Classification – Decision Trees

UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

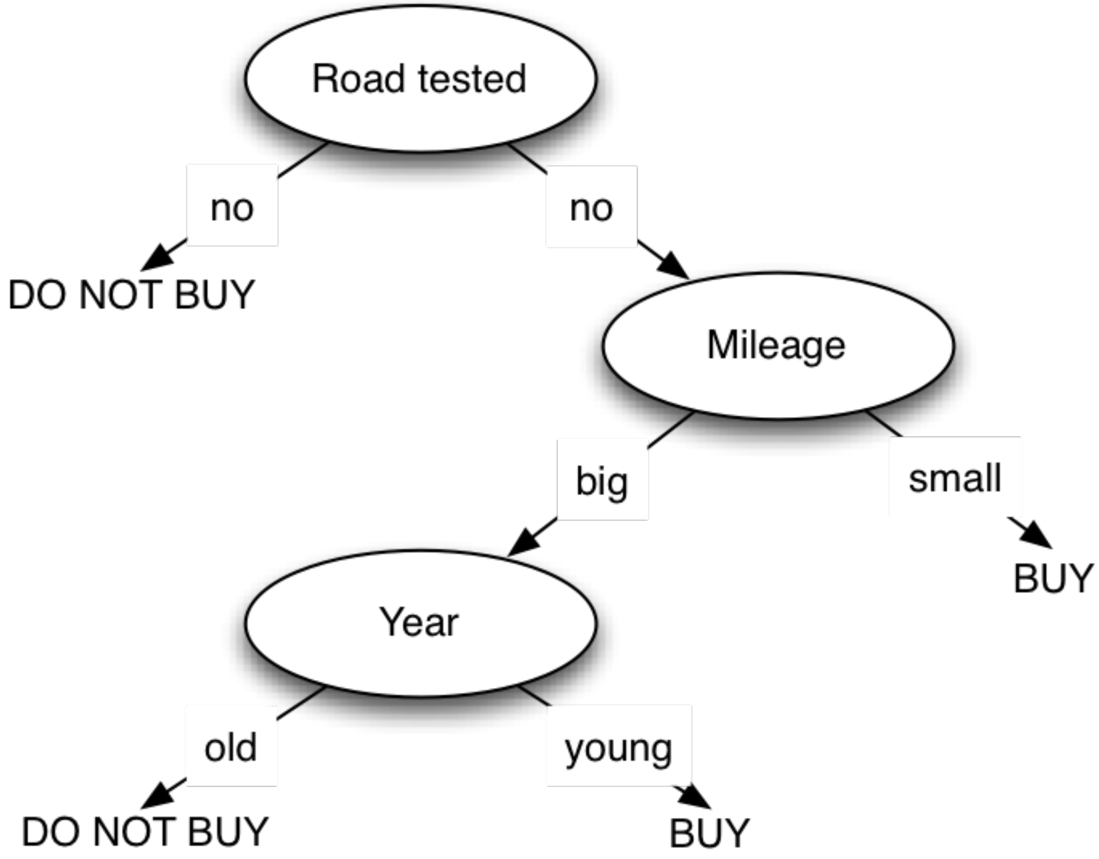
URL: http://www.krcadinac.com

What is classification?

- The task of defining a class which an instance belongs to
  - an instance is defined by a set of attributes;
  - a set of possible classes is given

Decision trees

Example: Deciding whether to buy a car



ID3 algorithm

- ID3 - Iterative Dichotomiser 3
- One of the best known algorithms for generating decision trees based on the set of examples (dataset)
- Resulting tree can be used for classifying future (unknown) instances

Example – Forecasting whether the play will be played

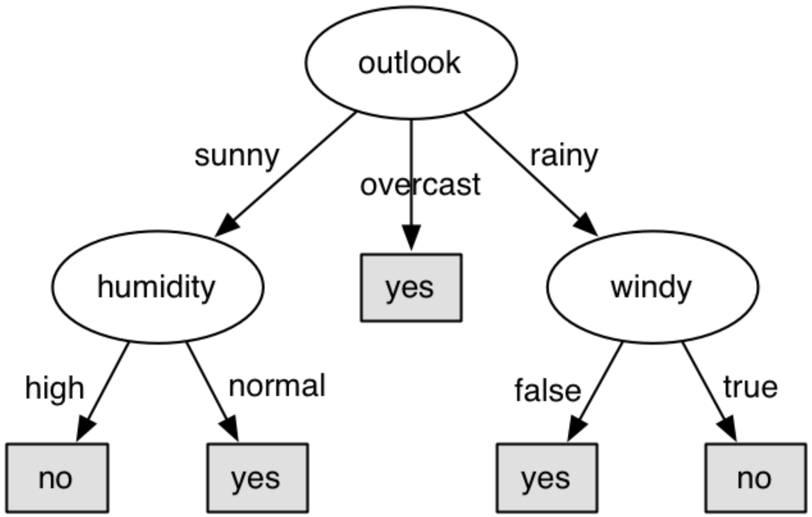
ToPlayOtNotToPlay.arff dataset

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

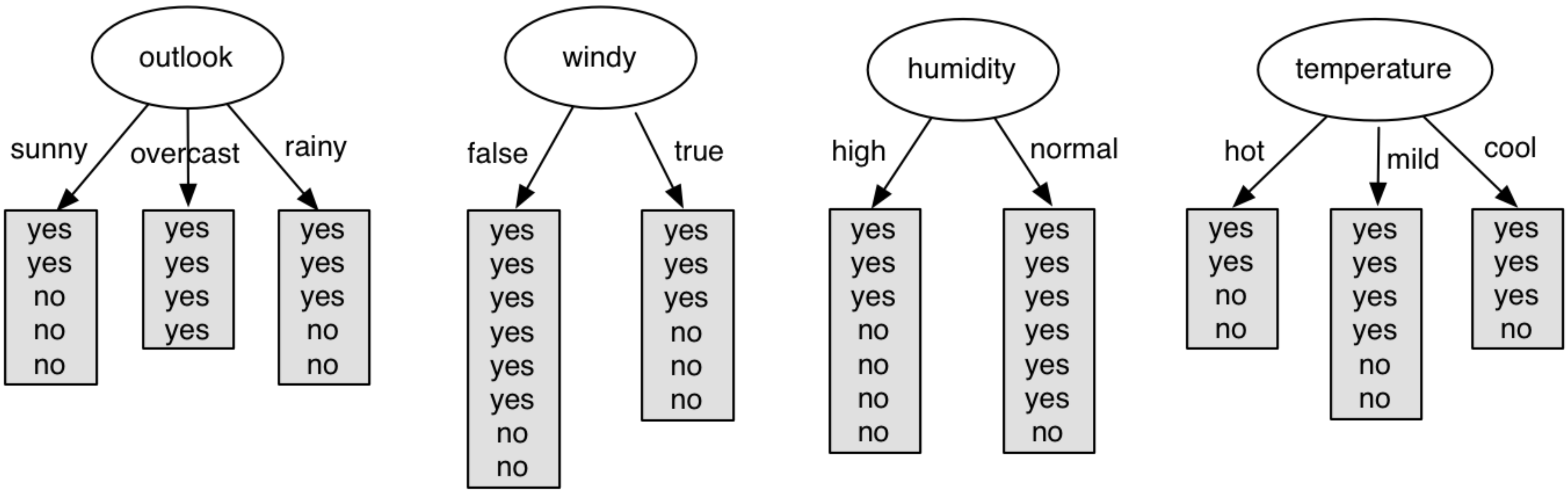
Top-down approach

Recursive divide-and-conquer:

- **Select** attribute for root node
  - Create branch for each possible attribute value
- **Split** instances into subsets
  - One for each branch extending from the node
- **Repeat** recursively for each branch
  - using only instances that reach the branch
- **Stop**
  - if all instances have the same class



Which attribute to select?



