



# 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](mailto:hyszeto@cse.cuhk.edu.hk)
- ▶ Language: Cantonese, English, Mandarin (A little)
- ▶ You are welcome to discuss with me any materials related to CSCI 3230.



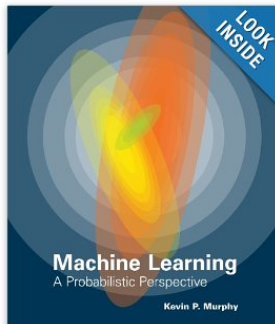
# Data Mining – Business Application: Consumer Behavior Discovery

---

- ▶ 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



## Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series) Hardcover

by Kevin P. Murphy (Author)

★★★★☆ 22 customer reviews

See all 2 formats and editions

Kindle  
\$53.49

Hardcover  
**\$81.00**

13 Used from \$70.44  
40 New from \$68.99

Today's Web-enabled deluge of electronic data calls for automated methods of data analysis. Machine learning provides these, developing methods that can automatically detect patterns in data and then use the uncovered patterns to predict future data. This textbook offers a comprehensive and self-contained introduction to the field of machine learning, based on a unified, probabilistic approach. The coverage combines breadth and depth, offering necessary background material on such topics as

[Read more](#)

Share    

### Frequently Bought Together



Price for both: **\$144.72**

[Add both to Cart](#) [Add both to Wish List](#)

[Show availability and shipping details](#)

☒ **This item:** Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series) by Kevin P. Murphy Hardcover **\$81.00**

☒ **Pattern Recognition and Machine Learning (Information Science and Statistics)** by Christopher M. Bishop Hardcover **\$63.72**

### Customers Who Bought This Item Also Bought



Foundations of Machine Learning (Adaptive ...



Bayesian Reasoning and Machine Learning



Pattern Recognition and Machine Learning ...



Probabilistic Graphical Models: Principles ...



Learning From Data  
Yaser S. Abu-Mostafa



The Elements of Statistical Learning: ...



Boosting: Foundations and Algorithms ...



Convex Optimization  
Stephen Boyd



Buy New **\$81.00**  
Qty: 1 List Price: \$90.00  
Save: \$9.00 (10%)

**FREE Shipping**

**In Stock.**

Ships from and sold by Amazon.com.  
Gift-wrap available.

☐ Yes, I want FREE Two-Day Shipping with Amazon Prime

[Add to Cart](#)

[Sign in to turn on 1-click ordering](#)

**Want it Wednesday, Oct. 23?** Order within **19 hrs 37 mins** and choose **One-Day Shipping** at checkout.  
[Details](#)

[Add to Wish List](#)

Trade in your item  
Get a **\$39.42** Gift Card.

[Trade in](#)

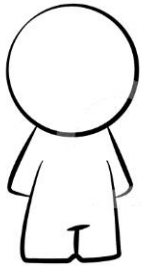
[Learn More](#)

Have one to sell?

[Sell on Amazon](#)

Page 1 of 13

# Data Mining – Biological Application: Genome-Wide Association Study



DNA Sequencing →



ACGT.....AC

220 million long

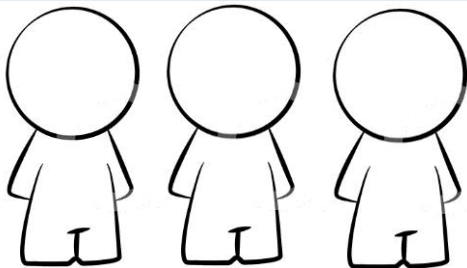
Case (with disease)



DNA Sequencing →



Control (without disease)



DNA Sequencing →



<http://www.sodahead.com>

[http://www.biol.unt.edu/~jajohnson/DNA\\_sequencing\\_process](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?

