CSCI3230
Introduction to Data Mining

Fall 2013

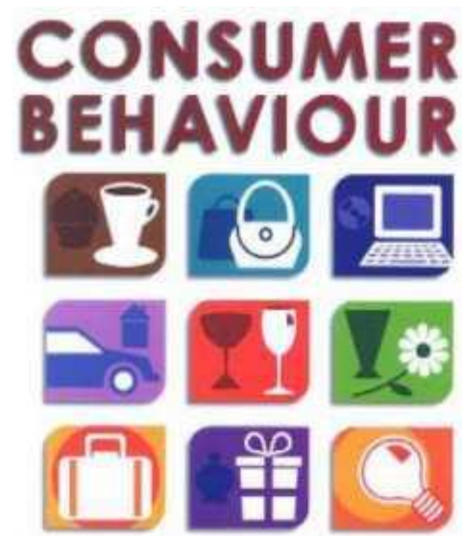
Week8, Antonio

Introduction

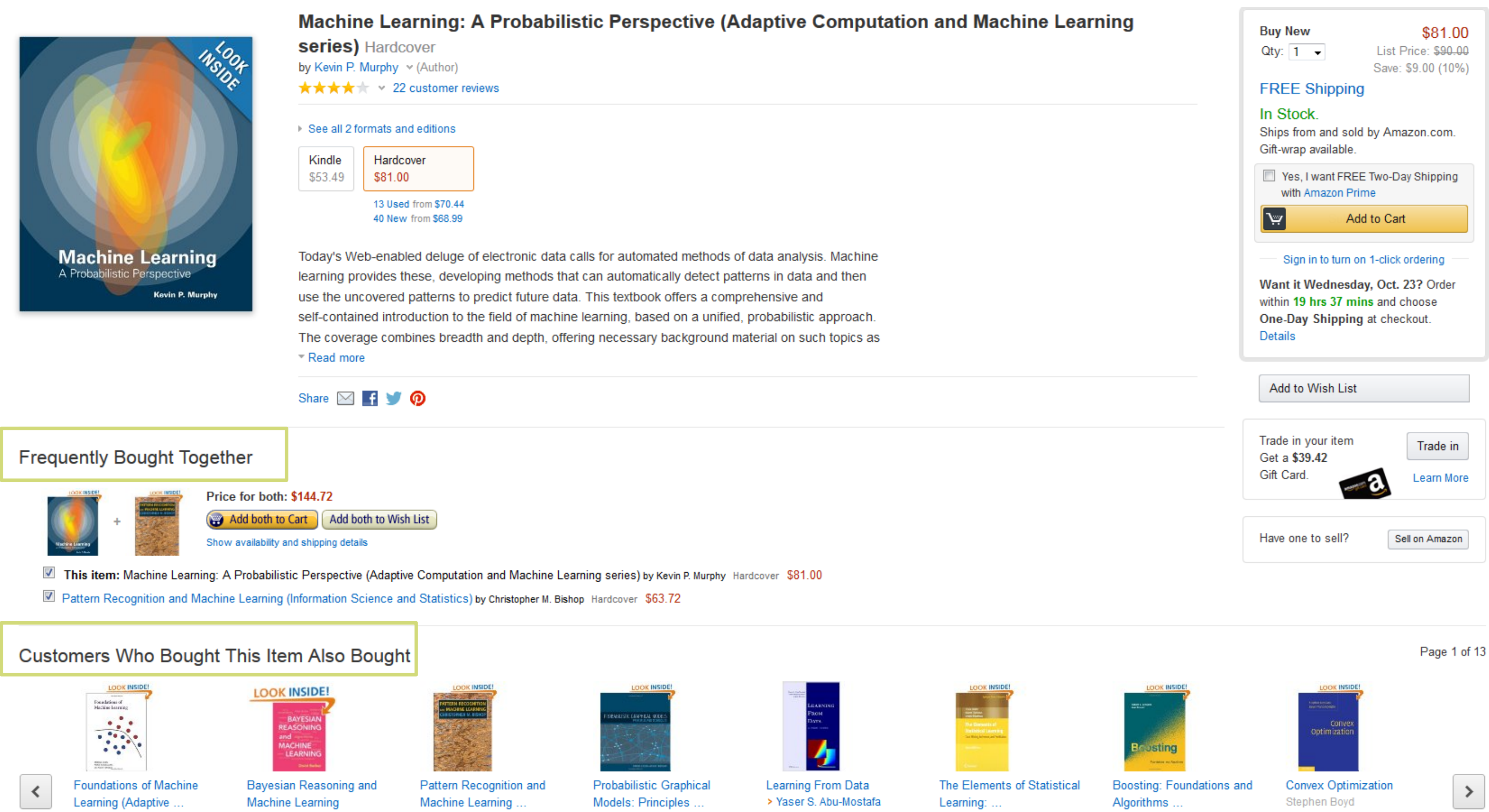
- Name: Sze-To Ho Yin, Antonio
- Office: SHB 1013
- Office hour: 14:30 – 16:30, Wednesday
- Email: hyszeto@cse.cuhk.edu.hk
- Language: Cantonese, English, Mandarin (A little)
- You are welcome to discuss with me any materials related to CSCI 3230.

Data Mining – Business Application: Consumer Behavior Discovery

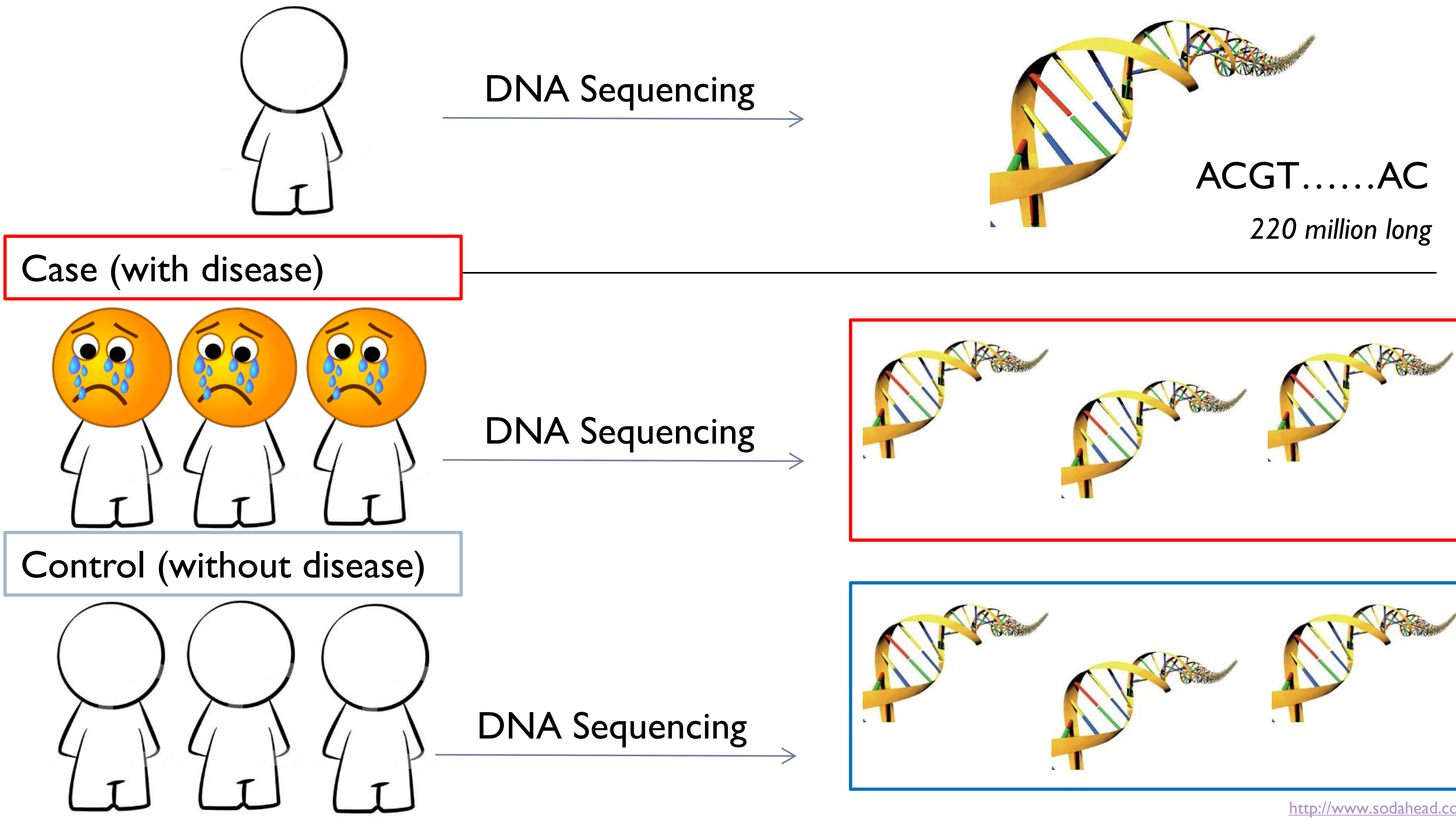
- Collect a huge amount of data from the users.
- Study when, why, how, and where people do or do not buy a product.
- We may discover new knowledge about your target customers.



Data Mining – Business Application: Consumer Behavior Discovery



Data Mining – Biological Application: Genome-Wide Association Study



http://www.sodahead.com
http://www.biol.unt.edu/~jajohnson/DNA_sequencing_process
http://www.illustrationsof.com/63130-royalty-free-human-factor-clipart-illustration

Tutorial outline

- Why Data Mining ?
- What is Data Mining ?
- How to mine data using WEKA?
  - Dataset and format
  - Data Preprocessing
  - Data Mining (Classification)
- How well are you doing?
  - Model Evaluation

Why Data Mining?

Data Mining helps us detect something new from data!!!