Which attribute to select?

- **Aim**: to get the smallest tree
- **Information theory**: measure information in bits. Founder is Claude Shannon, American mathematician and scientist 1916 - 2001
- Entropy H(S) can be calculated by using the formula:

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

where:

- S – set of all instances in the dataset
- N – number of distinct class values
- p_i – event probability

Dataset entropy

- From the total of 14 instances we have:
  - 9 instances "yes"
  - 5 instances "no"

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

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

Information gain

- Information gain Gain(A, S) of an attribute A over the set of instances S represents an amount of information we would gain by knowing the value of the attribute A. Information gain represents the difference between an entropy before branching and entropy after branching over the attribute A.

Information gain

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

where:

- $H(S)$ – entropy of the whole dataset S
- $|S_j|$ – number of instance with j value of an attribute A
- $|S|$ – total number of instances in dataset S
- v – set of distinct values of an attribute A
- $H(S_j)$ – Entropy of subset of instances for attribute A
- $H(A, S)$ – entropy of an attribute A

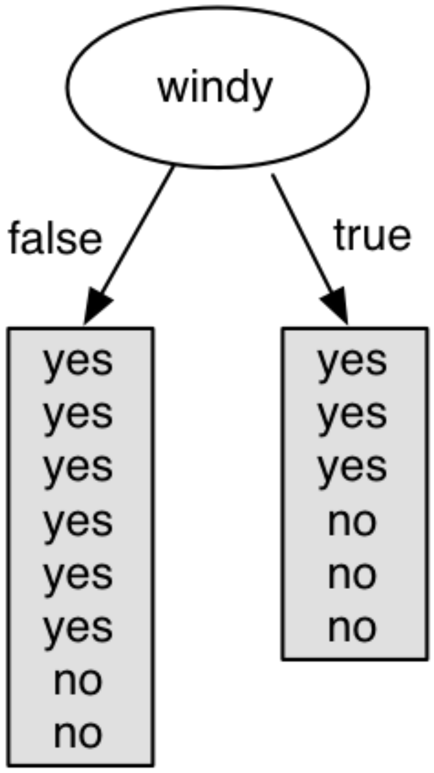
Choose an attribute with highest information gain.

Information gain of attribute “windy”

- From the total of 14 instances we have:
    - 6 instances “true”
    - 8 instances “false”

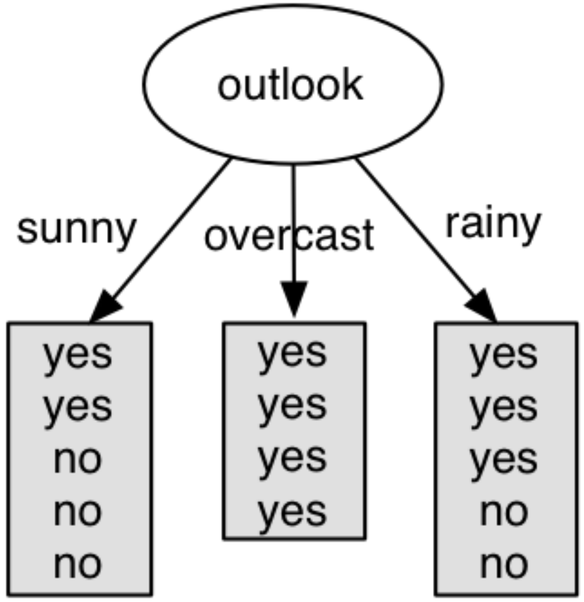
$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j)$$

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



Information gain of attribute “outlook”

- From the total of 14 instances we have:
  - 5 instances “sunny”
  - 4 instances “overcast”
  - 5 instances “rainy”

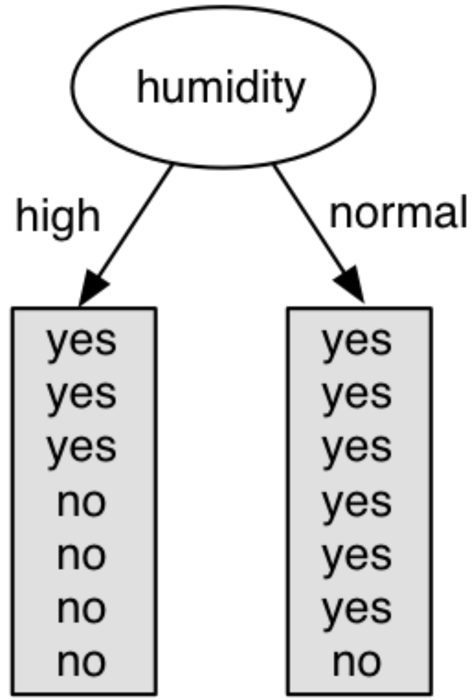


$$Gain(A_{Outlook}, S) = 0.940 - \frac{5}{14}\cdot\left(-\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5}\right)\right) + \frac{4}{14}\cdot\left(-\left(\frac{4}{4}\log_2\frac{4}{4}\right)\right) + \frac{5}{14}\cdot\left(-\left(\frac{3}{5}\cdot\log_2\frac{3}{5}+\frac{2}{5}\cdot\log_2\frac{2}{5}\right)\right)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

Information gain of attribute "humidity"

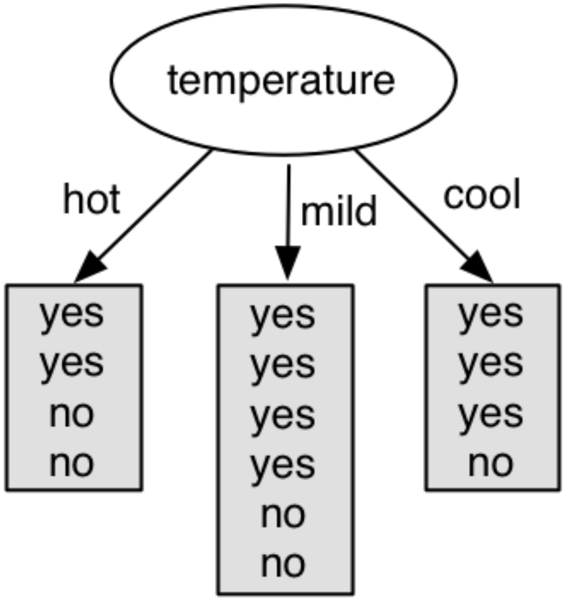
- From the total of 14 instances we have:
  - 7 instances "high"
  - 7 instances "normal"