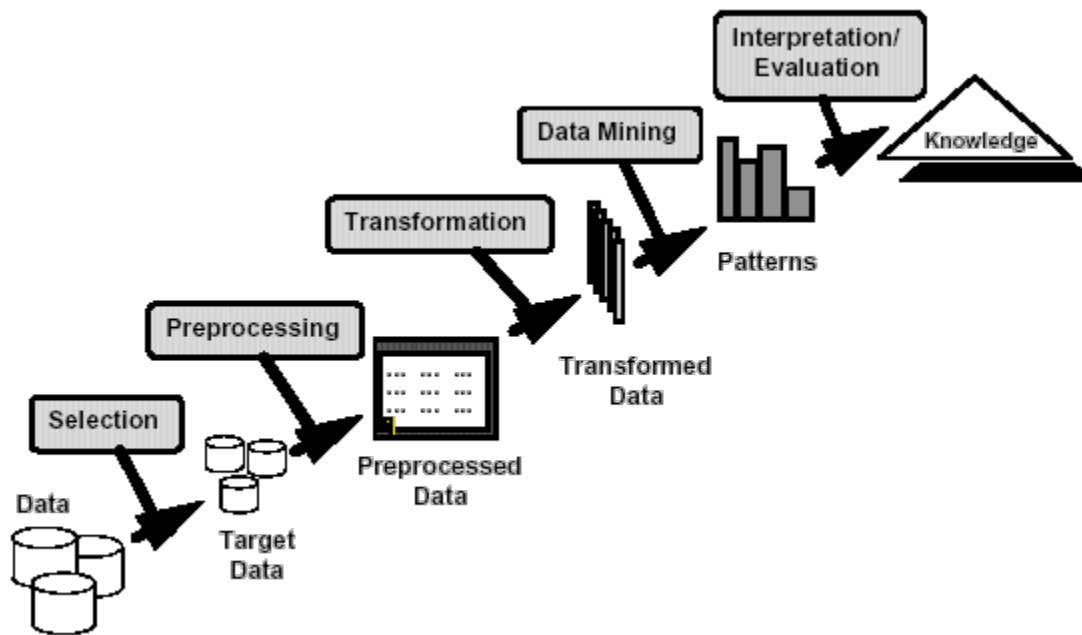
---

## ► 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?

---



# 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.  $f(X_1, X_2, \dots, X_n) = \{\text{male, female}\}$
- ▶ 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/>



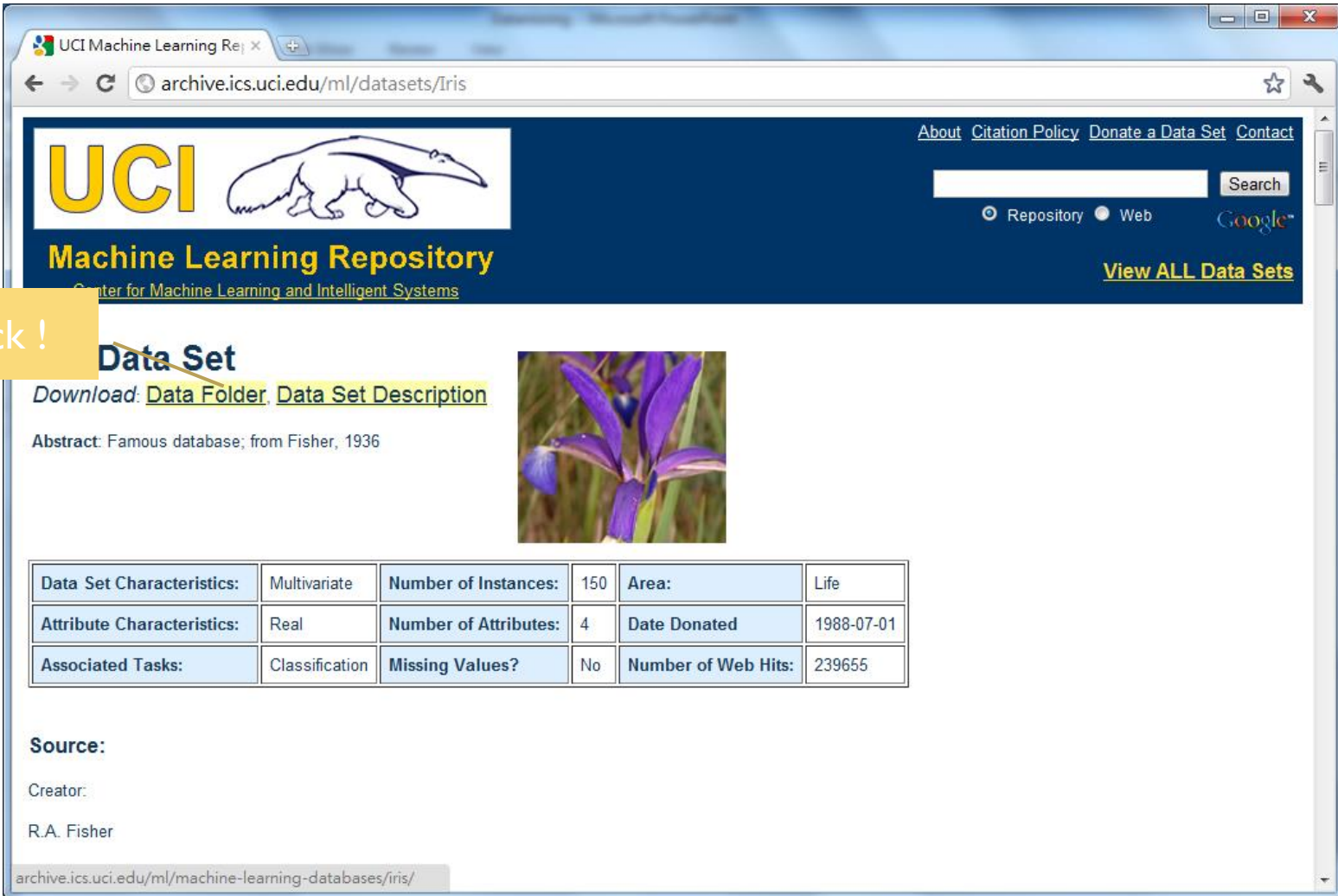
# Outline

---

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 !