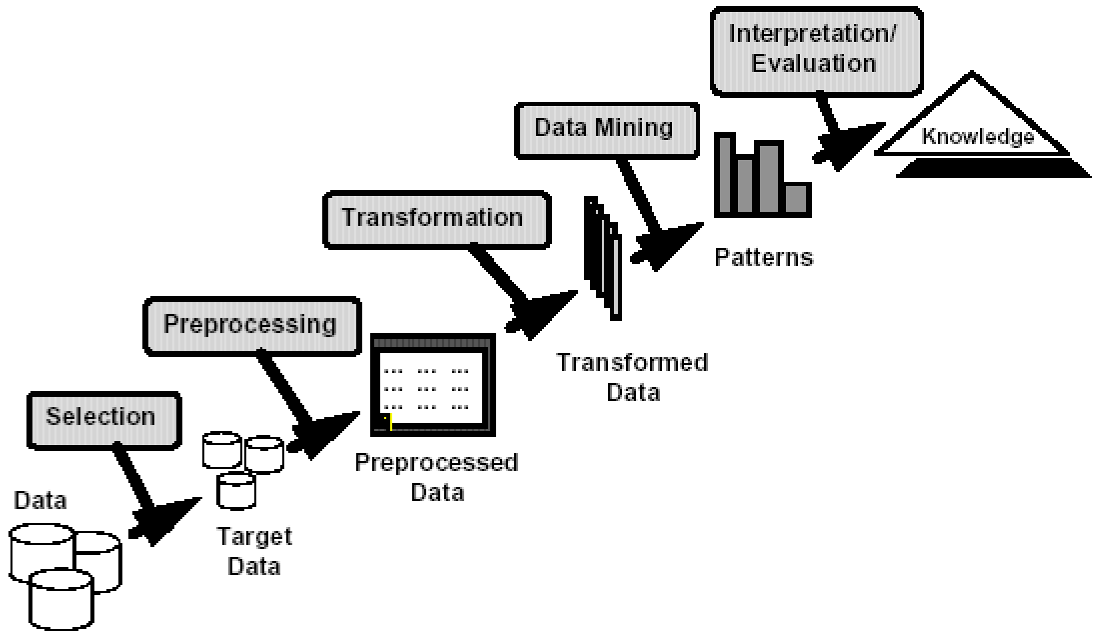
Why Data Mining?

- There is often hidden information behind data
- Human may take weeks or months to discover them
- The information discovered are useful for enhancing our understanding about an issue and predicting a future trend or event

What is Data Mining?

- Definitions:
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data.
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

What is Data Mining?



Also known as Knowledge Discovery in Databases (KDD)

Data Mining: Six Categories

- Anomaly detection
- Association rule learning
- **Classification**
- Clustering
- Regression
- Summarization

Classification (Informal Definition)

- Teach the computer how to classify objects by providing them examples.
- The computer can then classify unseen objects

- Provide some human photographs to the computer and tell them the gender
- Input an unseen photo to the computer. It will tell you if the person is male or female.

Classification (Informal Definition)

- Given a collection of records (training set), each record contains a set of attributes (such as height, weight,…). One of the attributes is the class (Male or female)

- Find a model for class attribute as a function of the values of other attributes. $f(X_1, X_2,\ldots X_n) = \{male, female\}$

- Our goal: Unseen records should be assigned a class as accurate as possible

How to Mine Data Using WEKA

Week 8, Fall 2013

WEKA

The *Weka* or *woodhen* (Gallirallus australis) is an endemic bird of New Zealand. (Source: *WikiPedia*)

WEKA

- Windows X86 with JVM (Download here)
- Windows X86 without JVM (Download here)

- Windows X64 with JVM (Download here)
- Windows X64 without JVM (Download here)

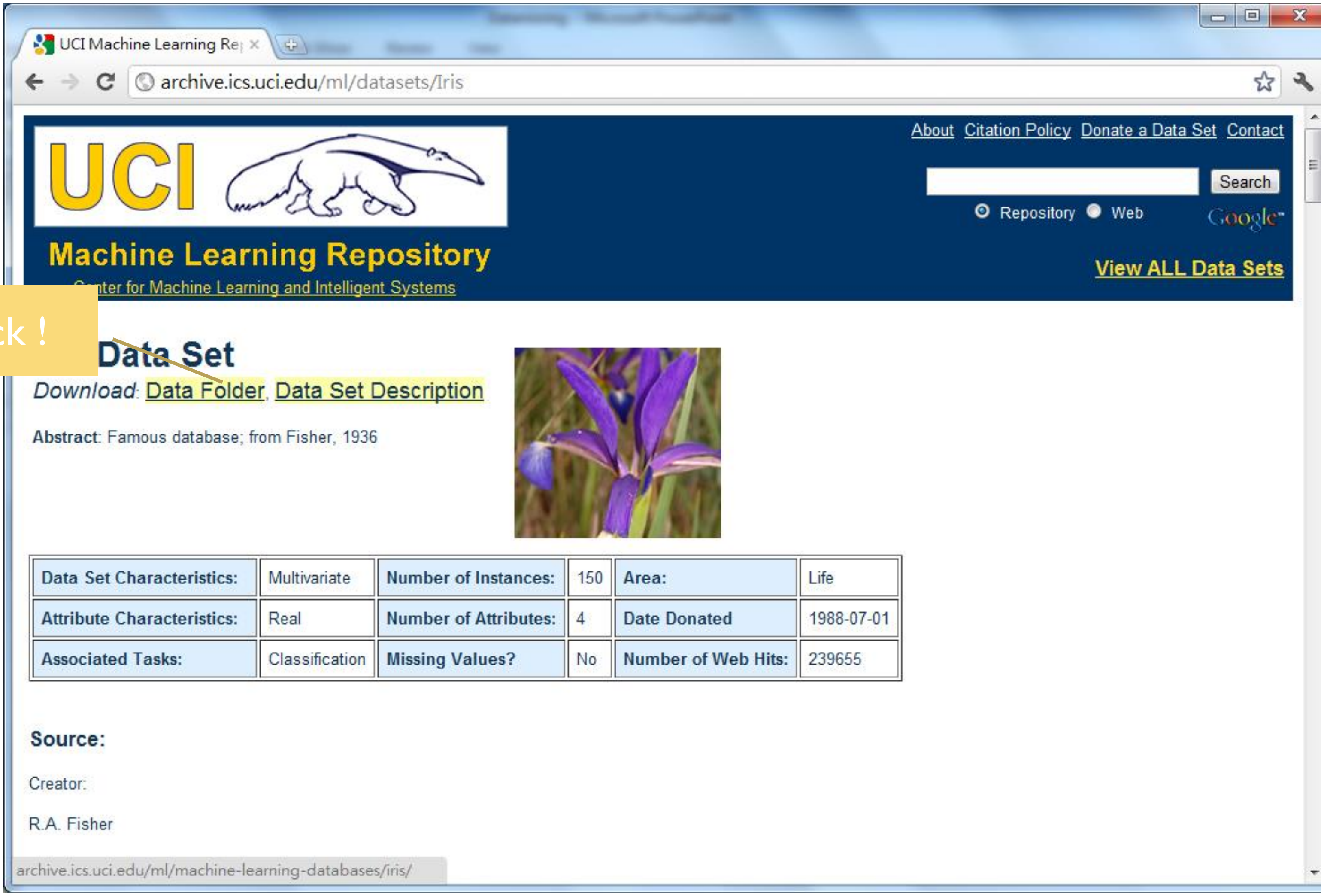
- Mac OS X with JVM (Download here)

- http://www.cs.waikato.ac.nz/ml/weka/

Outline

1. Download the dataset
2. Turn the dataset into ARFF format
3. Build a decision tree
4. Use WEKA to preprocess the data
5. Build a decision tree after preprocessing

