Klasifikacija – Stabla odlučivanja

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

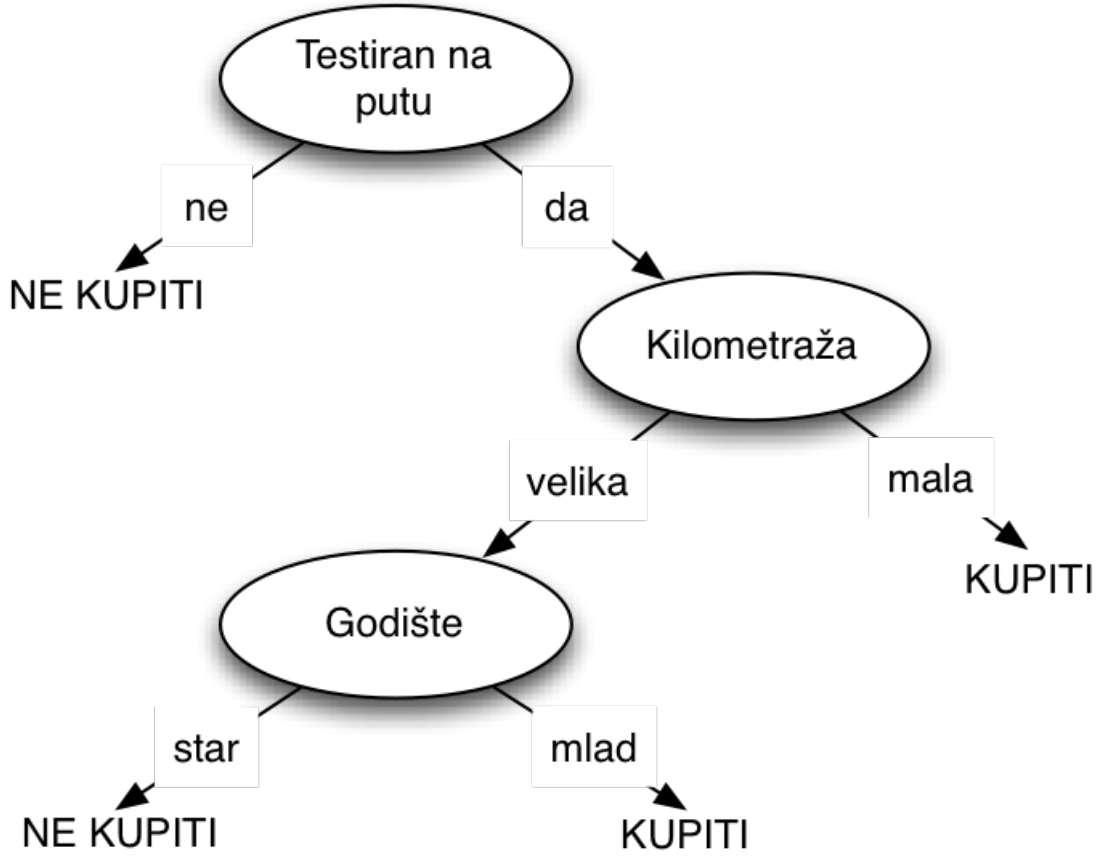
URL: http://nikola.milikic.info

Šta je klasifikacija?

- Zadatak određivanja klase kojoj neka instanca pripada
  - instanca je opisana vrednošću atributa;
  - skup mogućih klasa je poznat i dat

Stabla odlučivanja

Primer: Kupovina automobila



ID3 algoritam

- ID3 - Iterative Dichotomiser 3
- Jedan od najpoznatijih algoritama za generisanje stabla odlučivanja na osnovu skupa primera (dataseta)
- Rezultujuće stablo moguće je koristiti za klasifikovanje budućih (nepoznatih) instanci

Primer – Predviđanje da li će se predstava odigrati

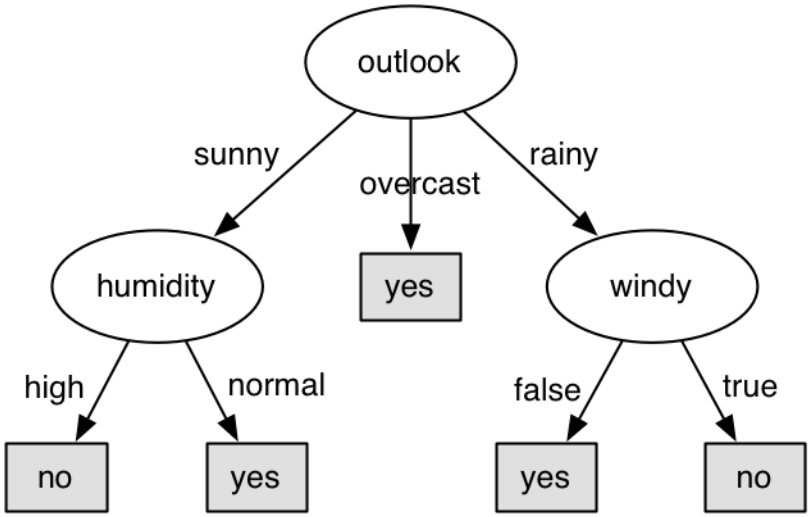
ToPlayOtNotToPlay.arff dataset

Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

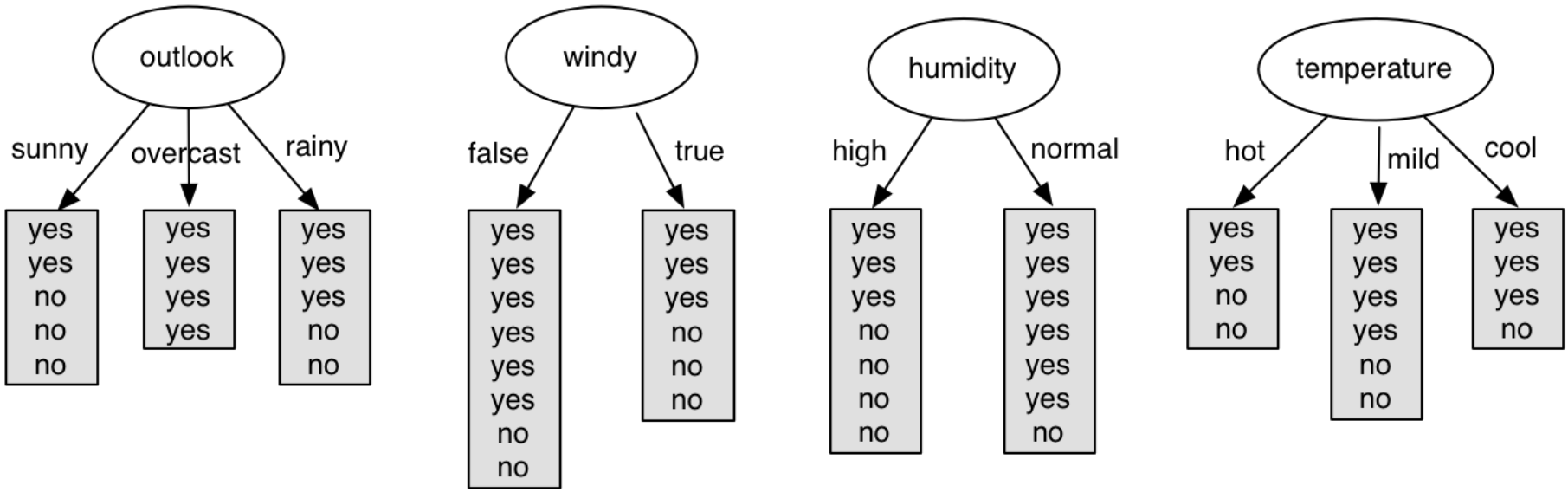
Top-down pristup

Rekurzivno podeli i vladaj:

- Odabrati atribut za čvor stabla
  - Kreirati granu za svaku vrednost atributa
- Podeliti instance na podskupove
  - Po jedan podskup za svaku vrednost atributa
- Ponavljati rekurzivno za svaku granu
  - koristeći samo instance date grane
- Stati
  - kada sve instance u datoj grani imaju istu klasu (čist čvor)



Koji atribut odabrati?