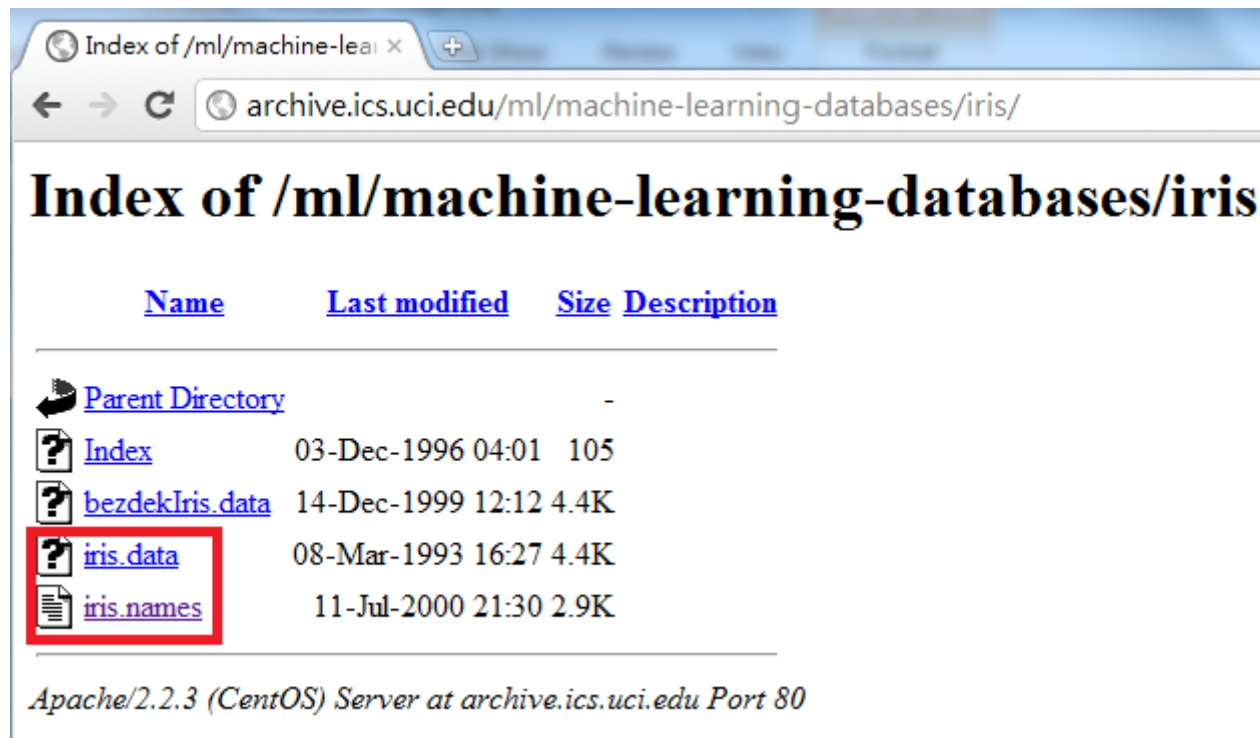
The screenshot shows the UCI Machine Learning Repository website. The header features the UCI logo and a navigation bar with links: About, Citation Policy, Donate a Data Set, and Contact. A search bar is present with a 'Search' button and radio buttons for 'Repository' and 'Web'. A link to 'View ALL Data Sets' is also visible. The main content area is titled 'Data Set' and includes a 'Download' section with links to 'Data Folder' and 'Data Set Description'. An abstract states: 'Famous database; from Fisher, 1936'. A photograph of a purple iris flower is displayed. Below the photo is a table with dataset characteristics.

Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	239655

**Source:**  
Creator:  
R.A. Fisher

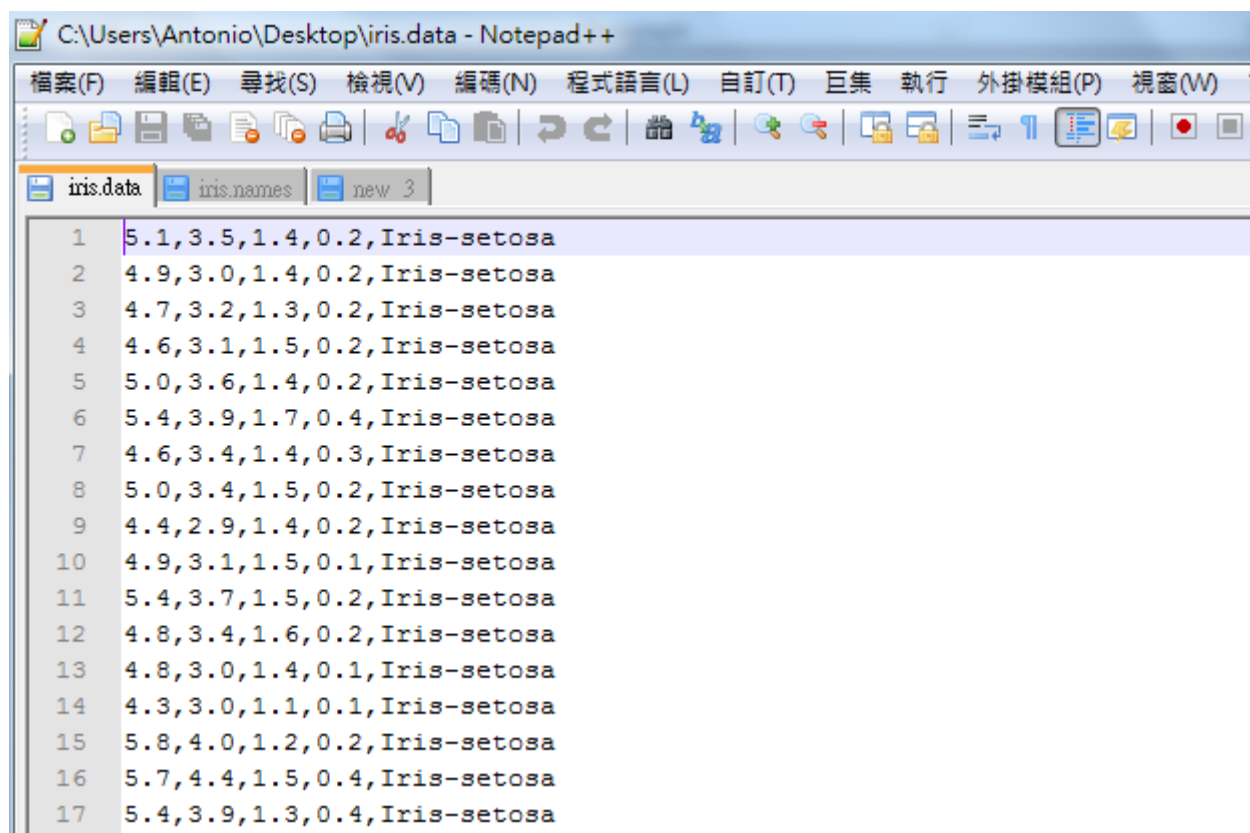
archive.ics.uci.edu/ml/machine-learning-databases/iris/