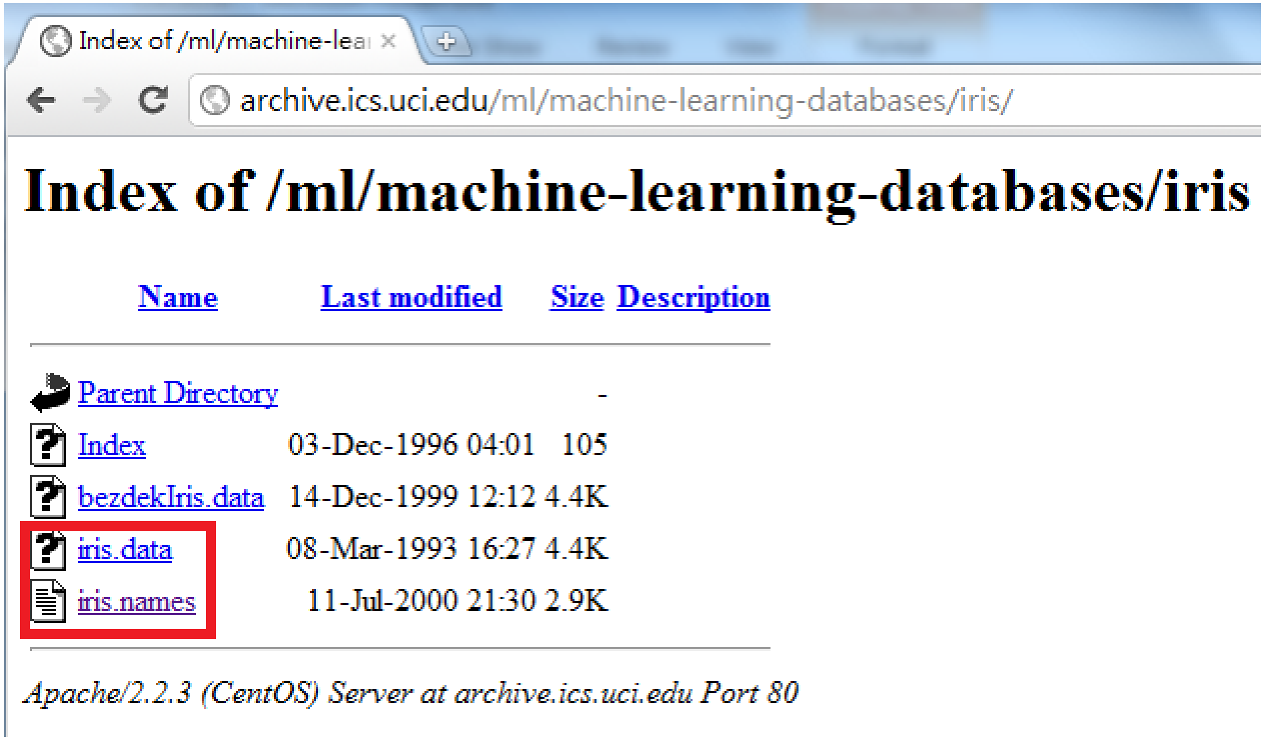
1. Download the dataset



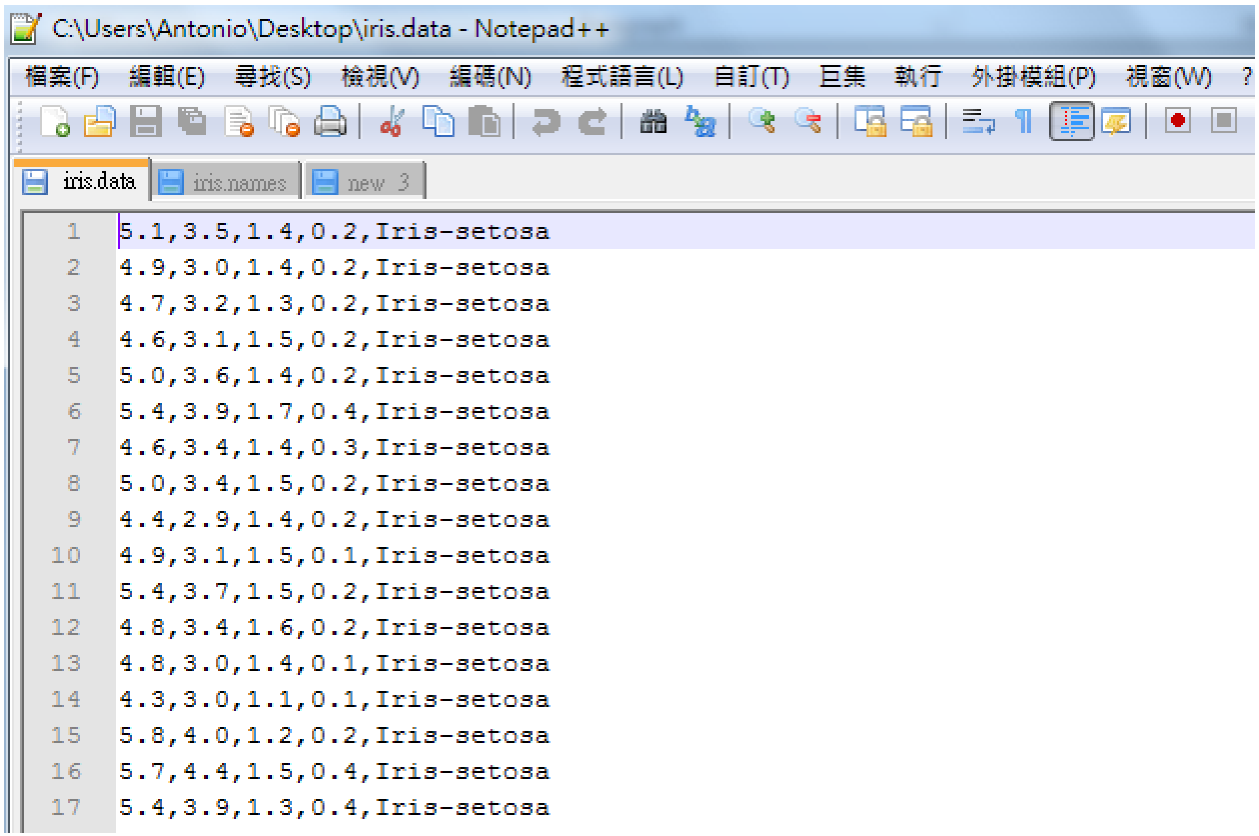
Click !

http://archive.ics.uci.edu/ml/datasets/Iris

1. Download the dataset



1. Download the dataset

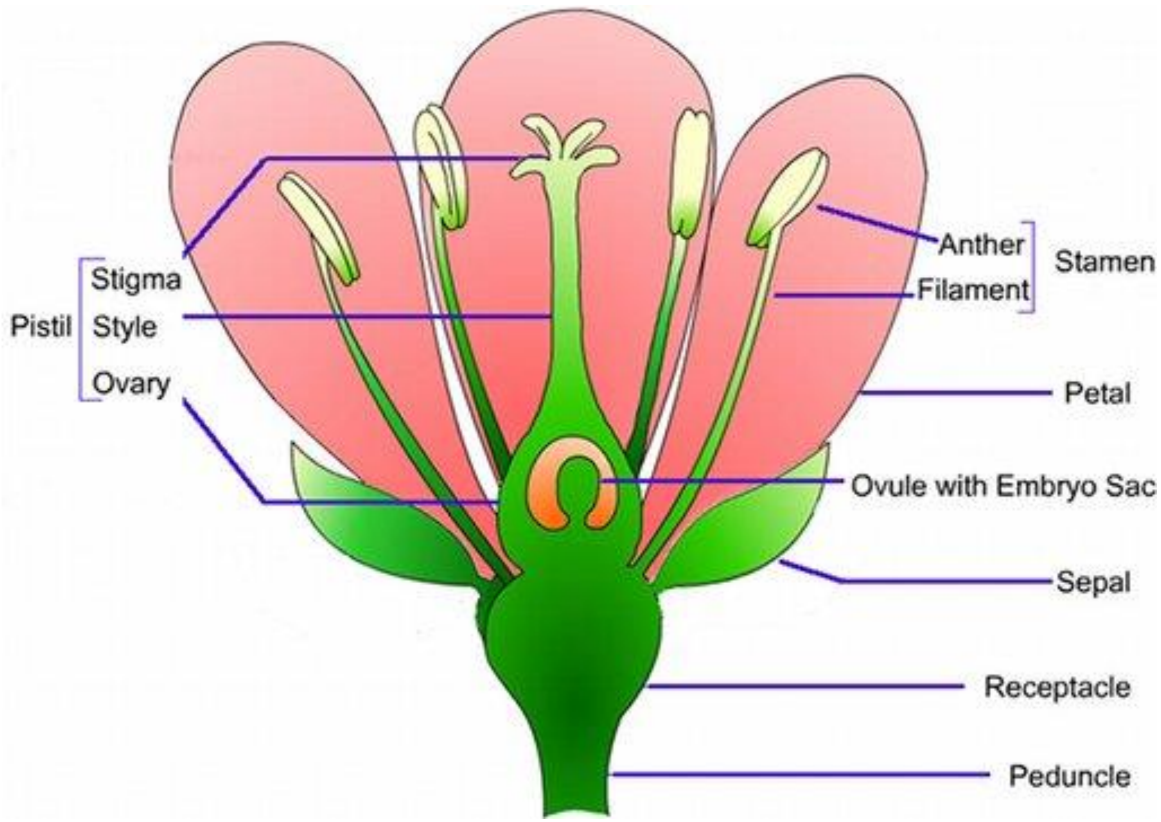


1. Download the dataset

iris.data | iris.names | new 3

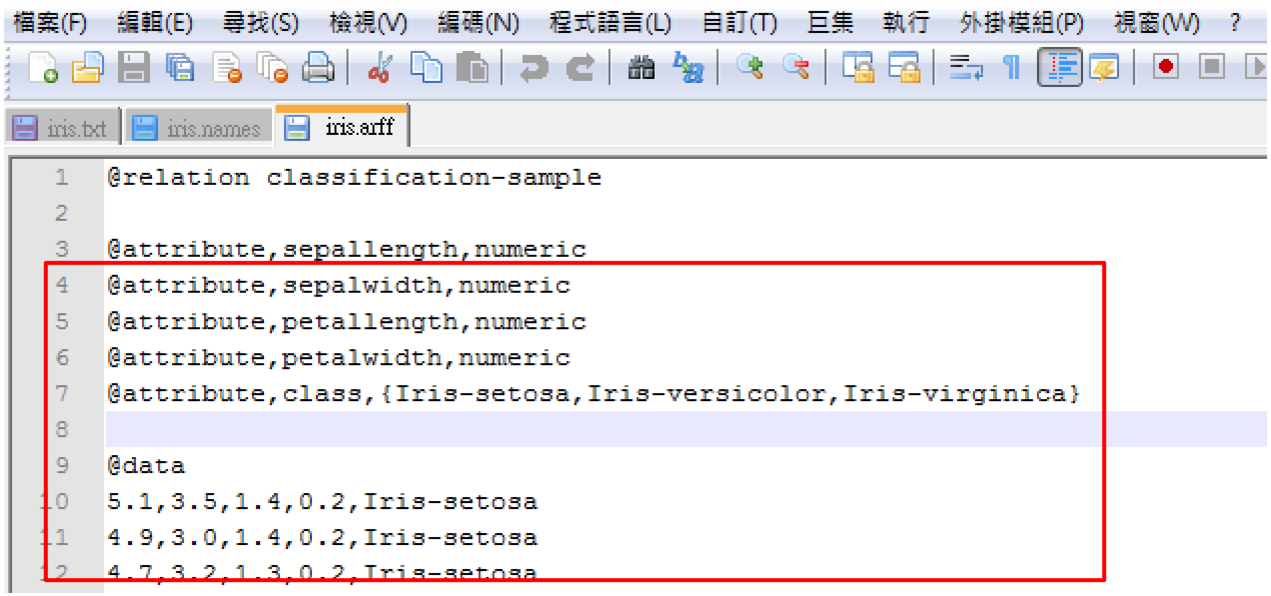
```
    The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa"
    where the errors are in the second and third features.

5. Number of Instances: 150 (50 in each of three classes)

6. Number of Attributes: 4 numeric, predictive attributes and the class

7. Attribute Information:
   1. sepal length in cm
   2. sepal width in cm
   3. petal length in cm
   4. petal width in cm
   5. class:
      -- Iris Setosa
      -- Iris Versicolour
      -- Iris Virginica

8. Missing Attribute Values: None
```

1. Download the dataset




```
7. Attribute Information:
   1. sepal length in cm
   2. sepal width in cm
   3. petal length in cm
   4. petal width in cm
   5. class:
      -- Iris Setosa
      -- Iris Versicolour
      -- Iris Virginica
```

2. Turn the dataset into ARFF format




```
@relation classification-sample

@attribute,sepallength,numeric
@attribute,sepalwidth,numeric
@attribute,petallength,numeric
@attribute,petalwidth,numeric
@attribute,class,{Iris-setosa,Iris-versicolor,Iris-virginica}

@data
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
```

Turn the dataset file into ARFF format before processing

2. Turn the dataset into ARFF format

- @relation classification-sample

- @attribute,sepallength,numeric
- @attribute,sepalwidth,numeric
- @attribute,petallength,numeric
- @attribute,petalwidth,numeric
- @attribute,class,{Iris-setosa,Iris-versicolor,Iris-virginica}

- @data

How to handle missing data ?