Koji atribut odabrati?

- **Cilj**: kreirati najmanje stablo
- **Teorija informacija**: meriti informacije u bitovima. Tvorac je Klod Šenon, američki matematičar i naučnik 1916 - 2001
- Entropija H(S) se računa prema formuli:

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

gde je:

- S – skup svih instanci u datasetu
- N – broj instanci u datasetu
- p_i – verovatnoća događaja

Entropija celog dataseta

- Od ukupno 14 instanci imamo:
  - 9 instanci “yes”
  - 5 instanci “no”

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

Informaciona dobit

- Informaciona dobit Gain(A, S) atributa A nad skupom instanci S predstavlja količinu informacija koja se dobija poznavanjem vrednosti atributa A. Informaciona dobit predstavlja razliku entropije pre grananja i entropije nakon grananja nad atributom A.

Informaciona dobit

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j) = H(S) - H(A,S)$$

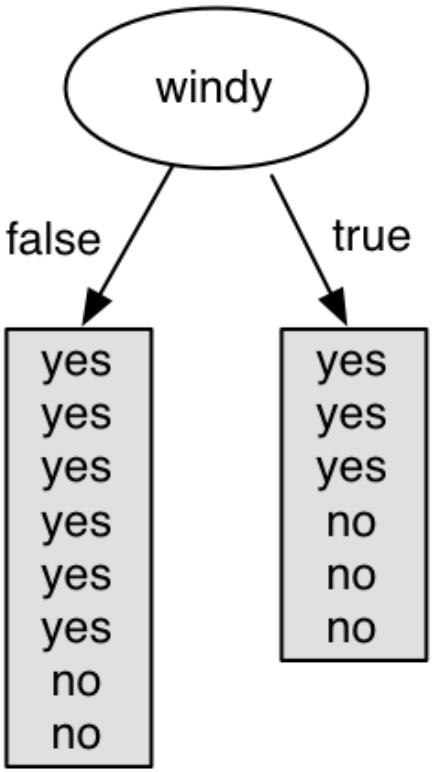
gde je:

- H(S) – entropija celog skupa instanci S
- $|S_j|$ – broj instanci sa j-tom vrednošću atributa A
- $|S|$ – ukupan broj instanci u skupu S
- v – skup vrednosti atributa A
- $H(S_j)$ – Entropija podskupa instanci sa atributom A
- H(A, S) – Entropija atributa A

Biramo atribut sa najvećom informacionom dobiti

Informaciona dobit atributa “windy”

- Od ukupno 14 instanci imamo:
  - 6 instanci “true”
  - 8 instanci “false”

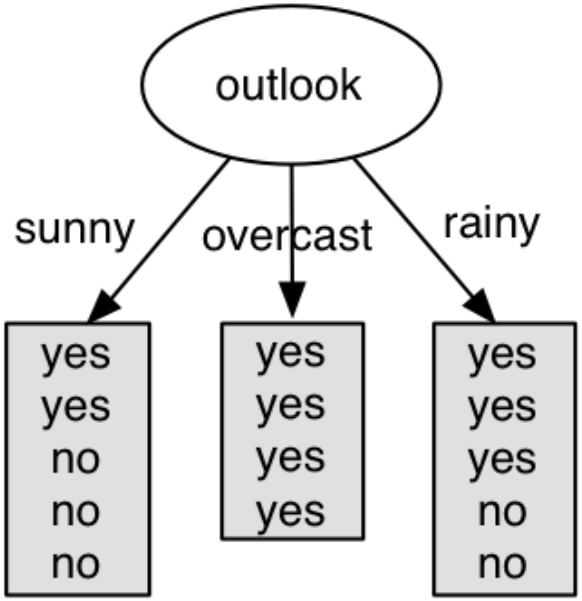


$$Gain(A_{Windy}, S) = 0.940 - \frac{8}{14} \cdot \left(-\left(\frac{6}{8} \cdot \log_2 \frac{6}{8} + \frac{2}{8} \cdot \log_2 \frac{2}{8} \right) \right) + \frac{6}{14} \cdot \left(-\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6} \right) \right) = 0.048$$

$$Gain(A, S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j)$$

Informaciona dobit atributa “outlook”

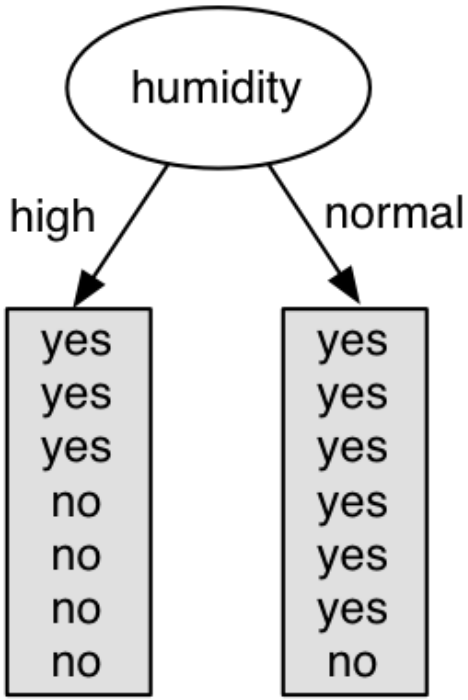
- Od ukupno 14 instanci imamo:
  - 5 instanci “sunny”
  - 4 instanci “overcast”
  - 5 instanci “rainy”



$$Gain(A_{Outlook}, S) = 0.940 - \frac{5}{14} \cdot \left(-\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right)\right) + \frac{4}{14} \cdot \left(-\left(\frac{4}{4}\log_2\frac{4}{4}\right)\right) + \frac{5}{14} \cdot \left(-\left(\frac{3}{5} \cdot \log_2\frac{3}{5} + \frac{2}{5} \cdot \log_2\frac{2}{5}\right)\right) = 0.247$$

Informaciona dobit atributa “humidity”