# 1. Download the dataset



# 1. Download the dataset

---

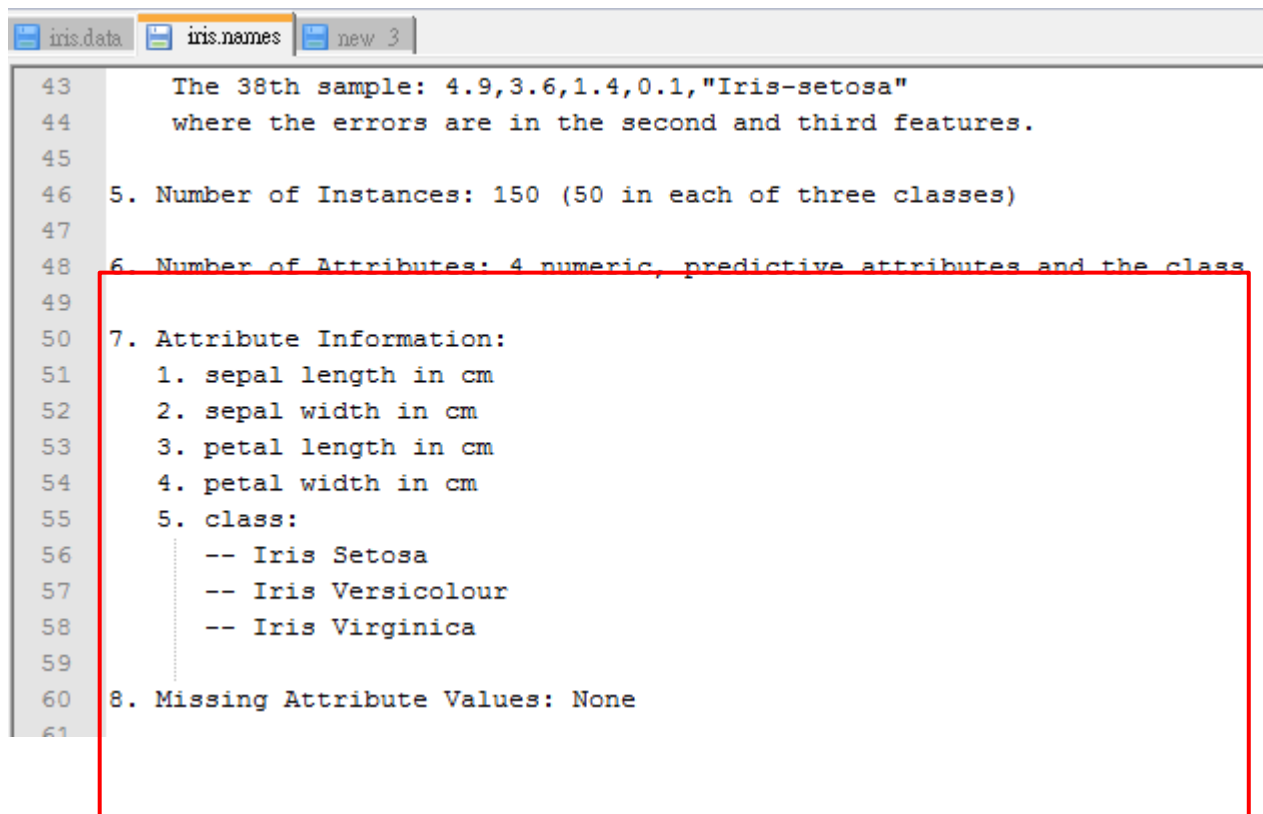


The screenshot shows a Notepad++ window titled "C:\Users\Antonio\Desktop\iris.data - Notepad++". The window contains a list of 17 lines of data, each representing an Iris-setosa flower. The data is formatted as a comma-separated list of five numerical values followed by the species name "Iris-setosa". The lines are numbered 1 through 17 on the left margin.

Line	5.1, 3.5, 1.4, 0.2, Iris-setosa
1	5.1, 3.5, 1.4, 0.2, Iris-setosa
2	4.9, 3.0, 1.4, 0.2, Iris-setosa
3	4.7, 3.2, 1.3, 0.2, Iris-setosa
4	4.6, 3.1, 1.5, 0.2, Iris-setosa
5	5.0, 3.6, 1.4, 0.2, Iris-setosa
6	5.4, 3.9, 1.7, 0.4, Iris-setosa
7	4.6, 3.4, 1.4, 0.3, Iris-setosa
8	5.0, 3.4, 1.5, 0.2, Iris-setosa
9	4.4, 2.9, 1.4, 0.2, Iris-setosa
10	4.9, 3.1, 1.5, 0.1, Iris-setosa
11	5.4, 3.7, 1.5, 0.2, Iris-setosa
12	4.8, 3.4, 1.6, 0.2, Iris-setosa
13	4.8, 3.0, 1.4, 0.1, Iris-setosa
14	4.3, 3.0, 1.1, 0.1, Iris-setosa
15	5.8, 4.0, 1.2, 0.2, Iris-setosa
16	5.7, 4.4, 1.5, 0.4, Iris-setosa
17	5.4, 3.9, 1.3, 0.4, Iris-setosa