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



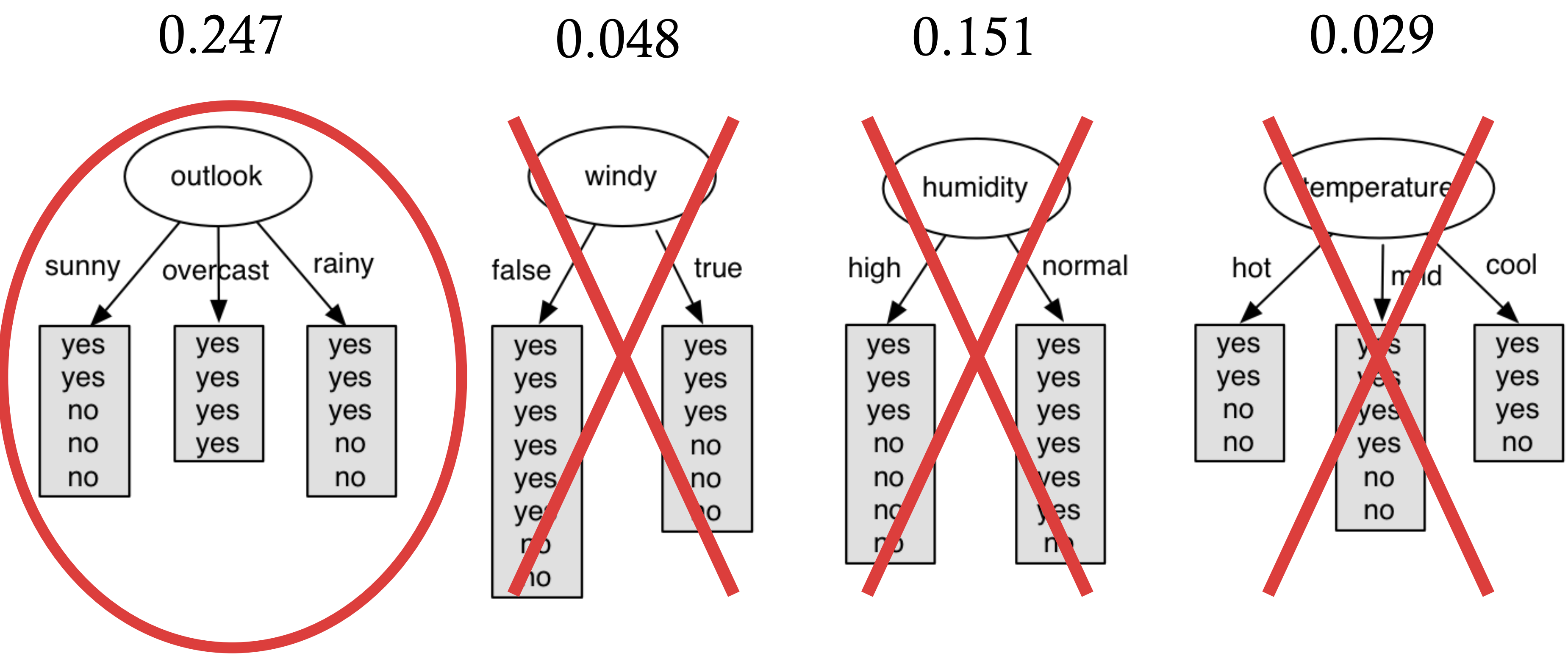
Information gain of attribute "temperature"

- From the total of 14 instances we have:
  - 4 instances "hot"
  - 6 instances "mild"
  - 4 instances "cool"