The instance data

Each instance is represented on a single line, with carriage returns denoting the end of the instance.

Attribute values for each instance are delimited by commas. They must appear in the order that they were declared in the header section (i.e. the data corresponding to the nth **@attribute** declaration is always the nth field of the attribute).

Missing values are represented by a single question mark, as in:

```
@data
4.4,?,1.5,?,Iris-setosa
```

Read more:
http://www.cs.waikato.ac.nz/ml/weka/arff.html

3. Build a decision tree

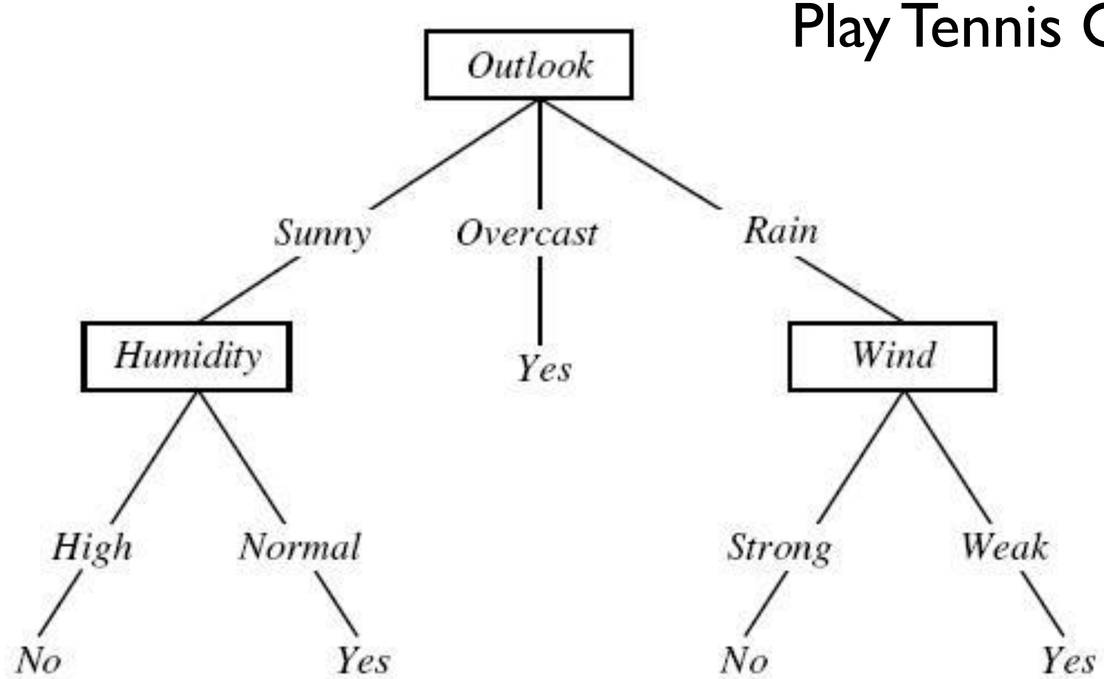
- Classifiers in WEKA are models for predicting nominal or numeric quantities

- Implemented learning schemes include:
  - Decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptron, logistic regression, Bayes' nets, …

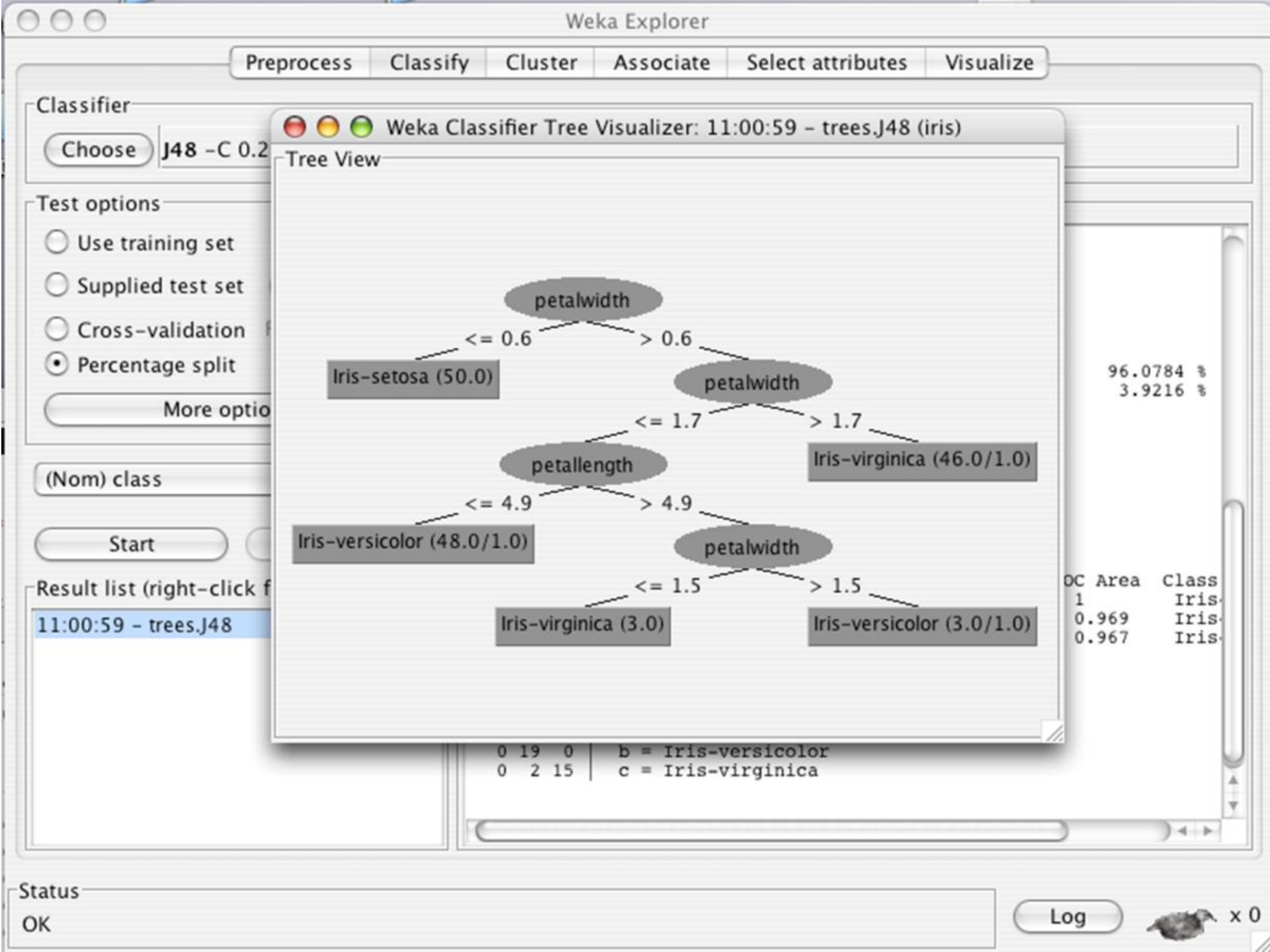
What is a decision tree?

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

Play Tennis Or Not?



3. Build a decision tree


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.2
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
More optio
(Nom) class
Start
Result list (right-click f
11:00:59 - trees.J48
Weka Classifier Tree Visualizer: 11:00:59 - trees.J48 (iris)
Tree View
petalwidth
<= 0.6
> 0.6
Iris-setosa (50.0)
petalwidth
<= 1.7
> 1.7
Iris-virginica (46.0/1.0)
petallength
<= 4.9
> 4.9
Iris-versicolor (48.0/1.0)
petalwidth
<= 1.5
> 1.5
Iris-virginica (3.0)
Iris-versicolor (3.0/1.0)
96.0784 %
3.9216 %
OC Area Class
1 Iris
0.969 Iris
0.967 Iris
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
Status
OK
Log
x 0