# 1. Download the dataset

---

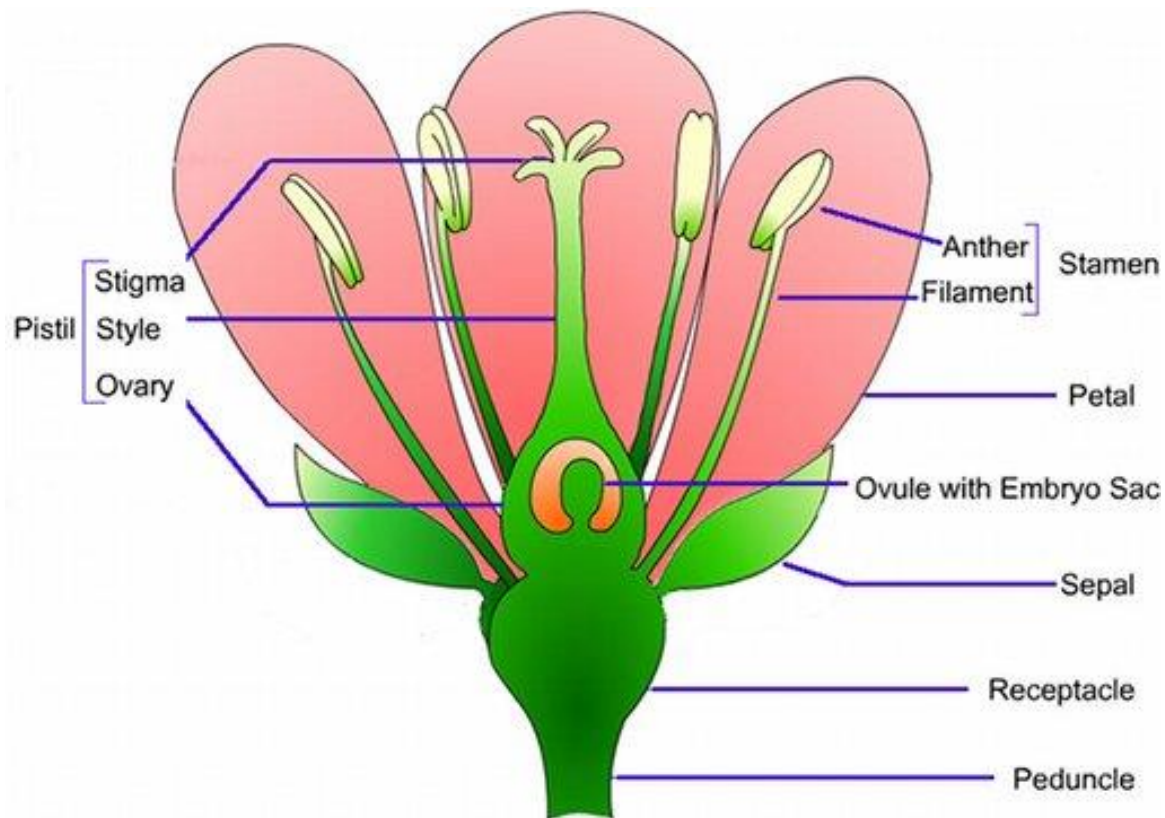


The screenshot shows a Jupyter Notebook interface with three tabs: 'iris.data', 'iris.names', and 'new 3'. The 'iris.names' tab is active. The code cell contains text describing the dataset, with a red rectangular box highlighting the section from line 50 to line 60. The text includes sample information, instance counts, attribute counts, and a detailed list of attributes and classes.

```
43 The 38th sample: 4.9,3.6,1.4,0.1,"Iris-setosa"
44 where the errors are in the second and third features.
45
46 5. Number of Instances: 150 (50 in each of three classes)
47
48 6. Number of Attributes: 4 numeric, predictive attributes and the class
49
50 7. Attribute Information:
51 1. sepal length in cm
52 2. sepal width in cm
53 3. petal length in cm
54 4. petal width in cm
55 5. class:
56 -- Iris Setosa
57 -- Iris Versicolour
58 -- Iris Virginica
59
60 8. Missing Attribute Values: None
61
```

