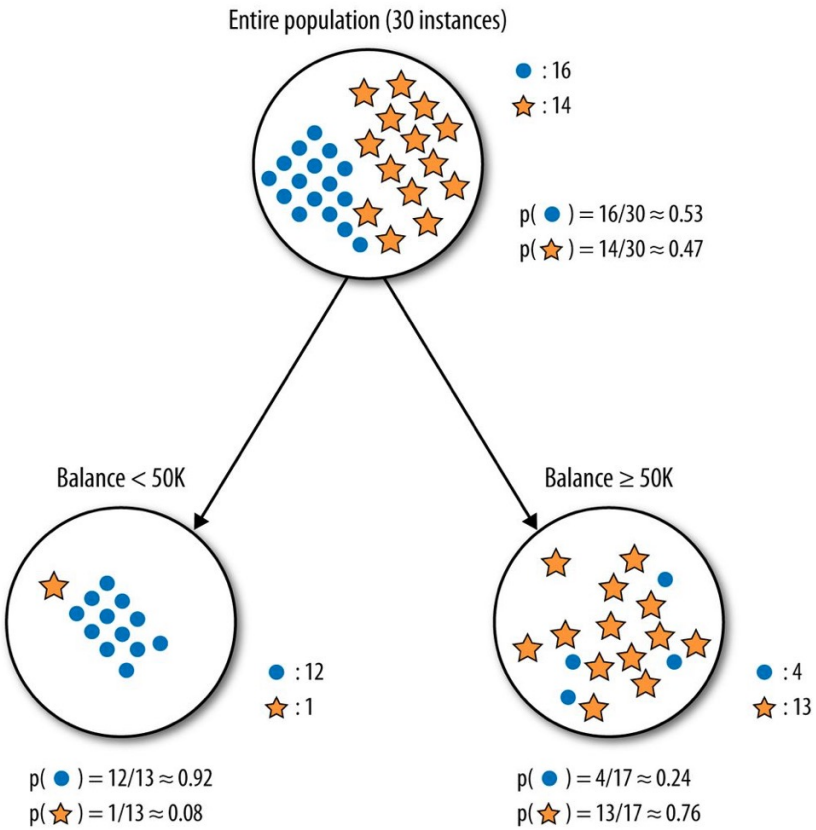
REKURZIVNA, BINARNA PODELA:

METRIKE KVALITETA KLASIFIKACIJE

Sve metrike se baziraju na proporciji (trening) instanci u regionu R_i koje pripadaju klasi k: $\hat{p}_{ik}$

REKURZIVNA, BINARNA PODELA:
METRIKE KVALITETA KLASIFIKACIJE

Sve metrike se baziraju na proporciji (trening) instanci u regionu R_i koje pripadaju klasi k: $\hat{p}_{ik}$



Izvor slike:
https://ekamperi.github.io/images/decision_trees/pure_vs_impure_node.png

STOPA GREŠKE PRI KLASIFIKACIJI

Proporcija instanci (iz skupa za trening) u regionu R_i koje ne pripadaju dominantnoj klasi tog regiona

$$CER = 1 - \max_k \hat{p}_{ik}$$

$\hat{p}_{ik}$ - proporcija (trening) instanci u regionu *i* koje pripadaju klasi *k*

CER – Classification Error Rate

GINI INDEX

$$G = \sum_{k=1}^{K} \hat{p}_{ik} (1 - \hat{p}_{ik})$$

$\hat{p}_{ik}$ - proporcija trening instanci u regionu R_i koje pripadaju klasi k

Opisuje se kao mera ‘nečistoće’ (*impurity)* čvora / regiona

- ‘čisti’ čvorovi su oni u kojima veliki procenat instanci pripada istoj klasi
- mala vrednost za Gini indeks ukazuje na ‘čiste’ čvorove

CROSS-ENTROPY

$$D = -\sum_{k=1}^{K} \hat{p}_{ik} \, log \hat{p}_{ik}$$

$\hat{p}_{ik}$ - proporcija trening instanci u regionu R_i koje pripadaju klasi k

Cross-entropy takođe predstavlja meru ‘nečistoće’ ili (ne)homogenosti čvora => što je vrednost manja, to je čvor ‘čistiji’

REKURZIVNA, BINARNA PODELA:

METRIKE KVALITETA KLASIFIKACIJE

Pri ispitivanju mogućnosti podele regiona R_x na podregione R_{x1} i R_{x2}:

- Izabrana metrika (npr Gini index - GI) se najpre računa za svaki od podregiona R_{x1} i R_{x2} pojedinačno, a zatim
- Računa se ukupna mera kvaliteta (npr. ukupan GI), kao ponderisan prosek vrednosti metrike (npr. GI) na podregionima
  - Ponderi su određeni brojem instanci u podregionu