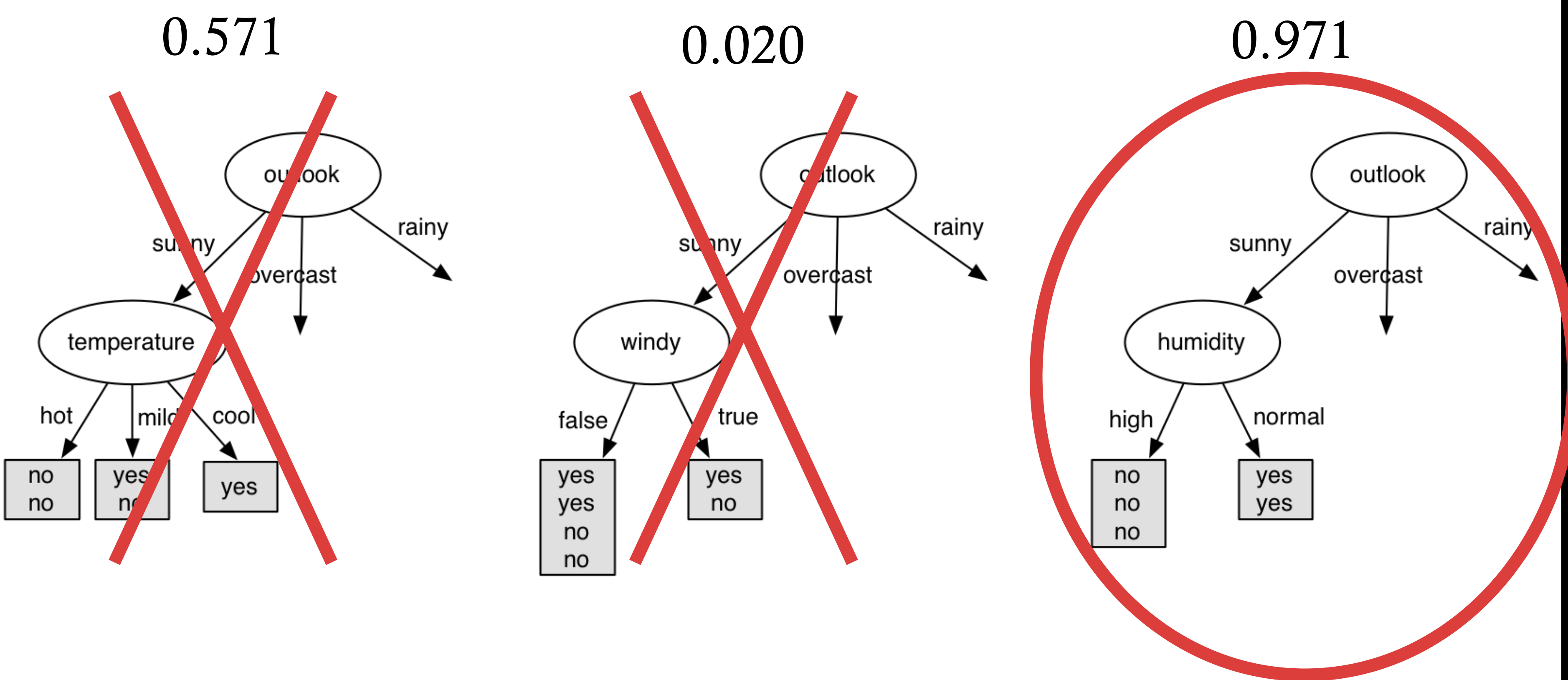


$$Gain(A_{Temperature}, S) = 0.940 - \frac{4}{14}\cdot\left(-\left(\frac{2}{4}\cdot\log_2\frac{2}{4} + \frac{2}{4}\cdot\log_2\frac{2}{4}\right)\right) + \frac{6}{14}\cdot\left(-\left(\frac{4}{6}\cdot\log_2\frac{4}{6} + \frac{2}{6}\cdot\log_2\frac{2}{6}\right)\right) + \frac{4}{14}\cdot\left(-\left(\frac{3}{4}\cdot\log_2\frac{3}{4}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

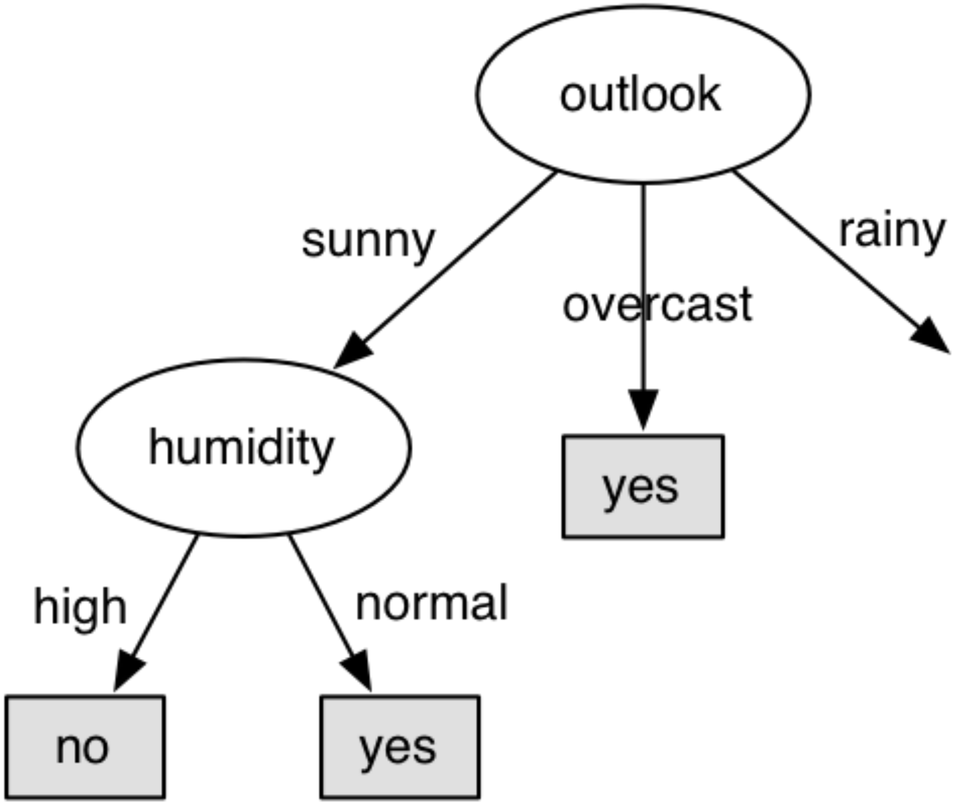
Which attribute to select?



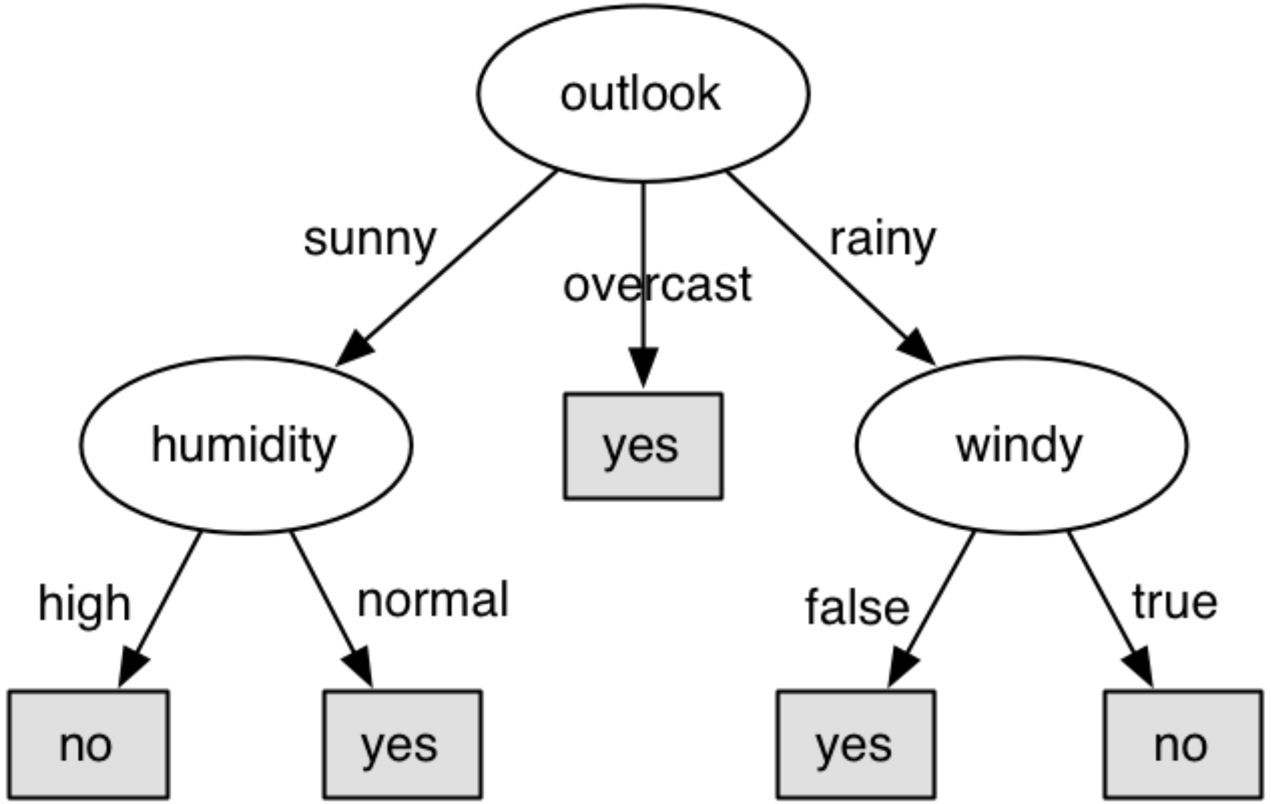
Iteration 2: Repeat recursively for each branch