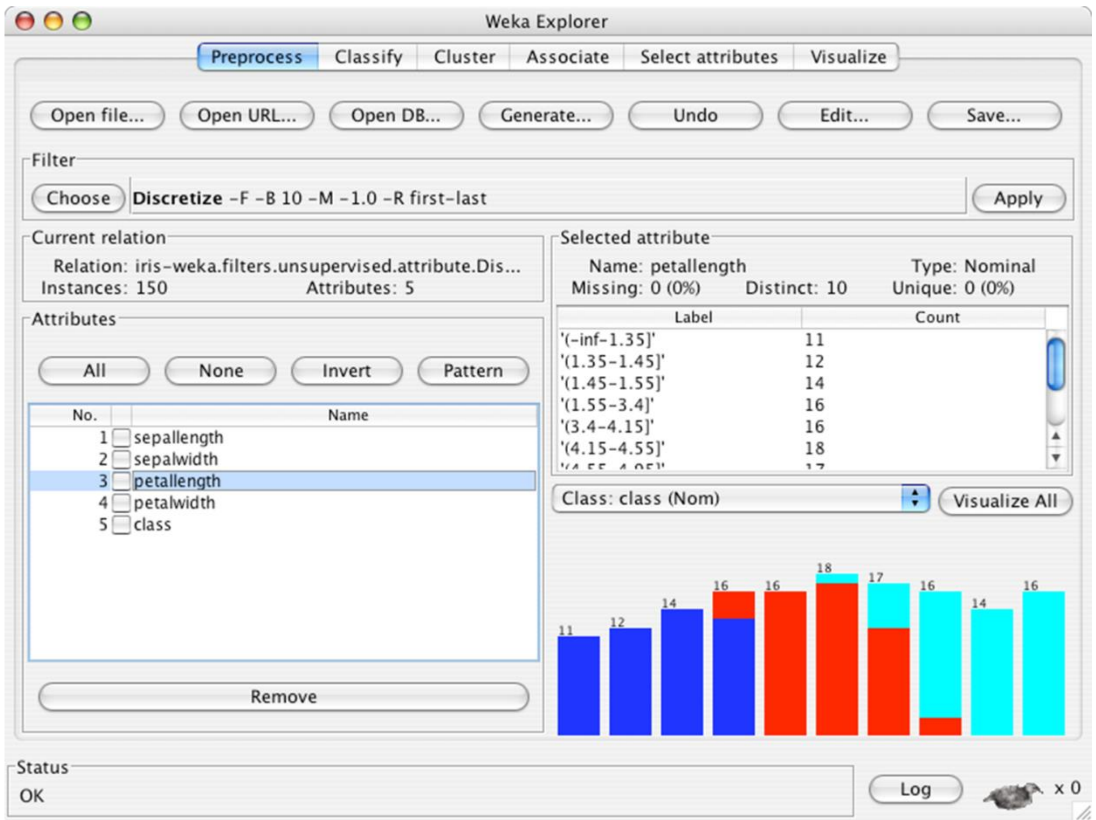
4. Use WEKA to preprocess the data

- Data can be imported from a file in various formats: ARFF, CSV, C4.5, binary or a URL or SQL Database

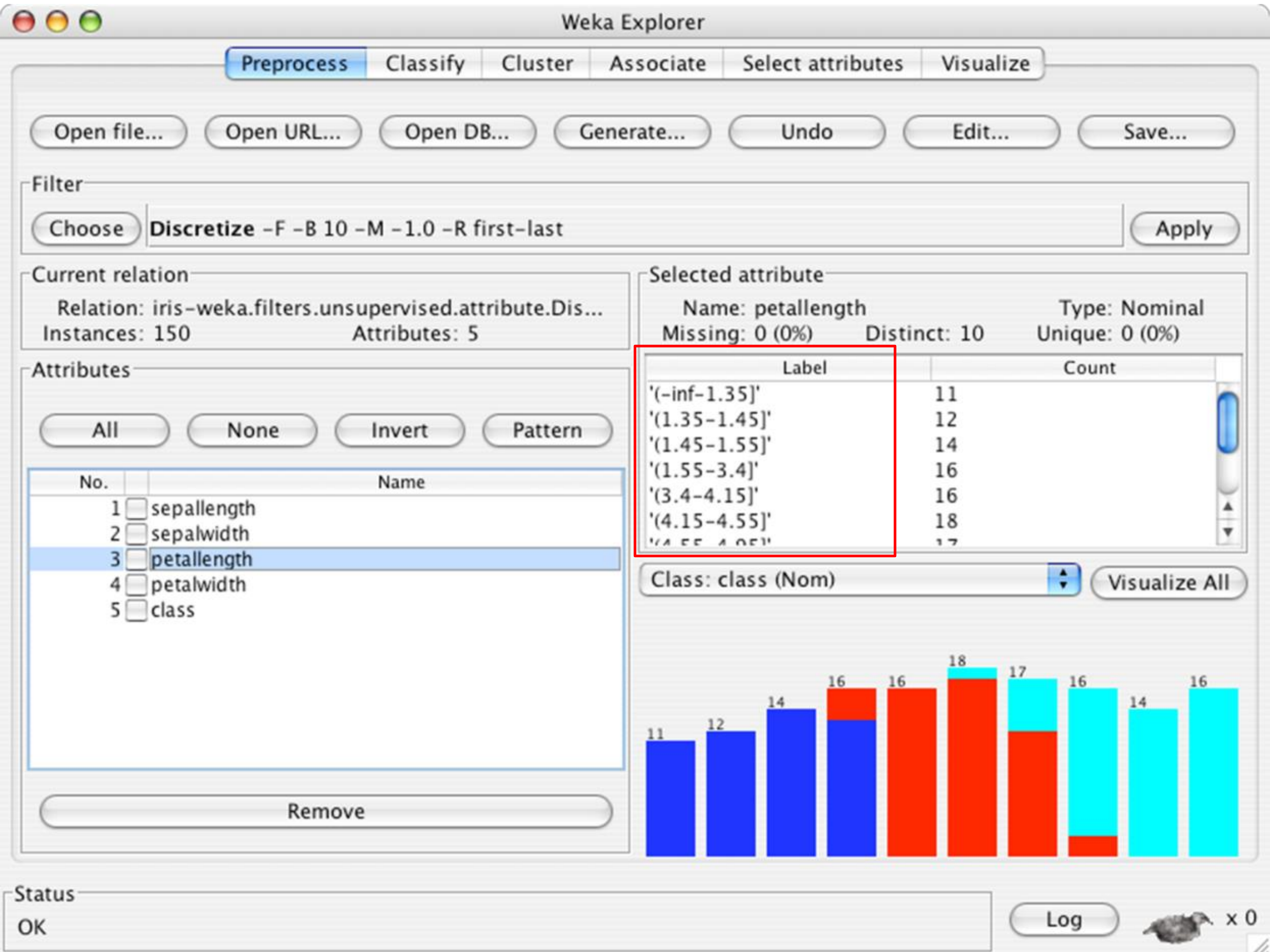
- Pre-processing tools in WEKA are called 'filters'

- WEKA contains filters for: <u>discretization</u>, normalization, resampling, attribute selection, transforming and combining attributes, …