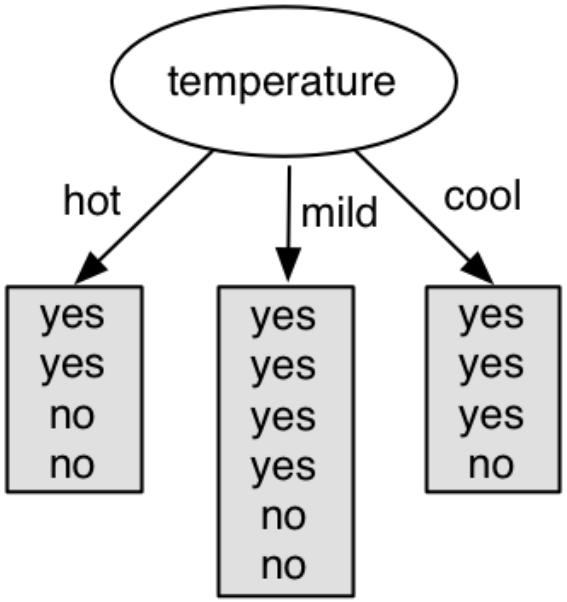
- Od ukupno 14 instanci imamo:
  - 7 instanci “high”
  - 7 instanci “normal”



$$Gain(A_{Humidity}, S) = 0.940 - \frac{7}{14} \cdot \left(-\left(\frac{3}{7} \cdot \log_2 \frac{3}{7} + \frac{4}{7} \cdot \log_2 \frac{4}{7}\right)\right) + \frac{7}{14} \cdot \left(-\left(\frac{6}{7} \cdot \log_2 \frac{6}{7} + \frac{1}{7} \cdot \log_2 \frac{1}{7}\right)\right) = 0.151$$

Informaciona dobit atributa “temperature”

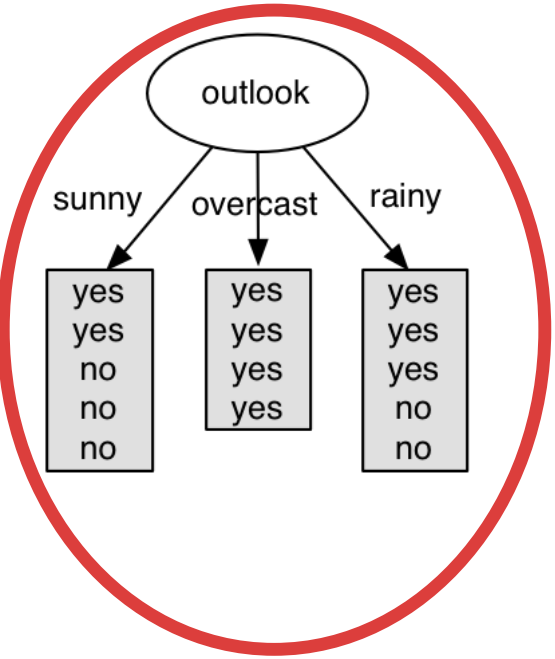
- Od ukupno 14 instanci imamo:
  - 4 instance “hot”
  - 6 instanci “mild”
  - 4 instanci “cool”



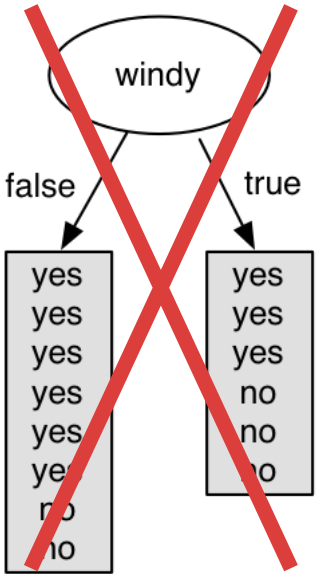
$$Gain(A_{Temperature}, S) = 0.940 - \frac{4}{14} \cdot \left(-\left(\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{2}{4} \cdot \log_2 \frac{2}{4}\right)\right) + \frac{6}{14} \cdot \left(-\left(\frac{4}{6} \cdot \log_2 \frac{4}{6} + \frac{2}{6} \cdot \log_2 \frac{2}{6}\right)\right) + \frac{4}{14} \cdot \left(-\left(\frac{3}{4} \cdot \log_2 \frac{3}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4}\right)\right) = 0.029$$

Koji atribut odabrati?