Iteration 2: Repeat recursively for each branch



Iteration 2: Repeat recursively for each branch



Weka

- Software for data mining in Java
- Set of algorithms for machine learning and data mining
- Developed at the University of Waikato, New Zealand
- Open-source
- Website: http://www.cs.waikato.ac.nz/ml/weka

ARFF file

- Attribute-Relation File Format – ARFF
- Textual file

Attributes can be:
- Numerical
- Nominal

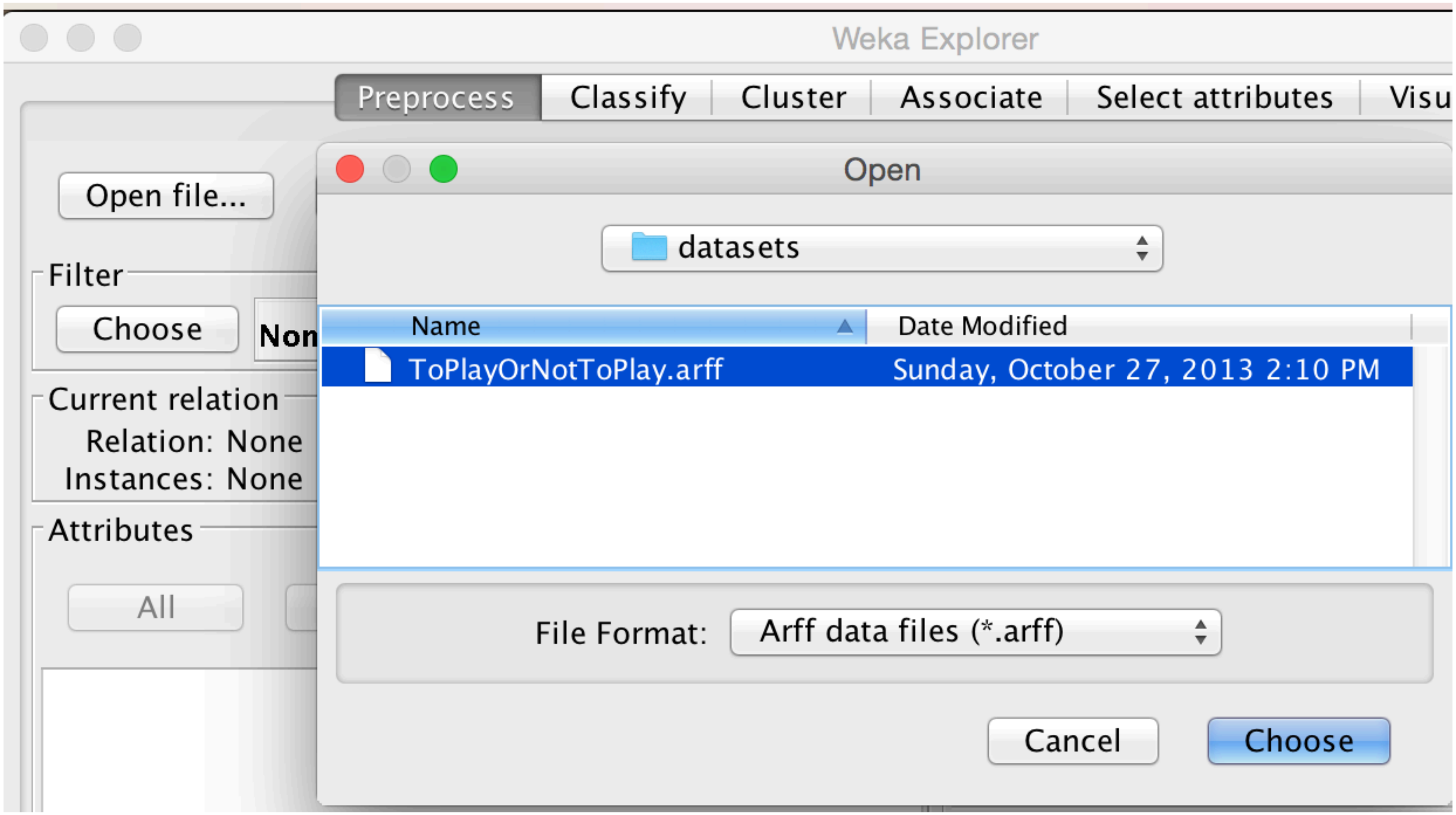
```
@relation TPONTPNom

@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
...
```

Datasets used for this class

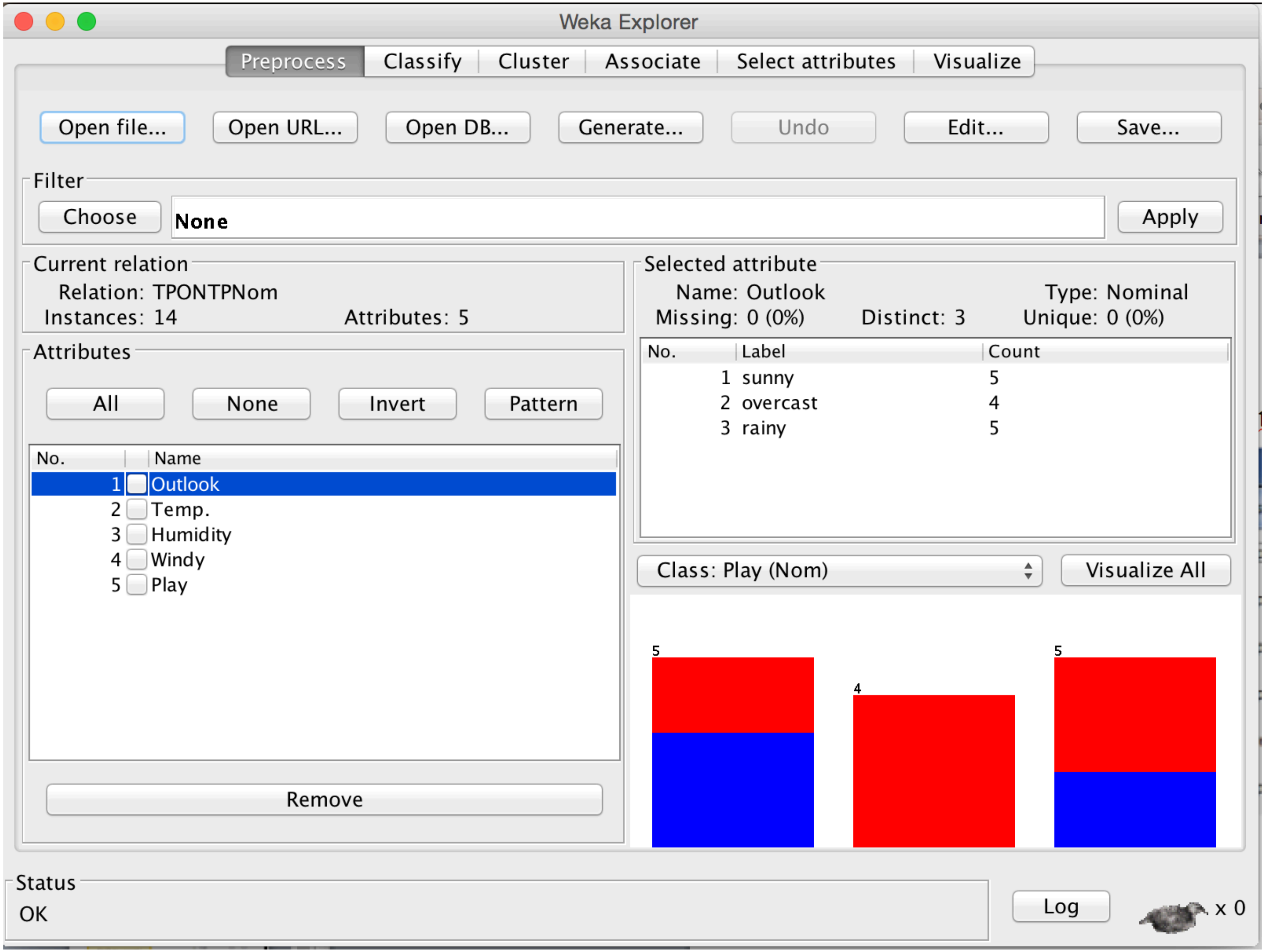
- Datasets from the webiste Technology Forge:

http://www.technologyforge.net/Datasets

Loading dataset



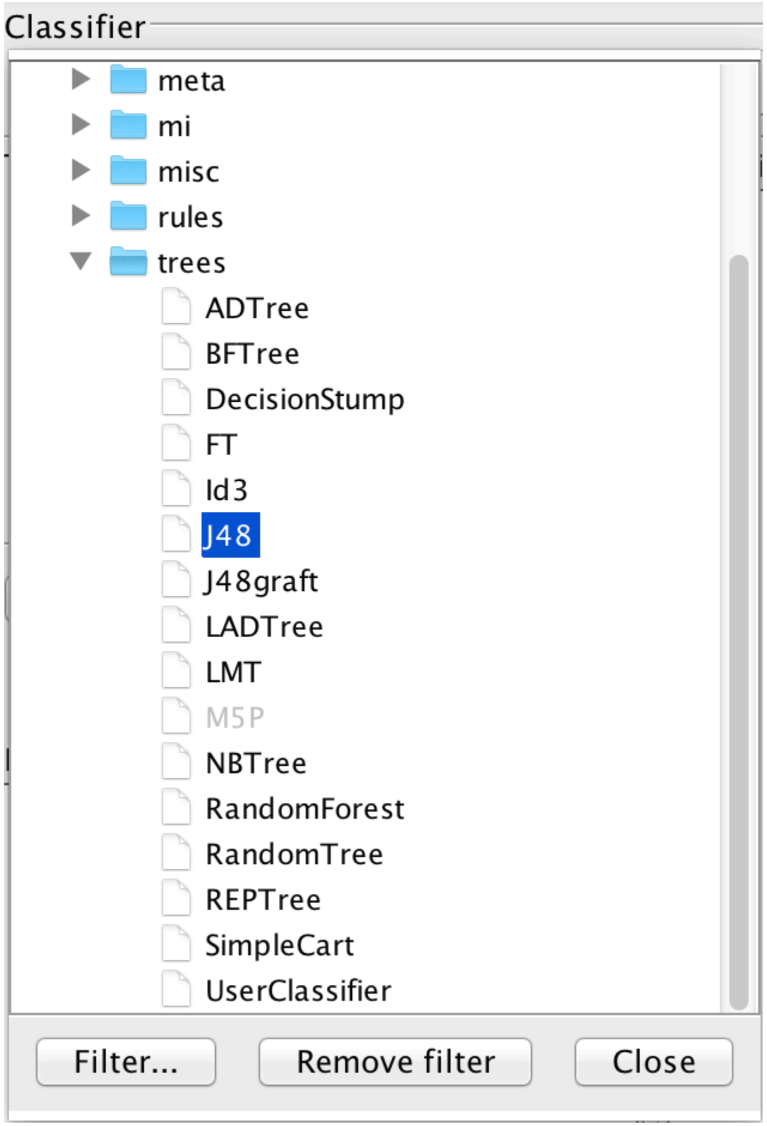
Dataset overview



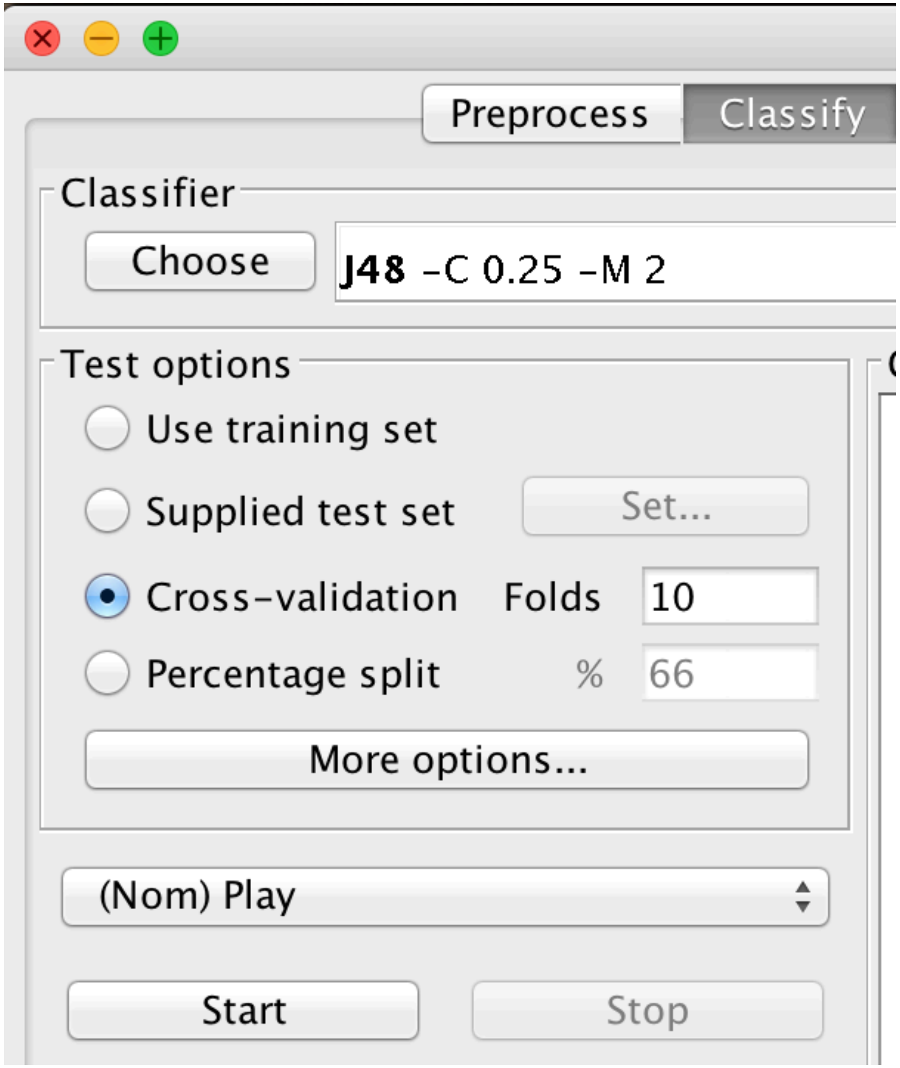
J48 class

- Implementation of C4.5 algorithm for generating decision trees.
- C4.5 algorithm is an extension of the ID3 algorithm.
- Extending theID3 algorithm by:
    - supporting continual and discrete attributes
    - supporting missing values (excludes instances with missing values when calculating entropy and information gain)
    - tree pruning
- Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