REKURZIVNA, BINARNA PODELA

Pri ispitivanju mogućnosti podele regiona R_x na podregione R_{x1} i R_{x2}:

- Izabrana metrika (npr Gini index - GI) se najpre računa za svaki od podregiona R_{x1} i R_{x2} pojedinačno,
- Zatim se računa ukupna mera kvaliteta (npr. ukupan GI), kao ponderisan prosek vrednosti metrike (npr. GI) na podregionima
  - Ponderi su određeni brojem instanci u podregionu

Postupak se ponavlja za sve atribute i njihove moguće granične vrednosti, i za podelu se bira [atribut – vrednost] sa najvećom ukupnom merom kvaliteta

OREZIVANJE STABLA (*TREE PRUNING*)

Velika klasifikaciona stabla, tj. stabla sa velikim brojem terminalnih čvorova (listova), imaju tendenciju over-fitting-a (tj. prevelikog uklapanja sa trening podacima)

Ovaj problem se može rešiti ‘orezivanjem’ stabla, odnosno odsecanjem nekih terminalnih čvorova

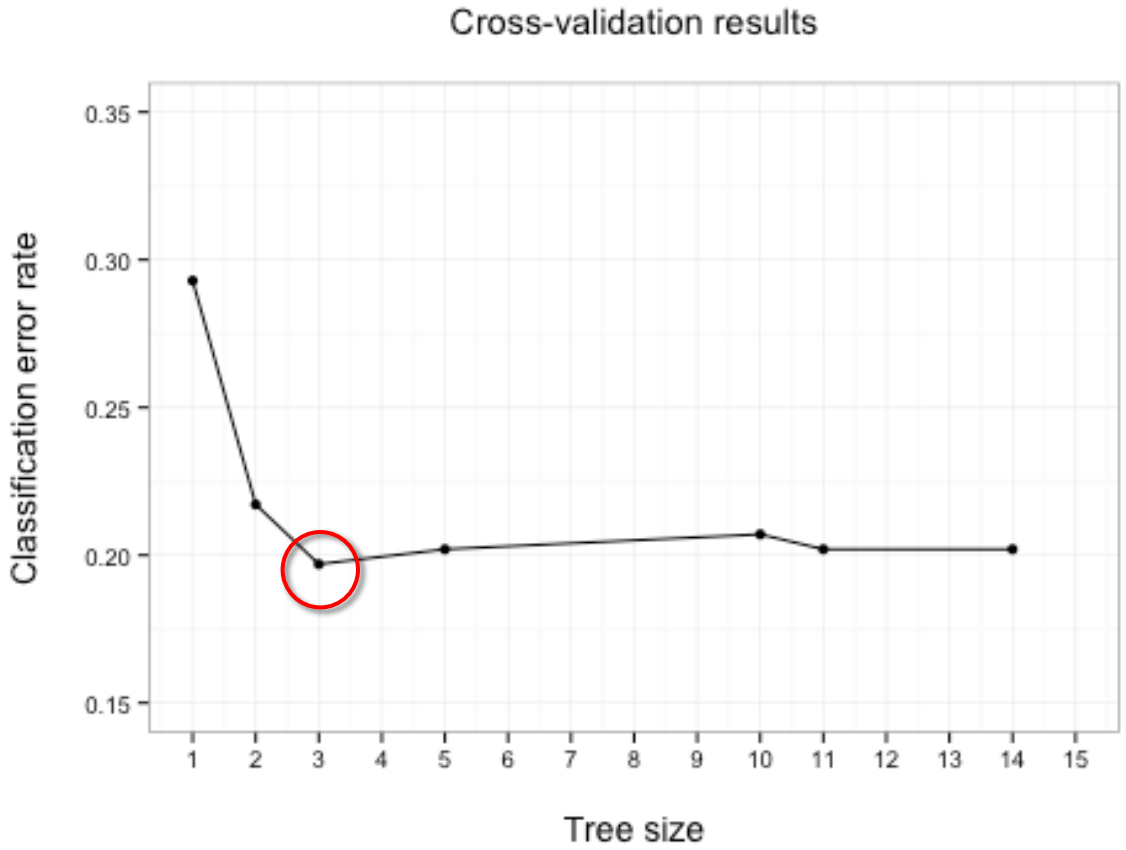
OREZIVANJE STABLA (*TREE PRUNING*)

Kako ćemo znati na koji način i u kojoj meri treba da ‘orežemo’ stablo?