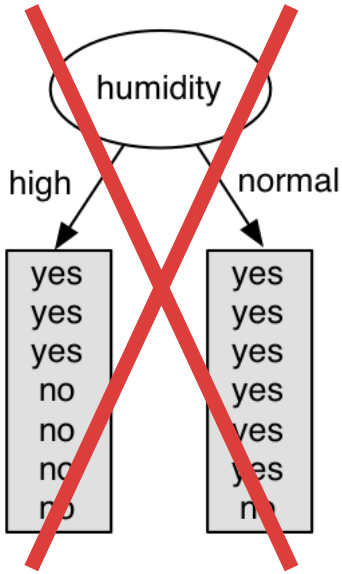
0.247



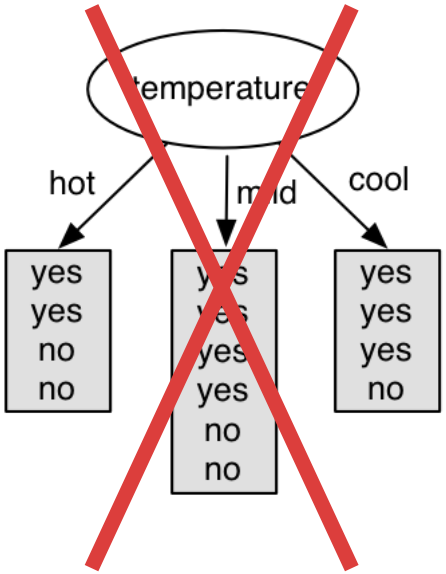
0.048



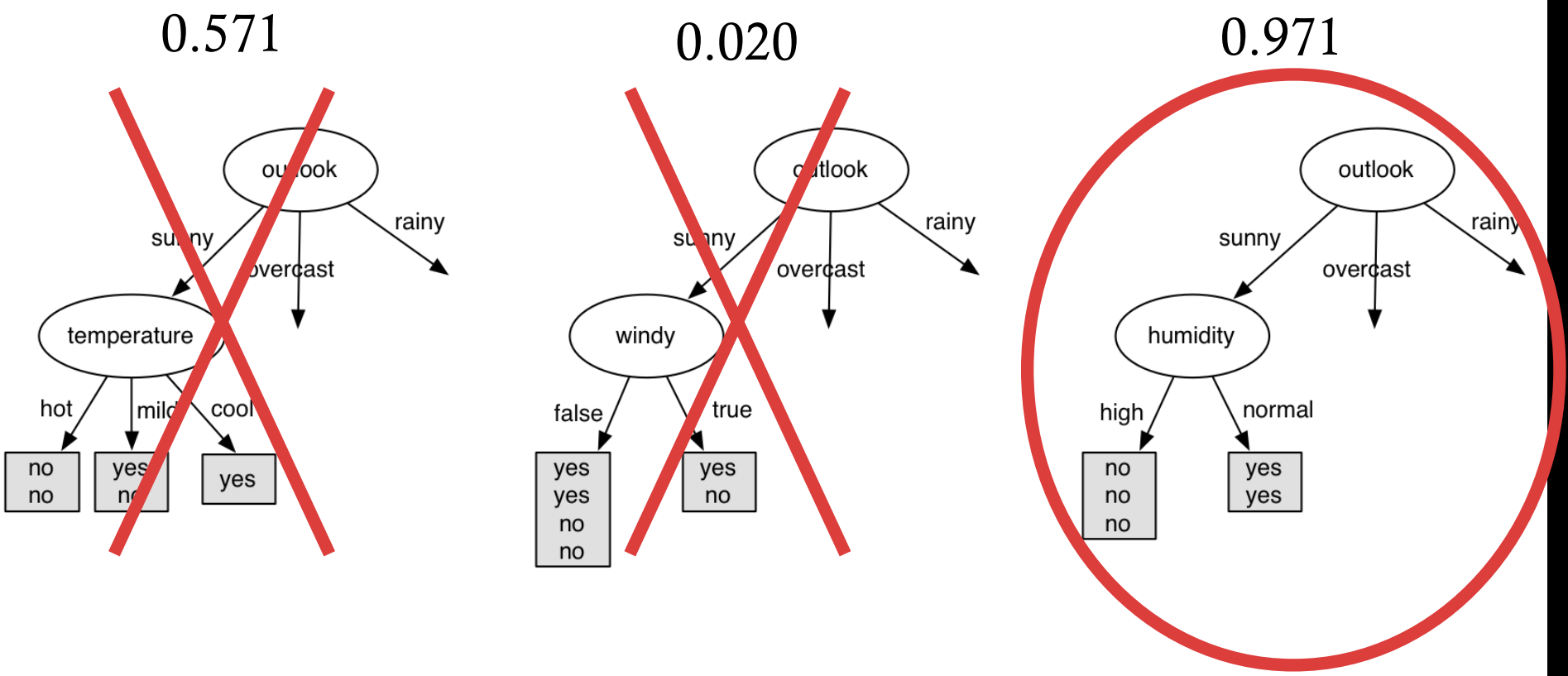
0.151



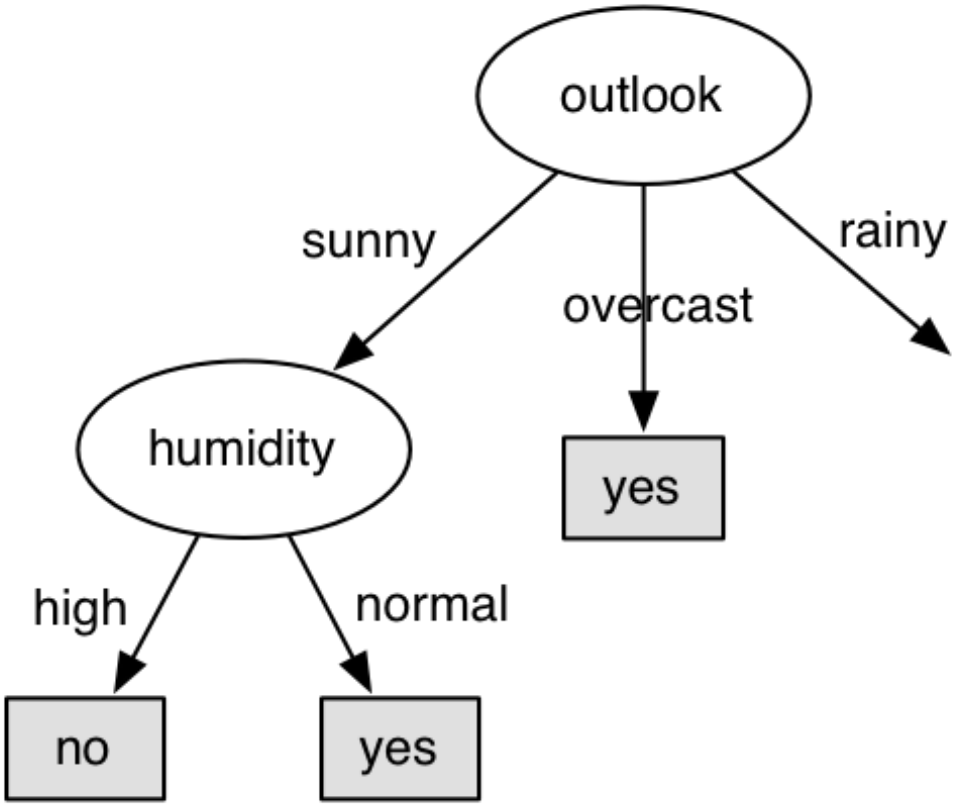
0.029



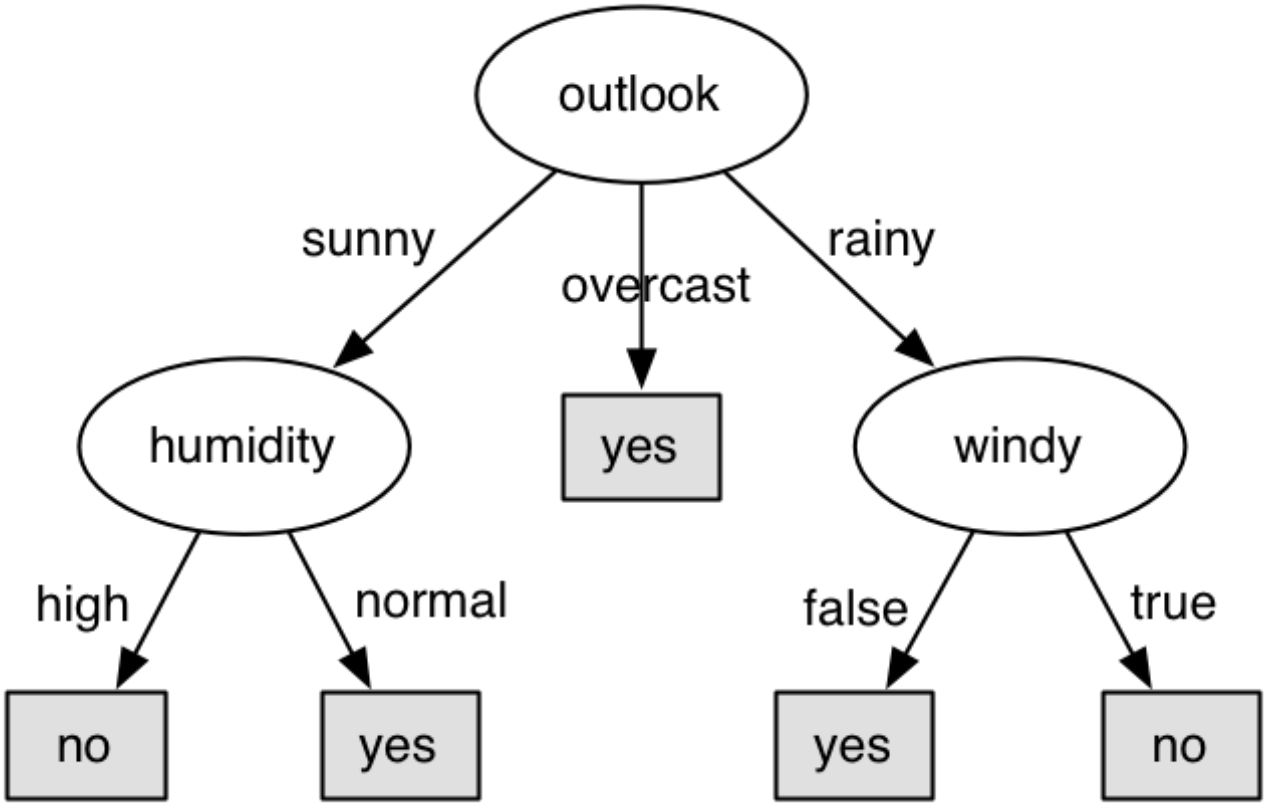
Iteracija 2: Ponavljamo postupak za svaku granu



Iteracija 2: Ponavljamo postupak za svaku granu



Iteracija 2: Ponavljamo postupak za svaku granu



Weka

- Softver za data mining u Javi
- Skup algoritama za mašinsko učenje i data mining
- Razvijen pri Univerzitetu Waikato, Novi Zeland
- Open-source
- Vebsajt: http://www.cs.waikato.ac.nz/ml/weka