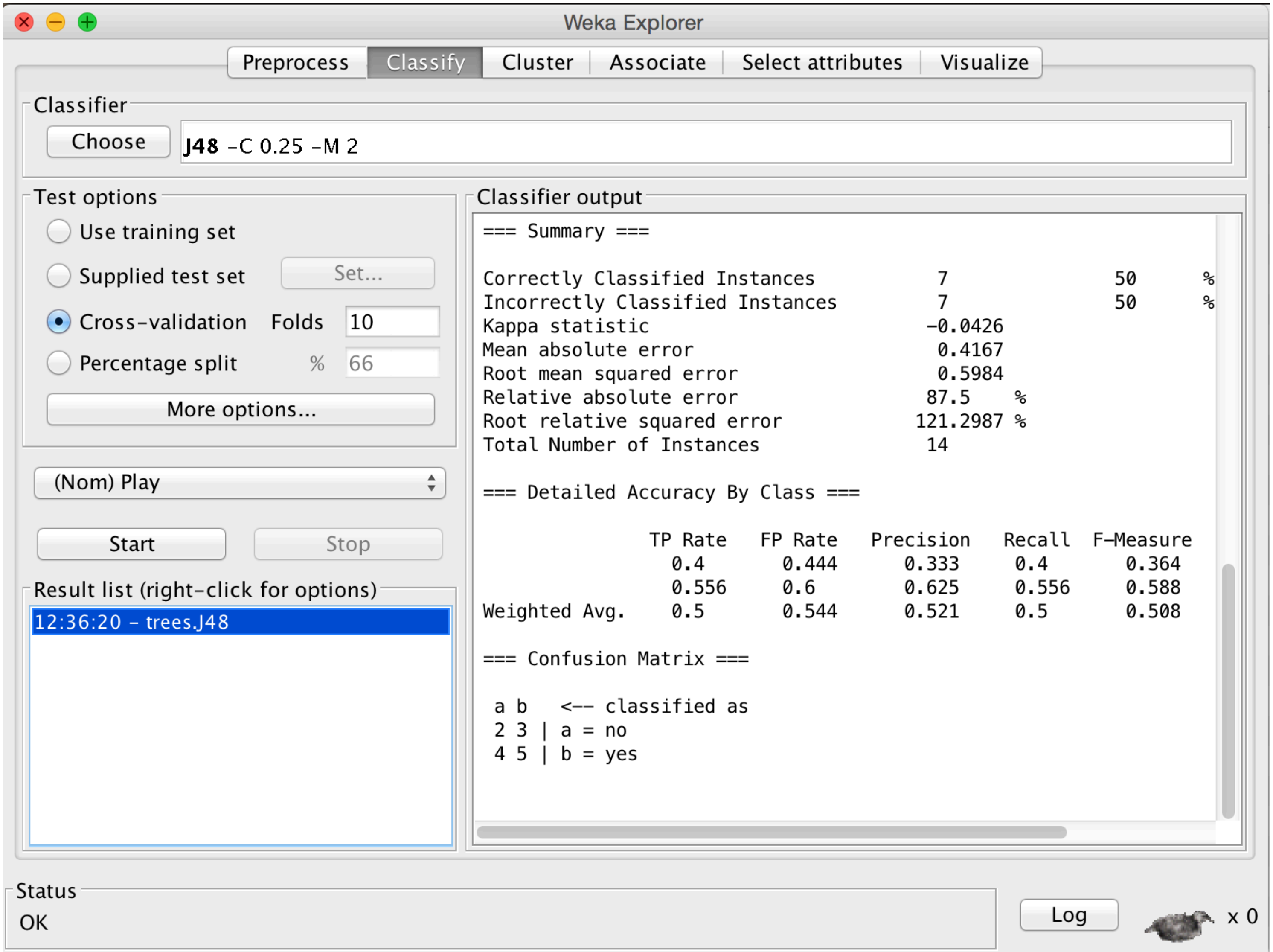
Choossing J48 classifier



Training the classifier



Overview of classification results



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 and 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