# 1. Download the dataset

---

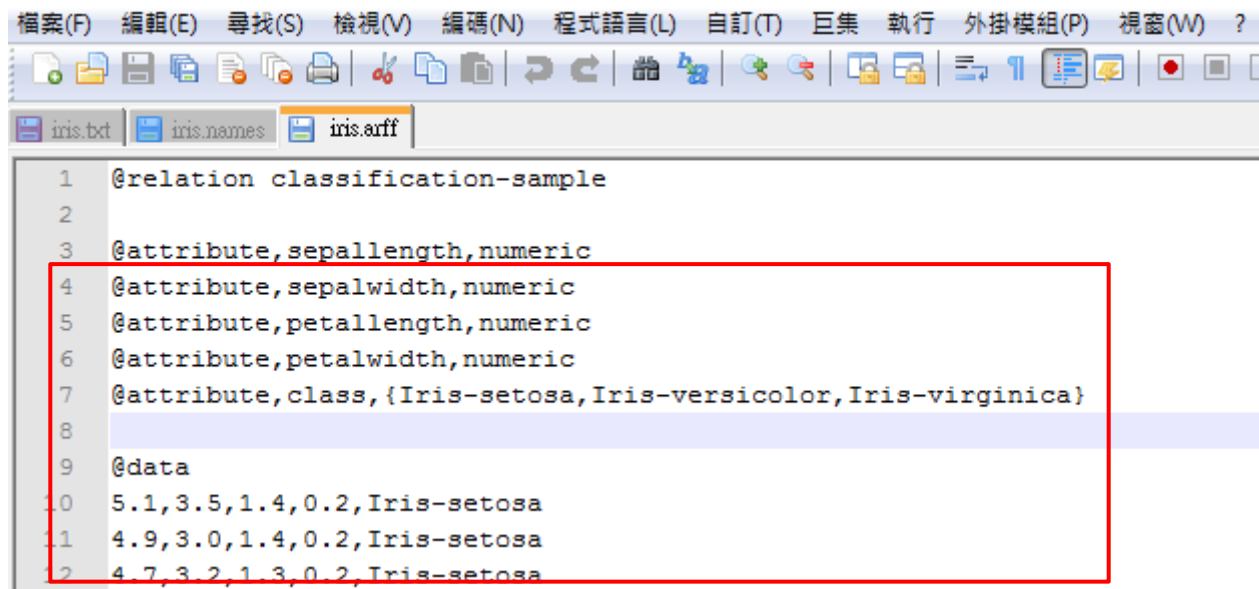


## 7. Attribute Information:

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

---



```
1 @relation classification-sample
2
3 @attribute, sepallength, numeric
4 @attribute, sepalwidth, numeric
5 @attribute, petallength, numeric
6 @attribute, petalwidth, numeric
7 @attribute, class, {Iris-setosa, Iris-versicolor, Iris-virginica}
8
9 @data
10 5.1, 3.5, 1.4, 0.2, Iris-setosa
11 4.9, 3.0, 1.4, 0.2, Iris-setosa
12 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,sepalength,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>

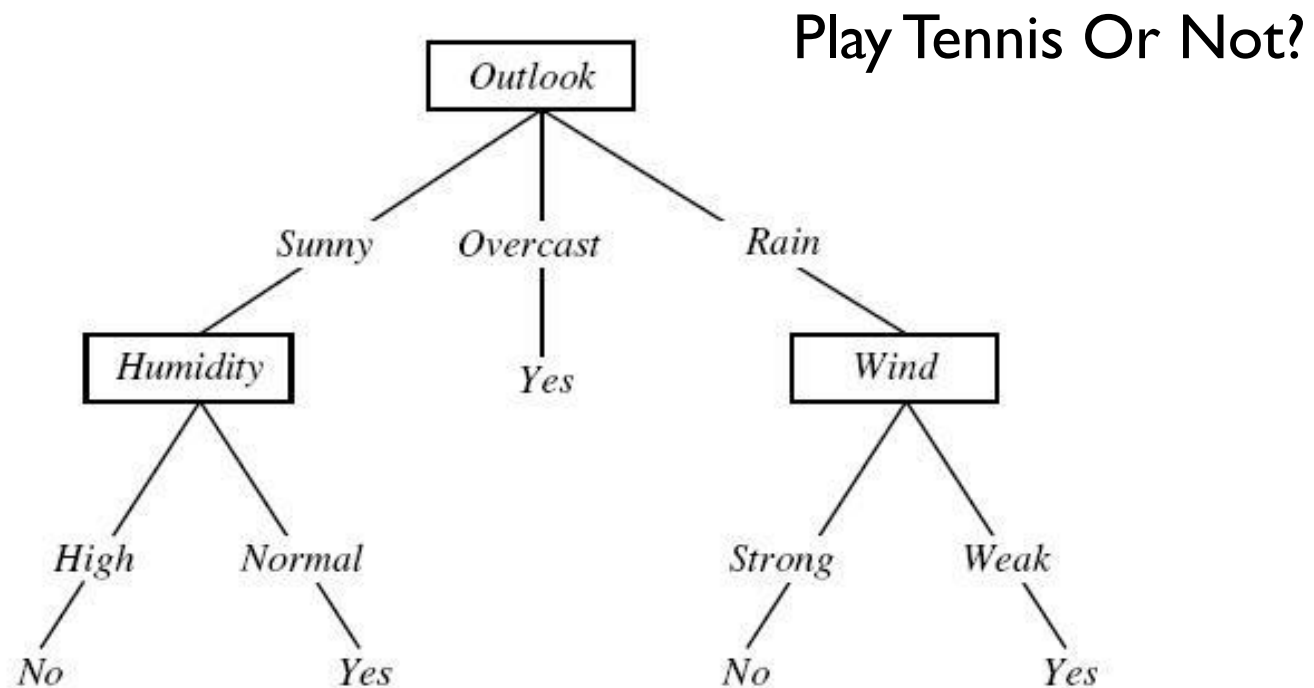