ARFF fajl

- Attribute-Relation File Format – ARFF
- Tekstualni fajl

Atributi mogu biti:
- Numerički
- Nominalni

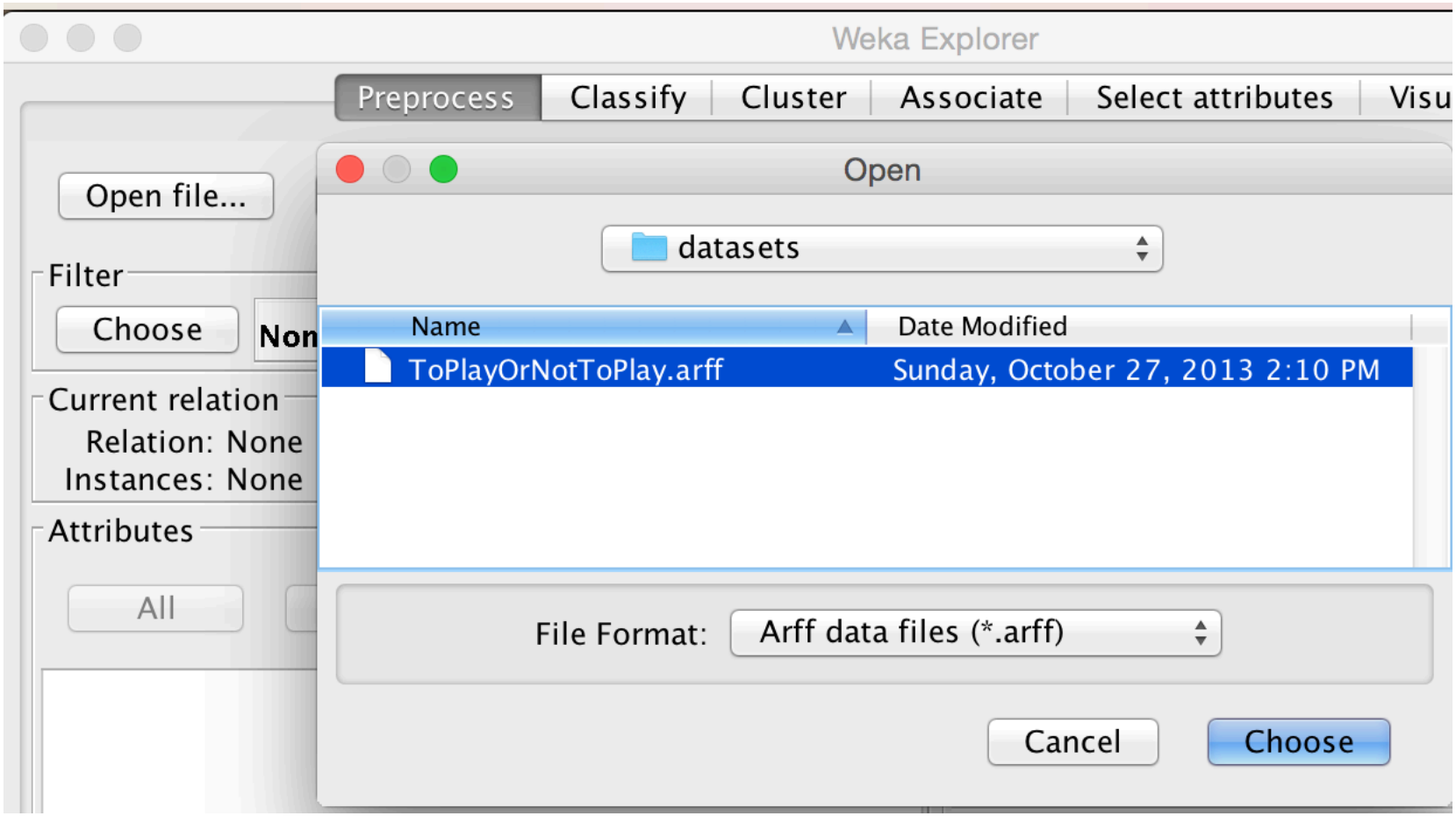
```
@relation TPONTPNom

@attribute Outlook {sunny, overcast, rainy}
@attribute Temp. {hot, mild, cool}
@attribute Humidity {high, normal}
@attribute Windy {'false', 'true'}
@attribute Play {no, yes}

@data
sunny, hot, high, 'false', no
sunny, hot, high, 'true', no
overcast, hot, high, 'false', yes
...
```

Skupovi podataka korišćeni na vežbama

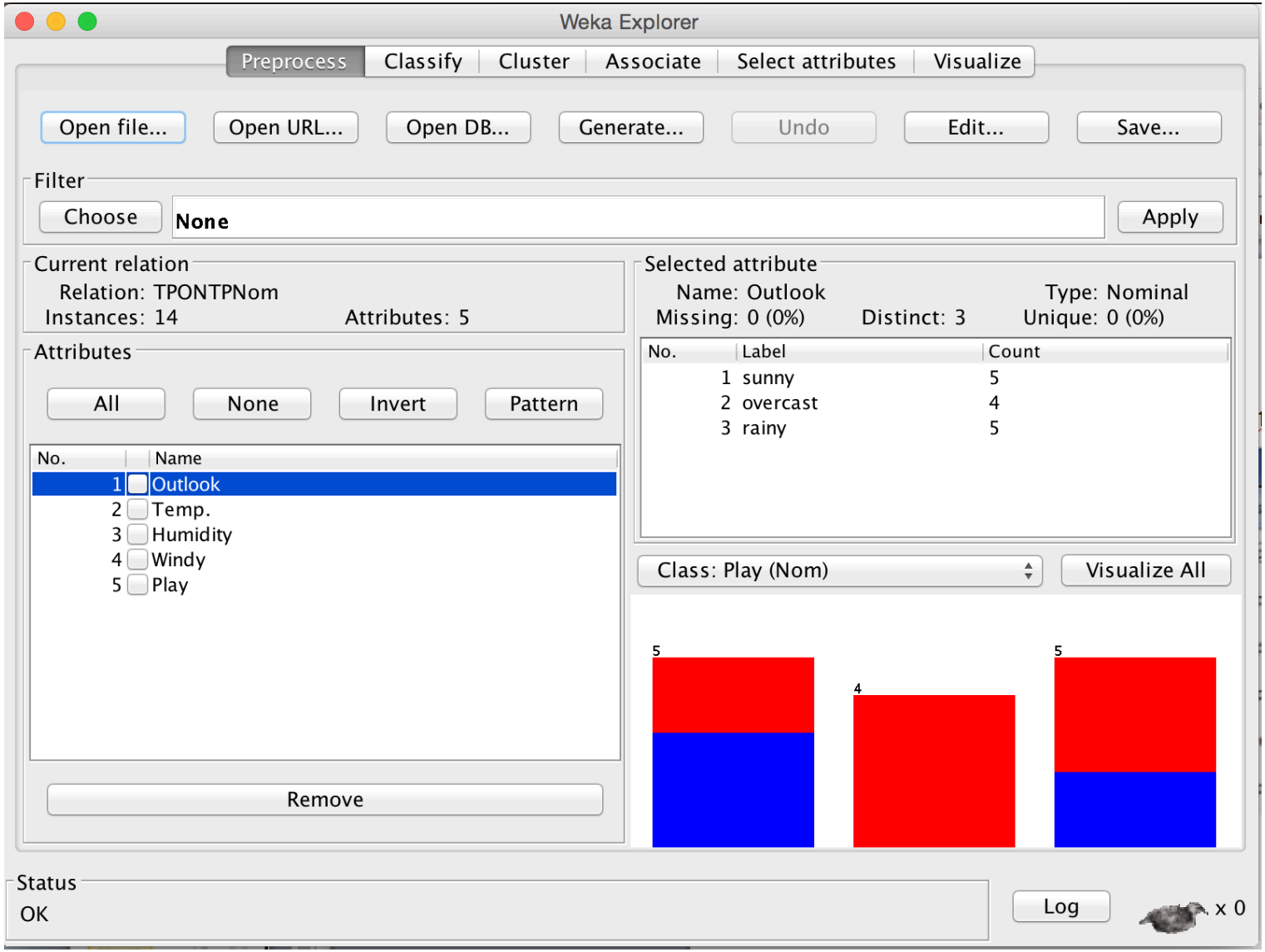
- Korišćeni skupovi podataka sa sajta Technology Forge:

http://www.technologyforge.net/Datasets

Učitavanje dataseta



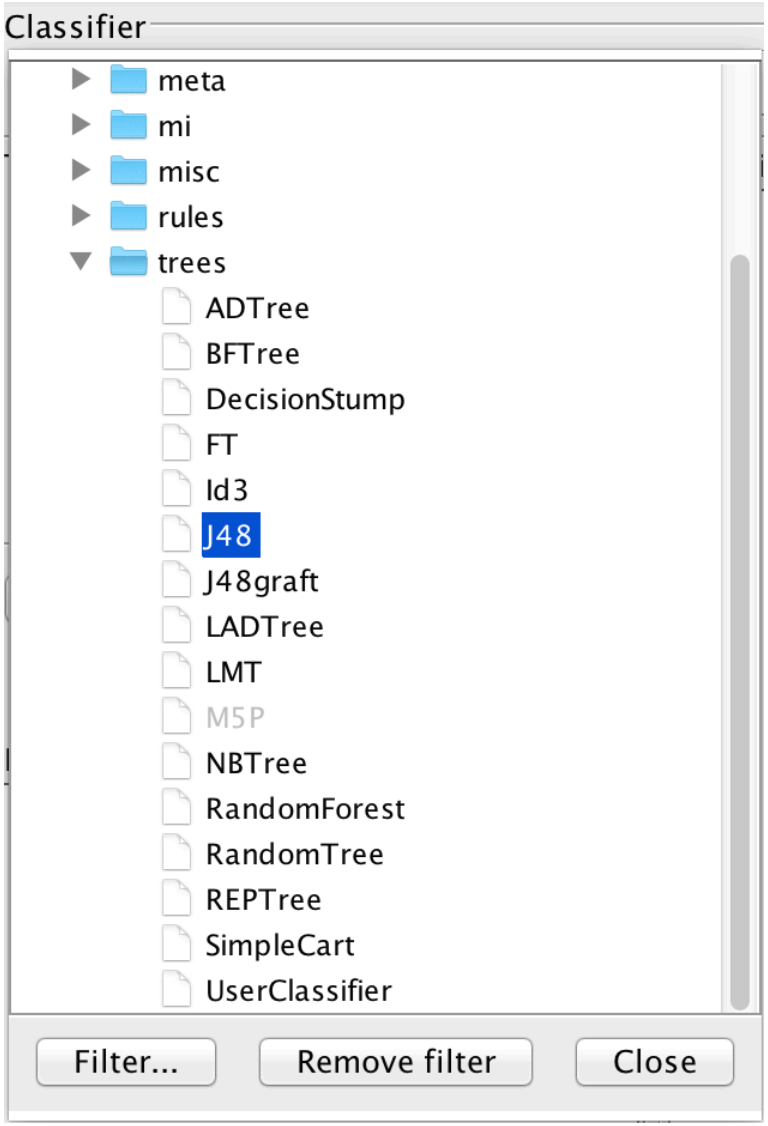
Pregled dataseta



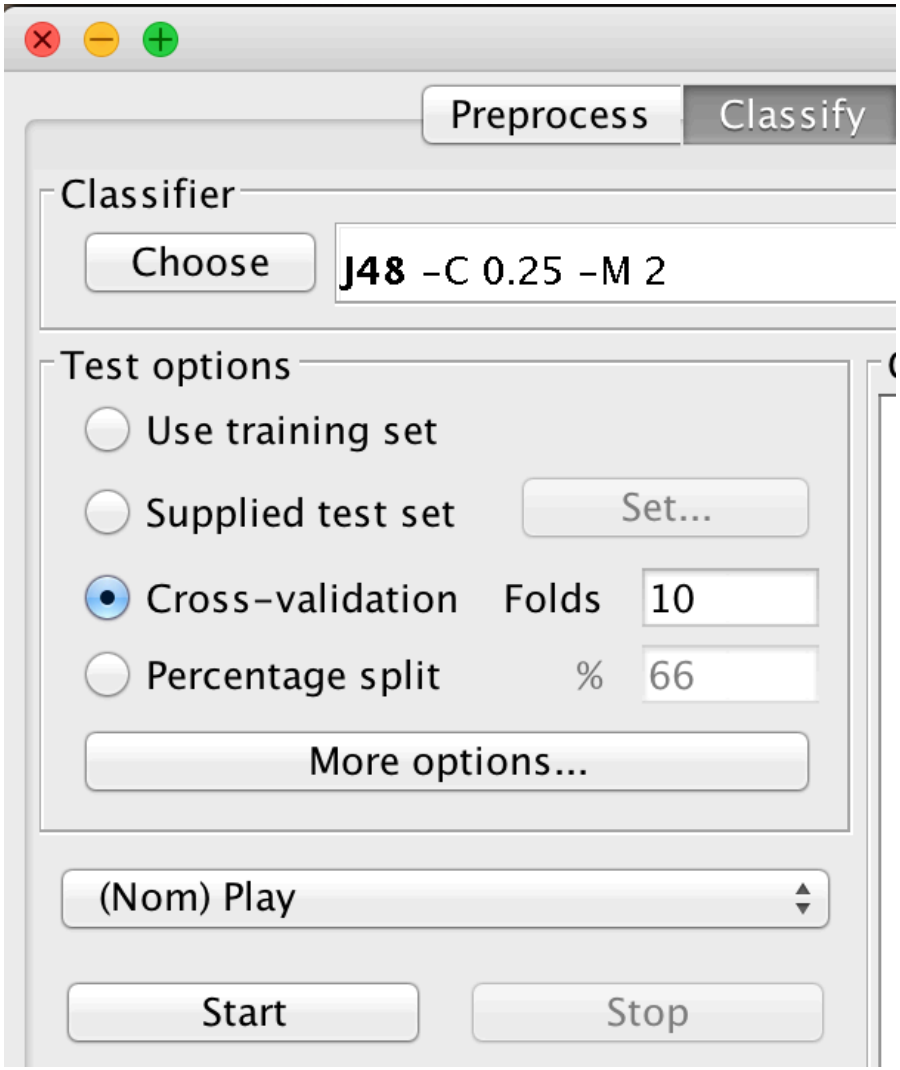
J48 klasa

- Implementacija C4.5 algoritma za generisanje stabla odlučivanja.
- C4.5 algoritam je proširenje ID3 algoritma
- Poboljšanja u odnosu na ID3 algoritam:
  - radi sa kontinualnim i diskretnim vrednostima atributa
  - podžava nedostajuće vrednosti atributa (ne uzima u obzir instance sa nedostajućim vrednostima prilikom računanja entropije i informacione dobiti)
  - orezuje stablo nakon kreiranja
- Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