4. Use WEKA to preprocess the data

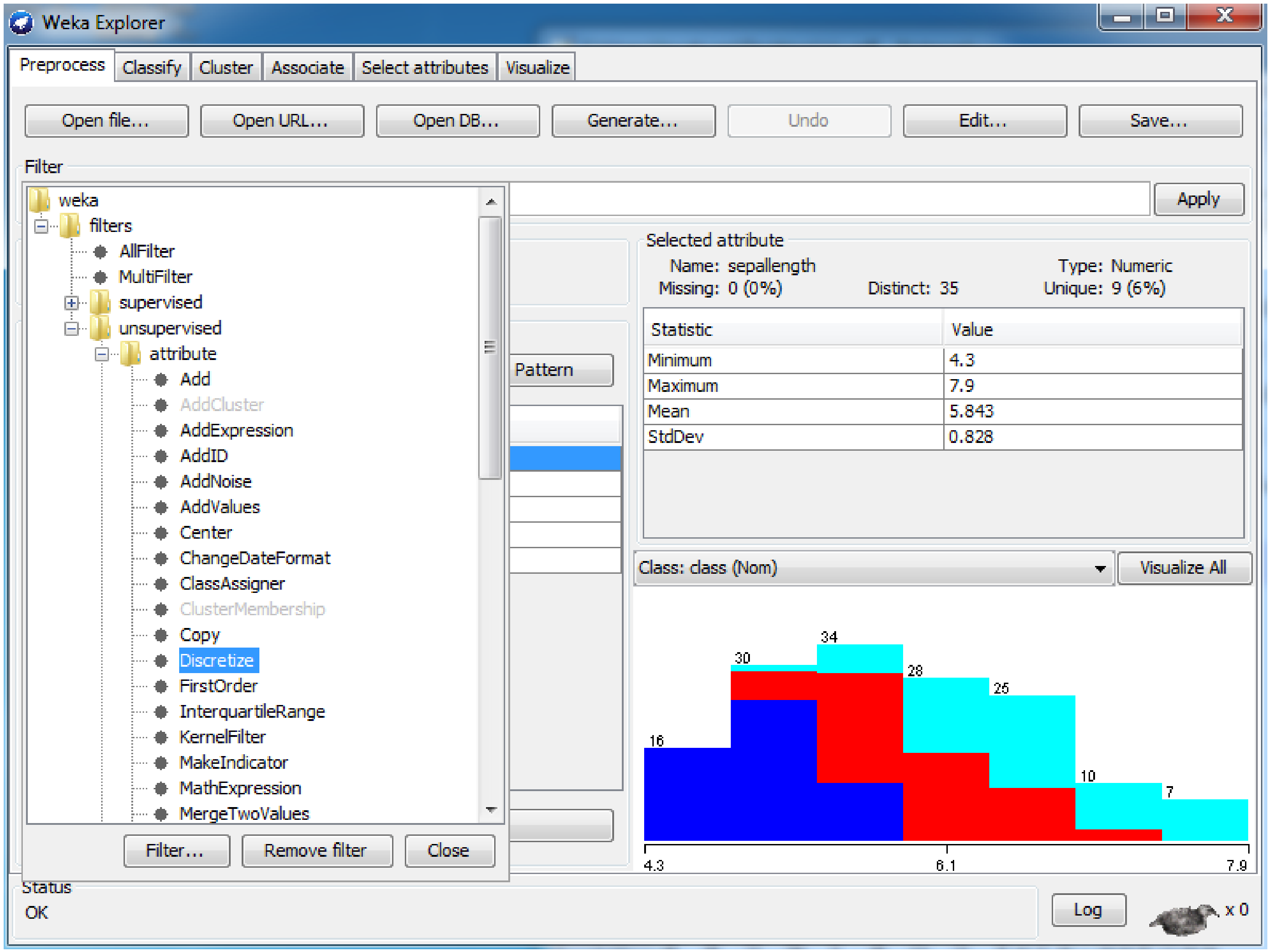


4. Data Preprocessing

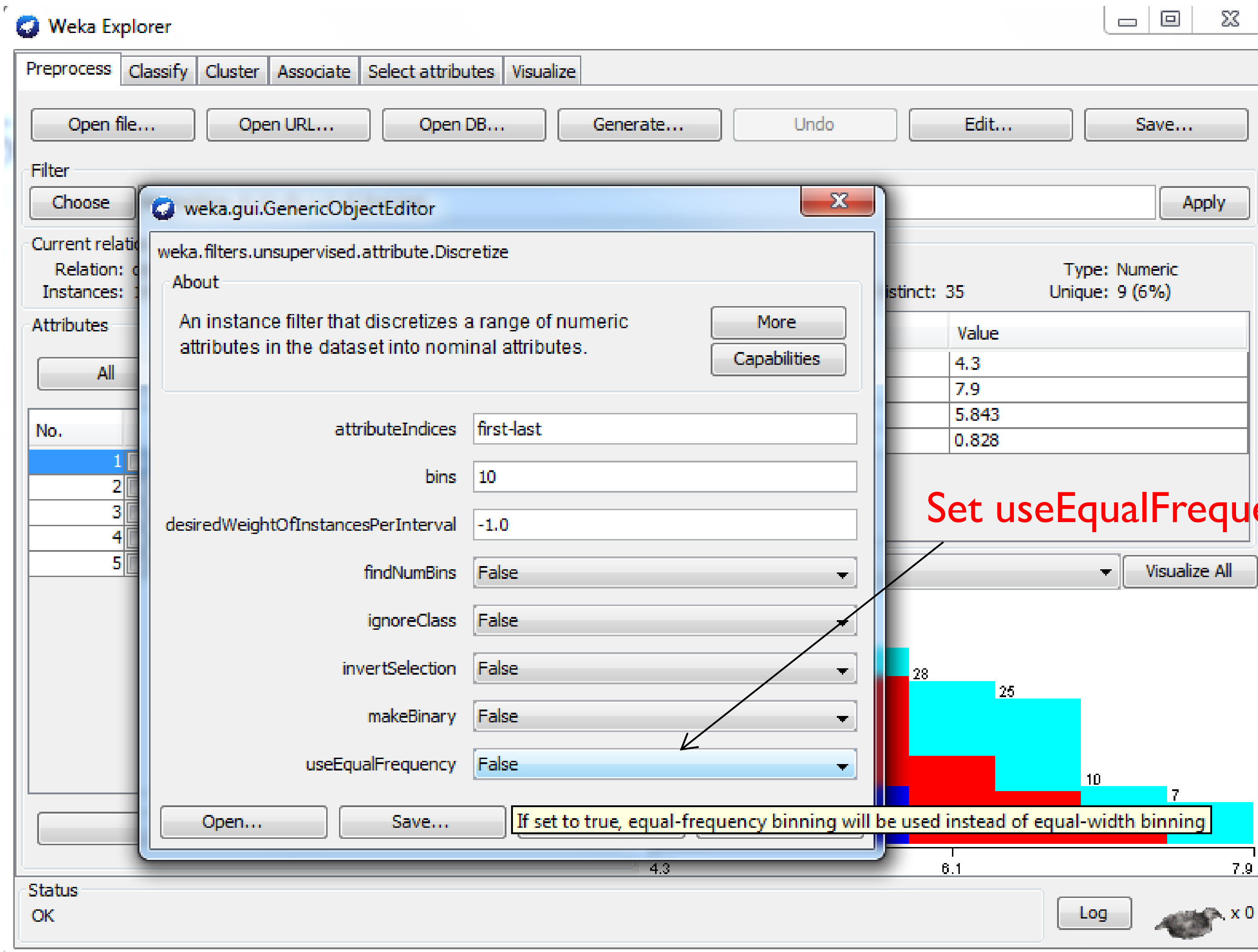


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
Discretize -F -B 10 -M -1.0 -R first-last
Apply
Current relation
Relation: iris-weka.filters.unsupervised.attribute.Dis...
Instances: 150
Attributes: 5
Selected attribute
Name: petallength
Type: Nominal
Missing: 0 (0%)
Distinct: 10
Unique: 0 (0%)
Label
Count
'(-inf-1.35]' 11
'(1.35-1.45]' 12
'(1.45-1.55]' 14
'(1.55-3.4]' 16
'(3.4-4.15]' 16
'(4.15-4.55]' 18
Attributes
All
None
Invert
Pattern
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Class: class (Nom)
Visualize All
11
12
14
16
16
18
17
16
14
16
Remove
Status
OK
Log
x 0