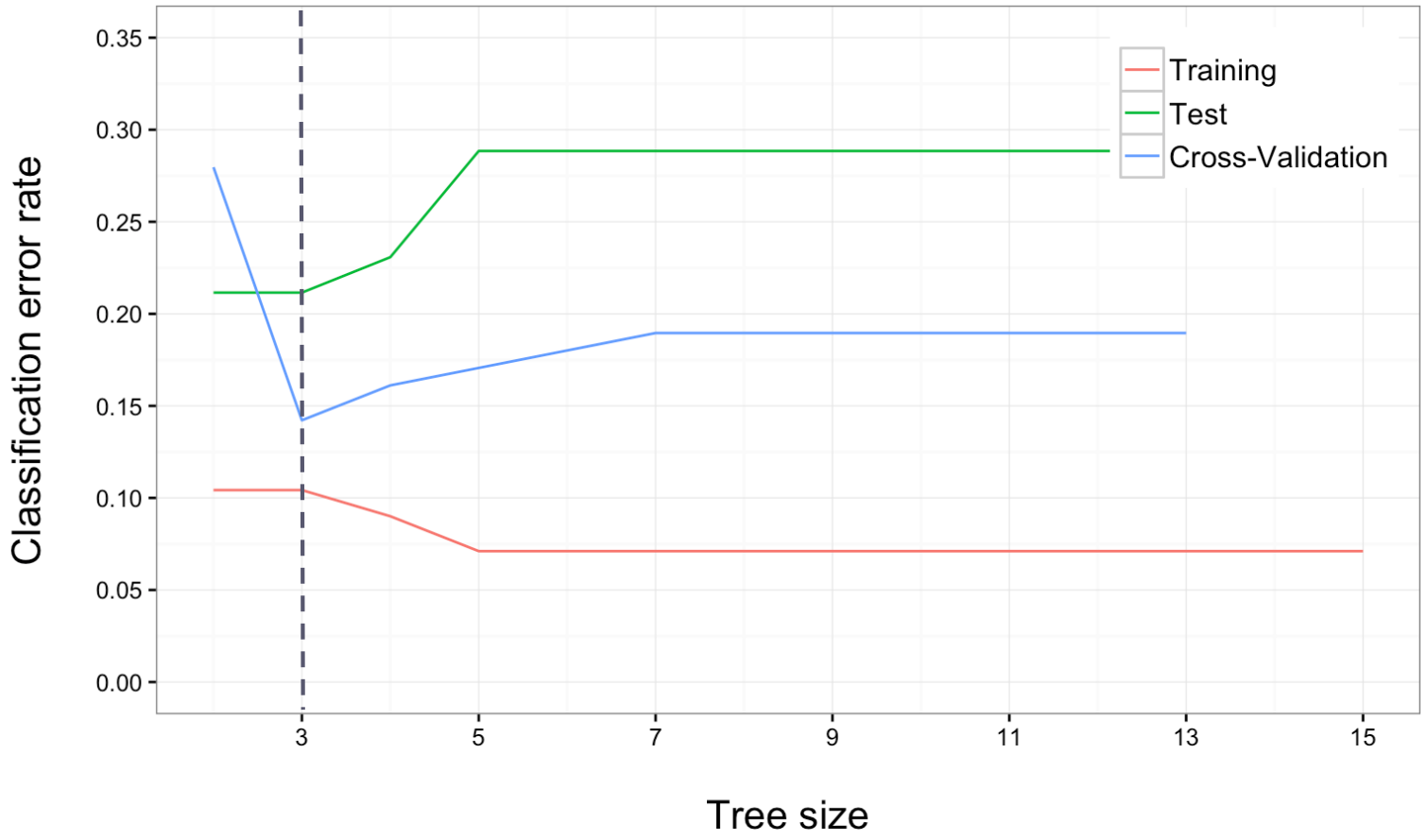
Preporuka je primenom kros validacije (*cross validation)* utvrditi grešku pri klasifikaciji za podstabla različite veličine (tj broja terminalnih čvorova) i izabrati podstablo koje daje najmanju grešku

OREZIVANJE STABLA KROZ KROS VALIDACIJU

U primeru klasifikacije igrača bejzbola, kros validacija pokazuje da se najmanja greška klasifikacije postiže u slučaju stabla veličine 3 (tj. stabla sa 3 terminalna čvora)



OREZIVANJE STABLA KROZ KROS VALIDACIJU



Grafikon potvrđuje da veličina stabla utvrđena kros validacijom (n=3), vodi minimalnoj grešci na test setu

PREDNOSTI I NEDOSTACI STABALA ODLUČIVANJA

- Prednosti:
  - Mogu se grafički predstaviti i jednostavno interpretirati
  - Mogu se primeniti kako na klasifikacione, tako i regresione probleme
  - Vrlo su fleksibilna po pitanju tipa atributa
- Nedostaci:
  - Daju slabije rezultate (manje tačne predikcije) nego drugi pristupi nadgledanog m. učenja