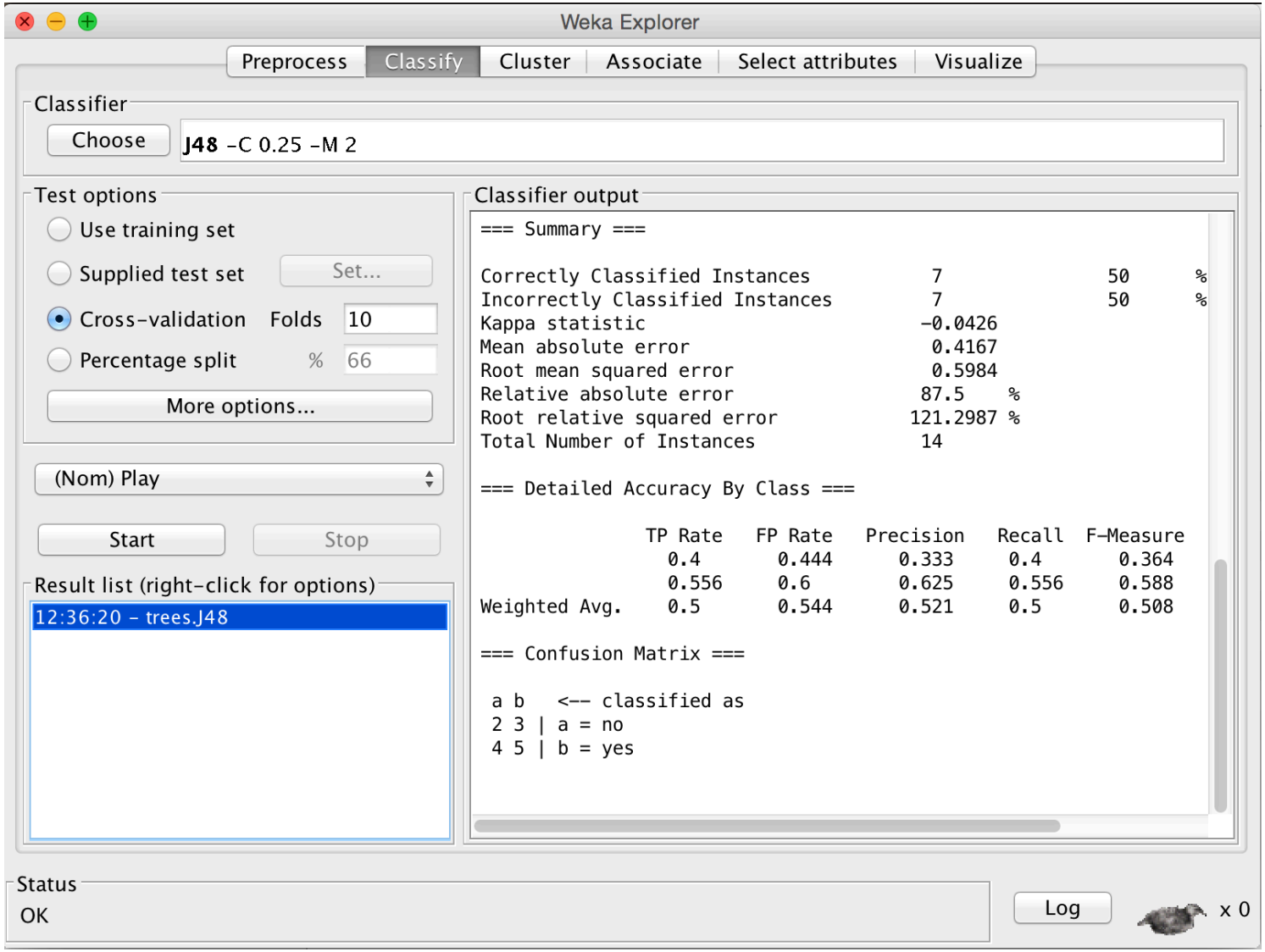
Odabir klasifikatora J48



Treniranje klasifikatora



Pregled rezultata klasifikacije



Matrica zabune (Confusion Matrix)

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative

```
=== Confusion Matrix ===

 a b   <-- classified as
 2 3 | a = no
 4 5 | b = yes
```