5. Build a decision tree (with preprocessing)

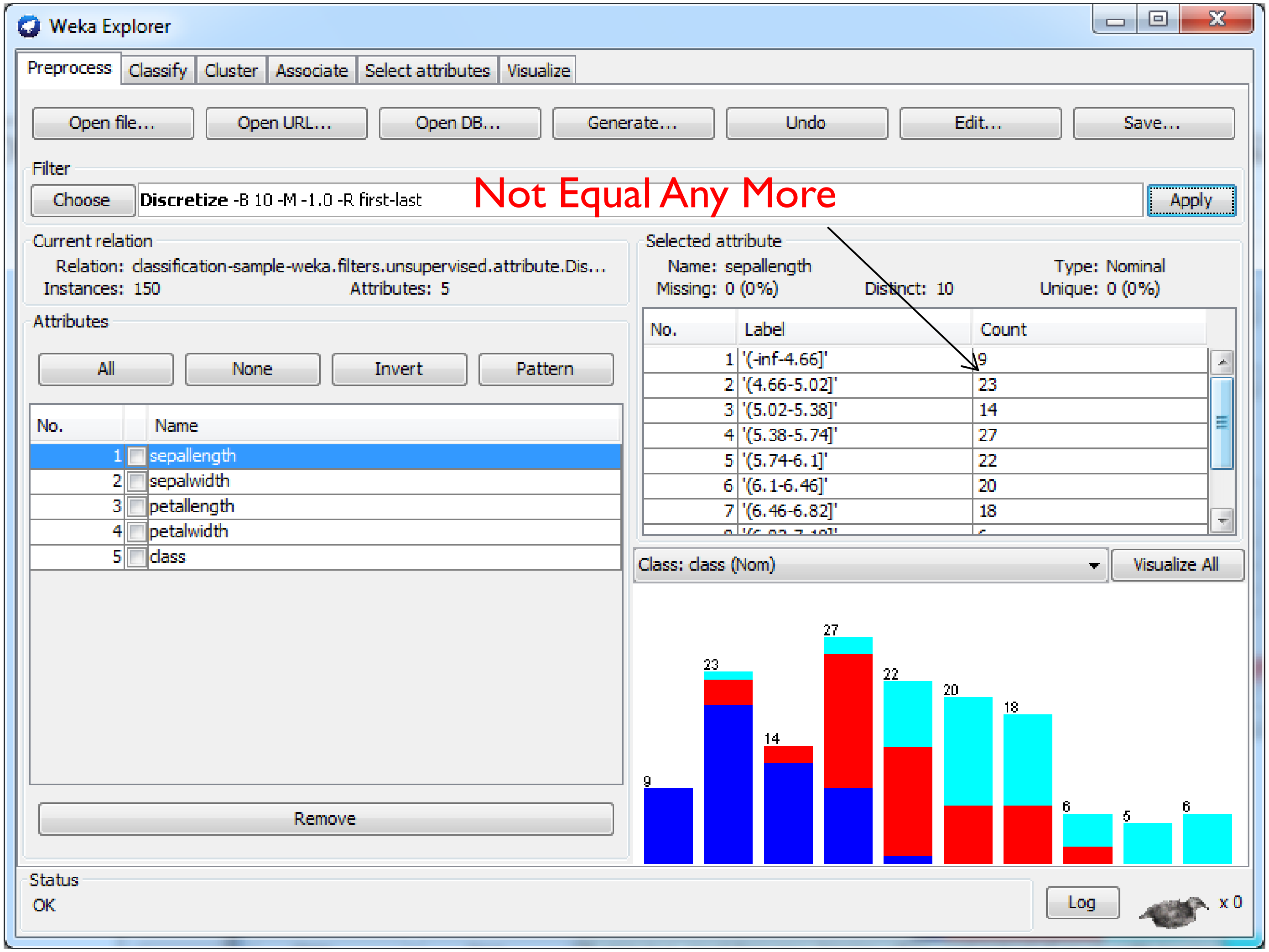


5. Build a decision tree (with preprocessing)

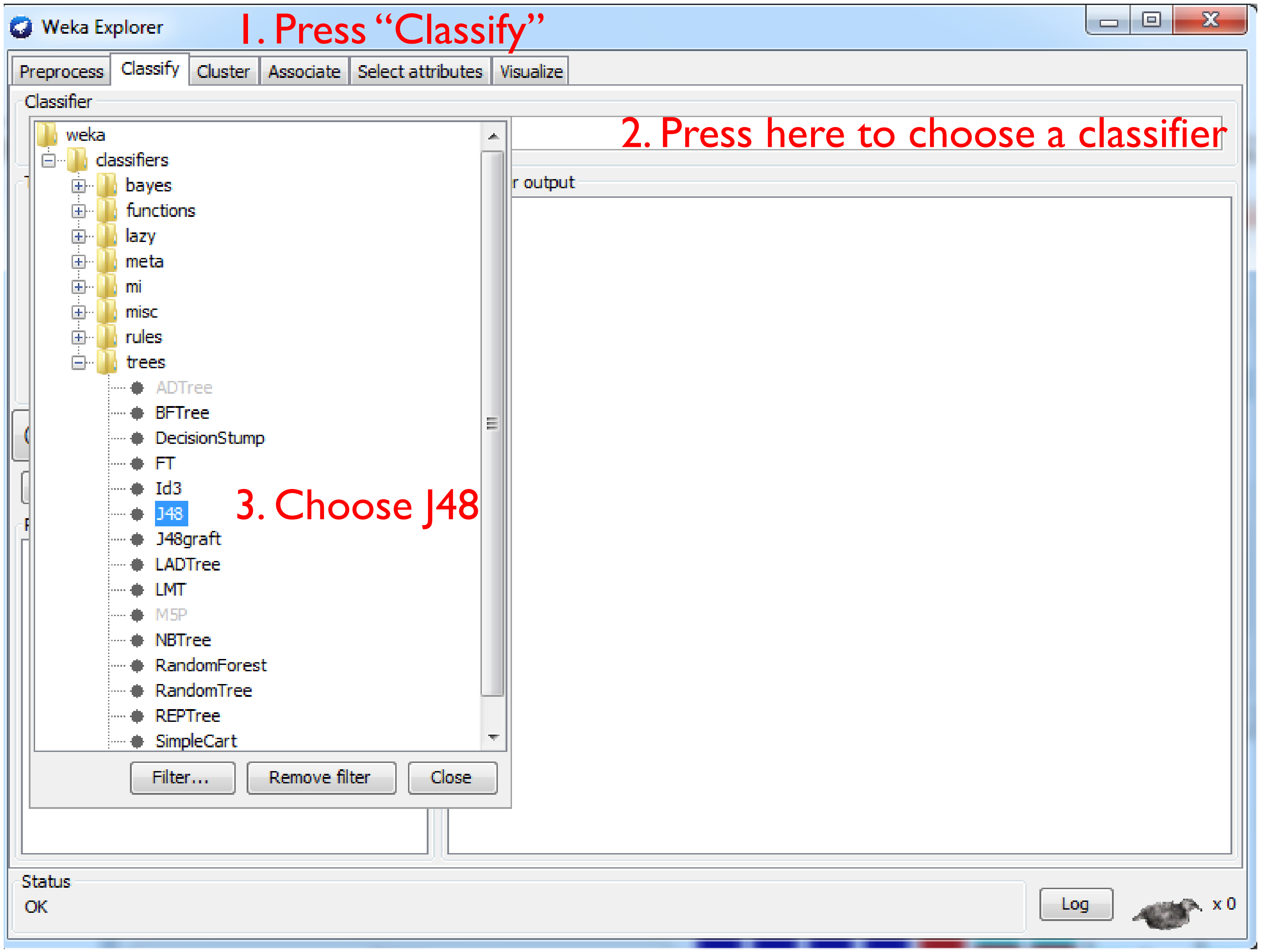


Set useEqualFrequency to False

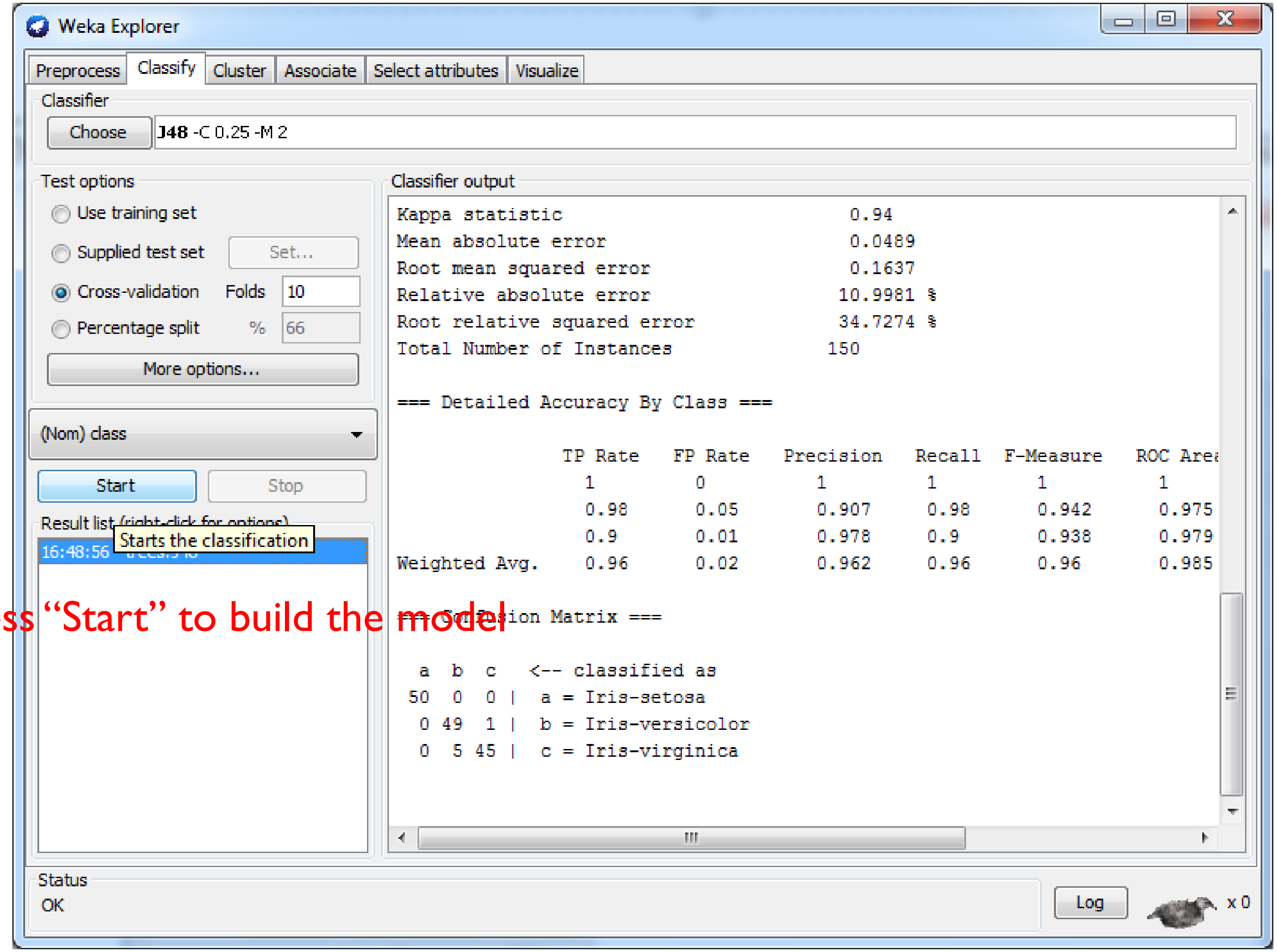
5. Build a decision tree (with preprocessing)