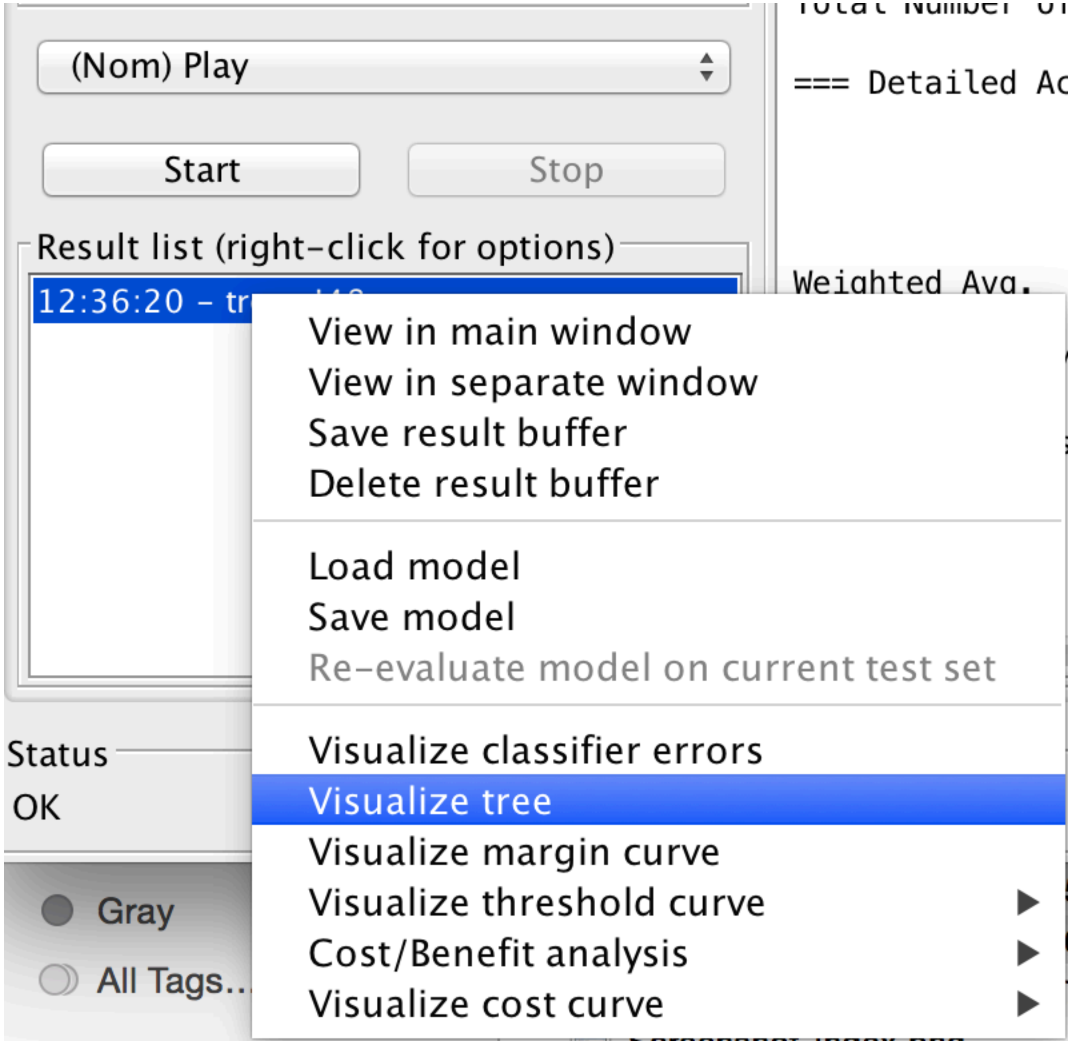
False Positives Rate

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

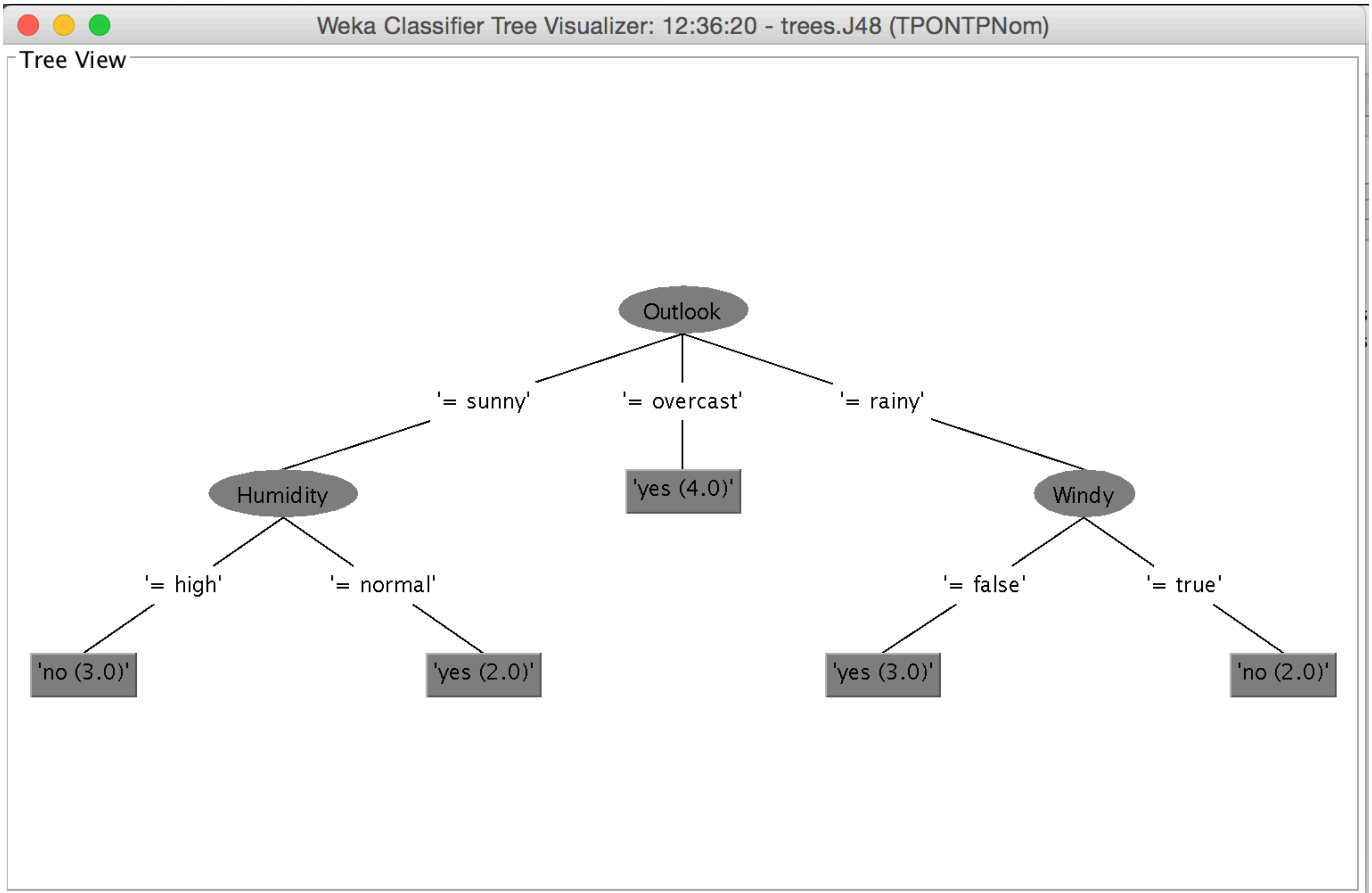
$$\text{Recall} = \frac{TP}{(TP + NP)}$$

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

Visualizing decision tree



Visualizing decision tree



Tree prunning

Tree pruning

- Pruning is the process of reducing the tree size by removing sub-trees that adds little to the efficiency of the decision tree. Sub-tree whose classification error is bigger than the error of a leaf node in its place is removed and replaced by the leaf node.

Example 2 – “Diabetes” dataset

- Dataset “Pima Indians Diabetes Database” contains data about female Pima Indians aged 21 years or higher and tested for diabetes. Dataset was donated by the Johns Hopkins University, Maryland, USA.

- There are total of 768 instances described by 8 numerical attributes about patient conditions and annotated with a class determining whether patients were positive or negative for diabetes.

- Our goal is to predict whether a new patient will be diagnosed positive or negative.

Example 3 – “Breast cancer” dataset

- “Breast cancer data” dataset contains information about patients diagnosed with breast cancer donated by Institute of Oncology, Ljubljana, Slovenia.
- This data set includes 201instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.
- Our goal is to predict whether there will be recurrent events or not.

Credits

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/

(Anonymous) survey for your comments ad suggestions:

http://goo.gl/cqdp3I

Questions?

UROŠ KRČADINAC

EMAIL: uros@krcadinac.com

URL: http://www.krcadinac.com