Precision, Recall i F measure

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.4       0.444      0.333      0.4       0.364      0.633    no
                   0.556     0.6        0.625      0.556     0.588      0.633    yes
Weighted Avg.      0.5       0.544      0.521      0.5       0.508      0.633
```

True Positives Rate

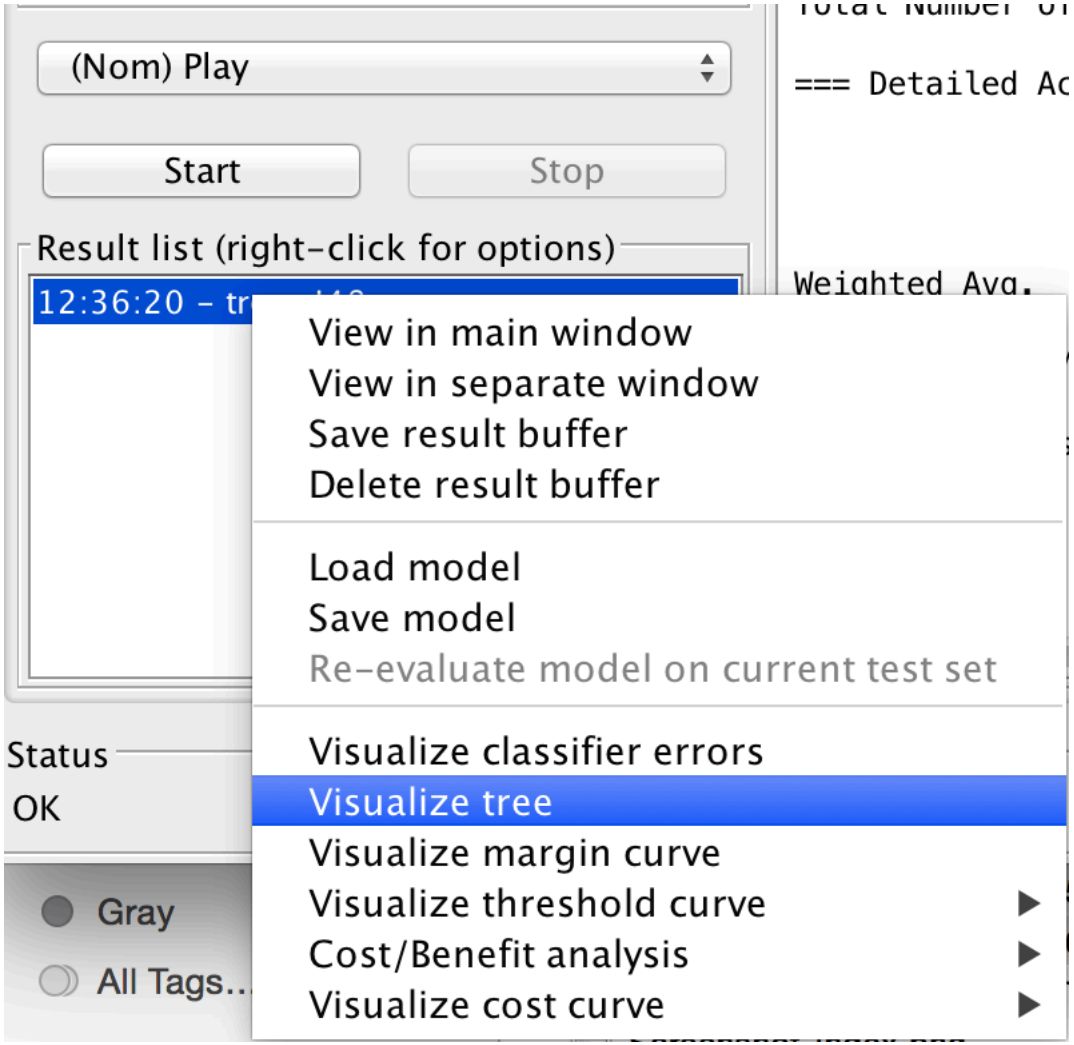
False Positives Rate

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

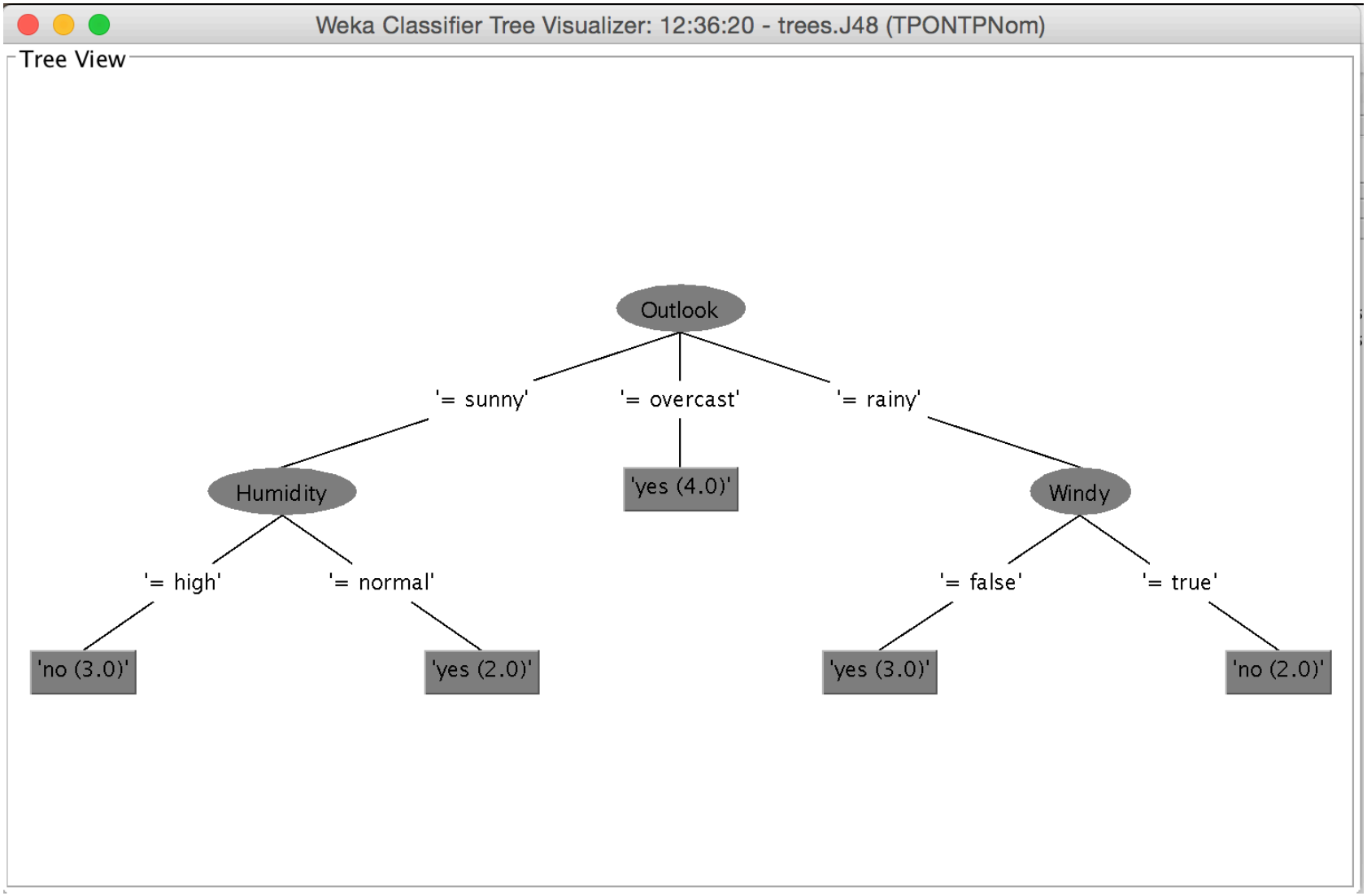
$$\text{Recall} = \frac{TP}{(TP + NP)}$$

$$\text{F measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Vizualizacija stabla odlučivanja



Vizualizacija stabla odlučivanja



Orezivanje

Orezivanje

- Orezivanje je proces smanjivanja stabla odlučivanja odstranjivanjem delova drveta (podstabla) koji u maloj meri doprinose tačnosti klasifikacije. Podstablo čija je greška prilikom klasifikovanja veća od greške koju bi doneo jedan čvor-list (terminalni čvor) se uklanja i na njegovo mesto se stavlja čvor-list.

Primer 2 – Skup podataka “Dijabetes”

- Skup podataka “Pima Indians Diabetes Database” sadrži podatke o ženama koje pripadaju grupi Pima Indijanaca koje imaju najmanje 21 godinu i testirane su na dijabetes. Dataset je kreirao Johns Hopkins University, Merilend, SAD.
- Postoji ukupno 768 instanci sa 8 atributa numeričke vrednosti koje opisuju karakteristike pacijenata i imaju podatak da li su osobe bile pozitivne ili negativne na dijabetes.
- Naš cilj je da predvidimo da li je neispitani pacijent pozitivan na dijabetes ili ne.

Primer 3 – Skup podataka “Rak dojke”

- Skup podataka “Breast cancer data” sadrži podatke o pacijentima obolelim od raka dojke sa Instituta za ongkologiju iz Ljubljane, Slovenija.
- Postoji ukupno 286 instanci sa 9 atributa koji su dobijeni od pacijenata obolelih od raka dojke.
- Naš cilj je da predvidimo da li će se tumor vratiti kod pacijenta.

Preporuke i zahvalnice

Weka Tutorials and Assignments @ The Technology Forge

- Link: http://www.technologyforge.net/WekaTutorials/

"Data Mining with Weka" and "More Data Mining with Weka": MOOCs from the University of Waikato. A self-paced session of "Data Mining with Weka" runs until 23 October June 2015.

- Link: https://www.youtube.com/user/WekaMOOC/

(Anonimni) upitnik za vaše kritike, komentare, predloge:

http://goo.gl/cqdp3I

Pitanja?

NIKOLA MILIKIĆ

EMAIL: nikola.milikic@fon.bg.ac.rs

URL: http://nikola.milikic.info