5. Build a decision tree (with preprocessing)

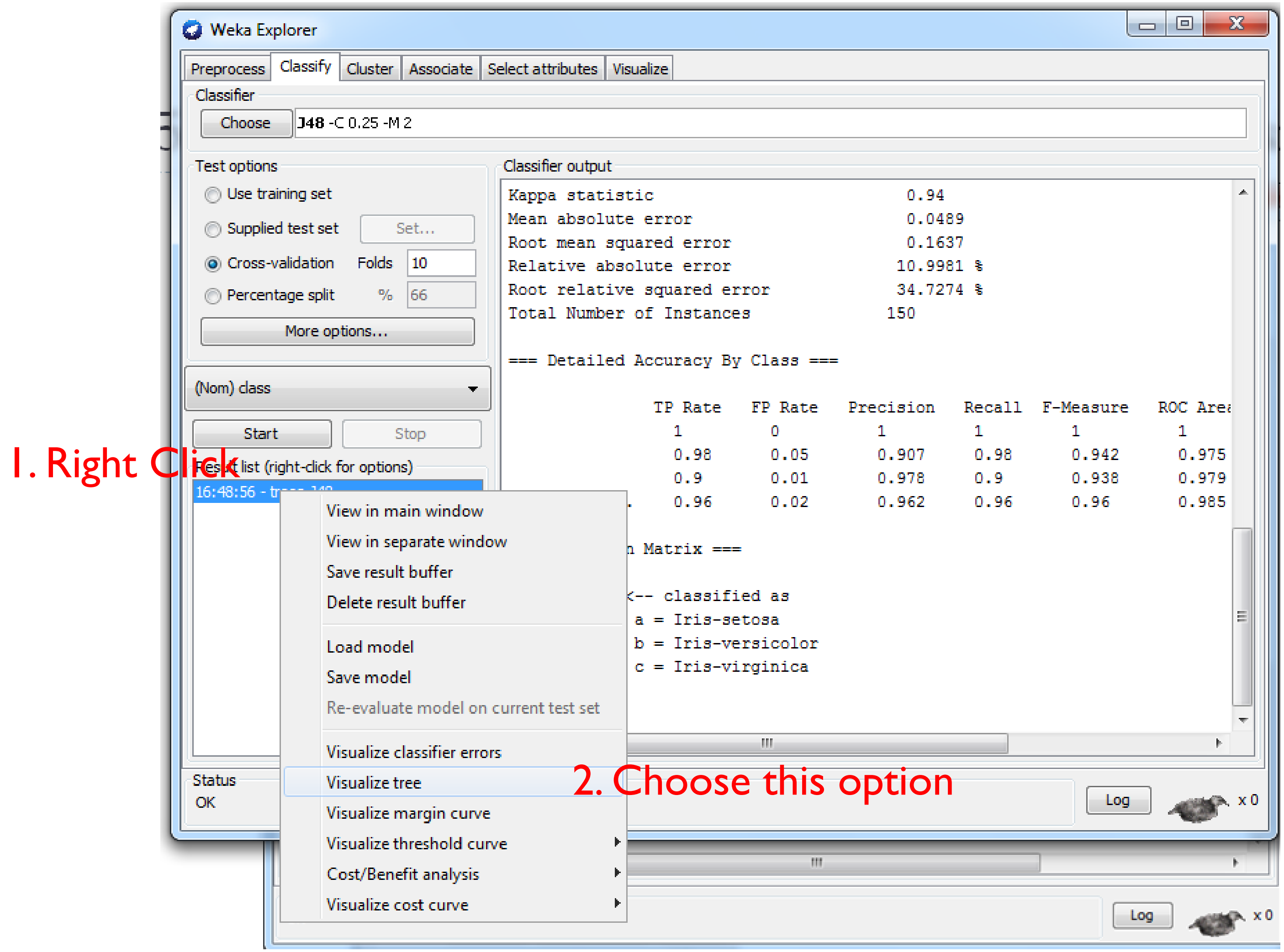


5. Build a decision tree (with preprocessing)

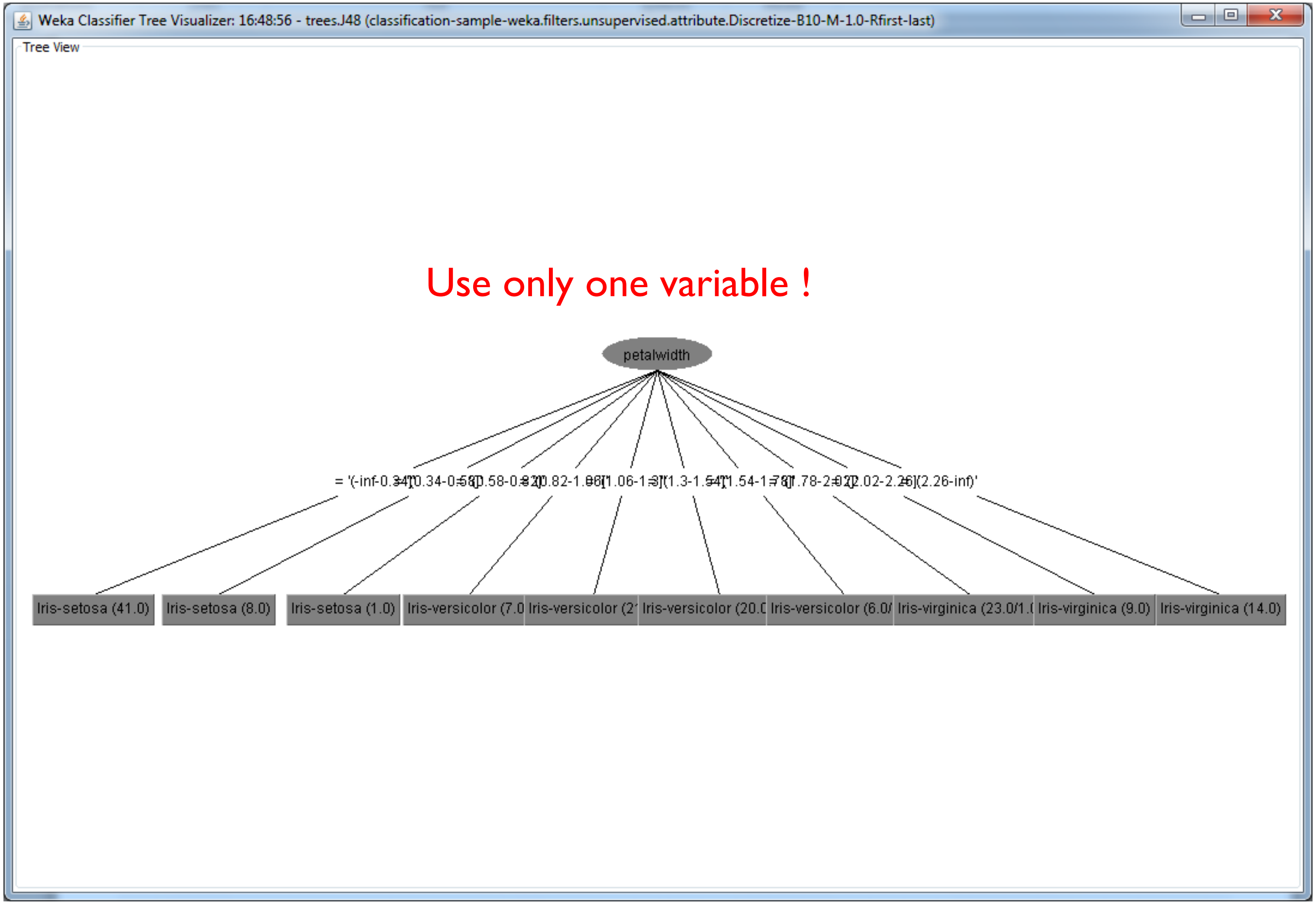


Press "Start" to build the model

5. Build a decision tree (with preprocessing)

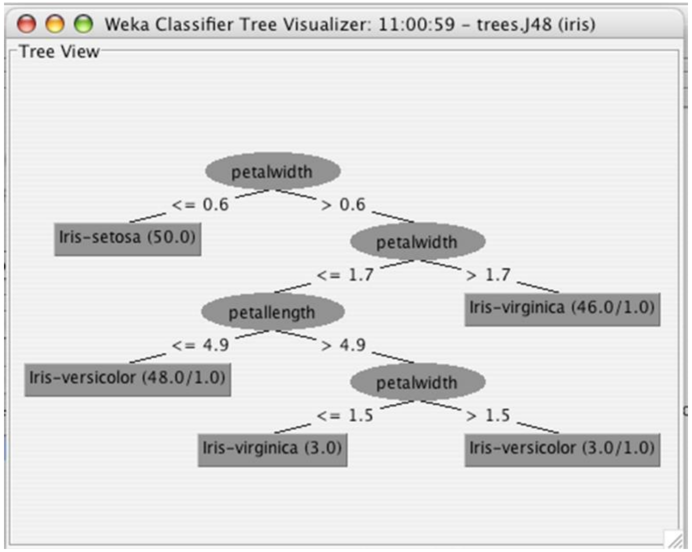


5. Build a decision tree (with preprocessing)

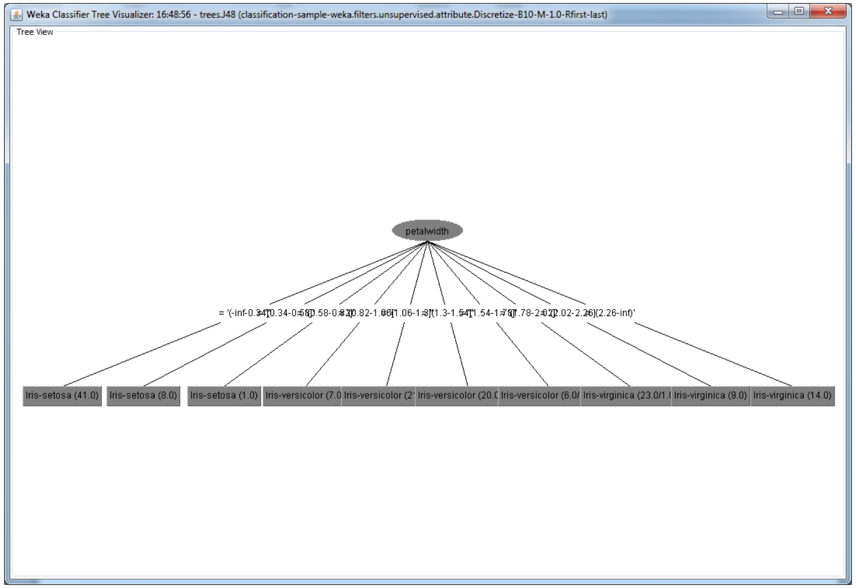


Comparison

Without preprocessing



With Preprocessing



Model Evaluation

- Confusion Matrix
  - TP – True Positive ; FP – False Positive
  - FN – False Negative; TN – True Negative

	Predicted Class		
Actual Class		Class = Yes	Class = No
	Class = Yes	a (TP)	b (FN)
	Class = No	c (FP)	d (TN)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Model Evaluation

- Given a set of records containing positive and negative results, the computer is going to classify the records to be positive or negative.

- Positive: The computer classifies the result to be positive
- Negative: The computer classifies the result to be negative
- True: What the computer classifies is true
- False: What the computer classifies is false

Model Evaluation

- Limitation of Accuracy
  - Consider a 2-class problem
    - Number of Class 0 examples = 9990
    - Number of Class 1 examples = 10
  - If a "stupid" model predicts everything to be class 0, accuracy is 9990/10000 = **99.9** %

- The accuracy is misleading because the model does not detect any example in class 1

Model Evaluation

- Cost-sensitive measures

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

$$\text{Recall (r)} = \frac{TP}{TP+FN} = \frac{a}{a+b}$$

$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

Harmonic mean of Precision and Recall
(Why not just average?)

Model Evaluation

- Given 30 human photographs, a computer predicts 19 to be male, 11 to be female. Among the 19 male predictions, 3 predictions are not correct. Among the 11 female predictions, 1 prediction is not correct.

	Predicted Class		
Actual Class		Male	Female
	Male	a = TP = 16	b = FN = 1
	Female	c = FP = 3	d = TN = 10

Model Evaluation

	Predicted Class		
Actual Class		Male	Female
	Male	a = TP = 16	b = FN = 1
	Female	c = FP = 3	d = TN = 10

- Accuracy = (16 + 10) / (16 + 3 + 1 + 10) = 0.867
- Precision = 16 / (16 + 3) = 0.842
- Recall = 16 / (16 + 1) = 0.941
- F-measure = 2 (0.842)(0.941) / (0.842 + 0.941)

  = 0.889

Discussion

- "In a specific case, precision cannot be computed." Is the statement true? Why?
- If the statement is true, can F-measure be computed in that case?

	a	b	c
a	TP	FN	FN
b	FP	TN	TN
c	FP	TN	TN

←Classified as

a: positive
b: negative
c: negative

- How about if b is positive, a and c are negative, or if c is positive, a and b are negative ?

Next tutorial

- The next tutorial will be a lab session held in 924A/B.
- You are required to finish a data mining task in the laboratory and answer relevant questions asked by tutors.
- This lab task costs 10% of your subject mark.
- No marks will be given to the absent students.
- The lab manual will be released prior to the lab session.
- Please get prepared !

Reference

- Text book:
  - Tan, Steinback, Kumar, "Introduction to Data Mining", Addision Wesley, 2006.
- Datasets
  - UC Irvine Machine Learning Repository