### 3. Build a decision tree

---

- ▶ 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.



### 3. Build a decision tree

---



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 options

(Nom) class

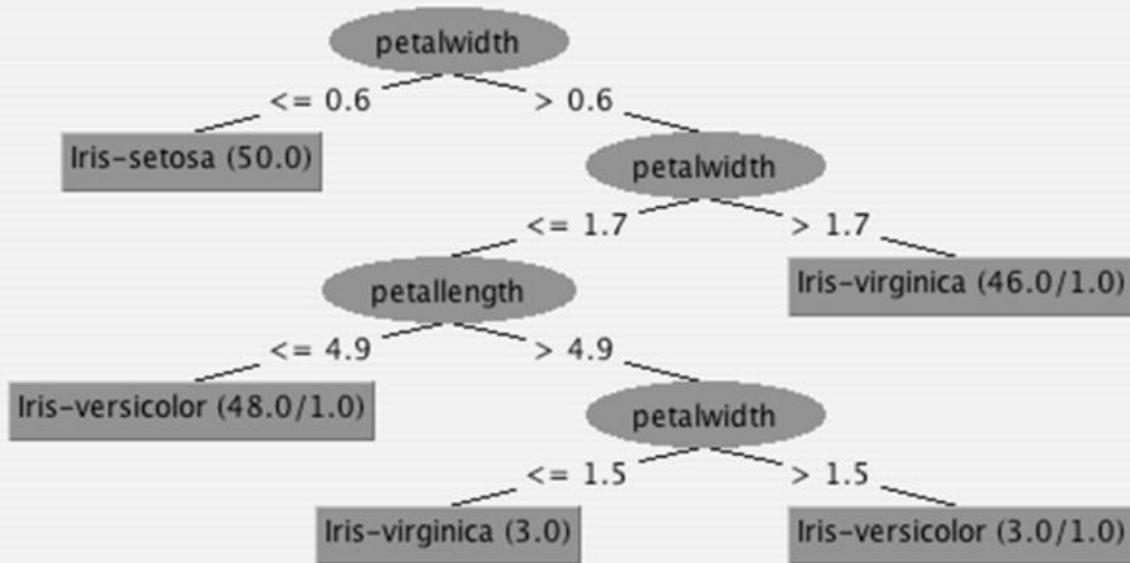
Start

Result list (right-click for details)

11:00:59 - trees.J48

Weka Classifier Tree Visualizer: 11:00:59 - trees.J48 (iris)

Tree View


 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

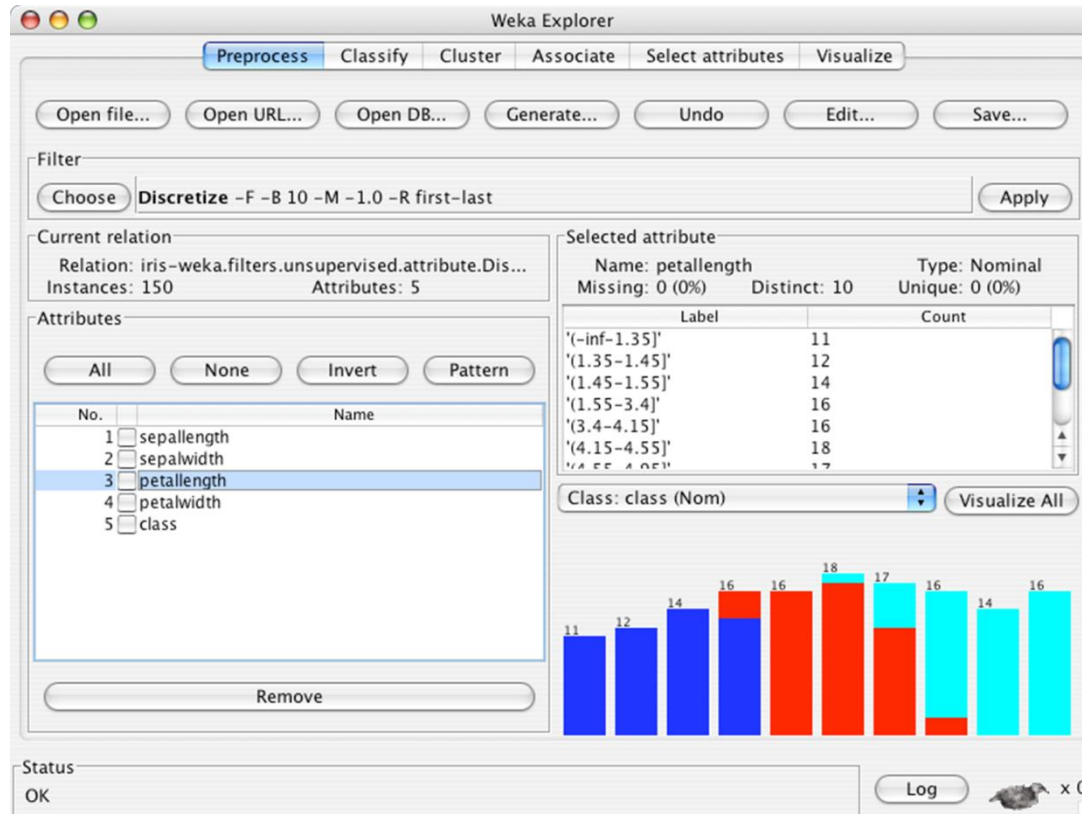
x 0

## 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: discretization, normalization, resampling, attribute selection, transforming and combining attributes, ...

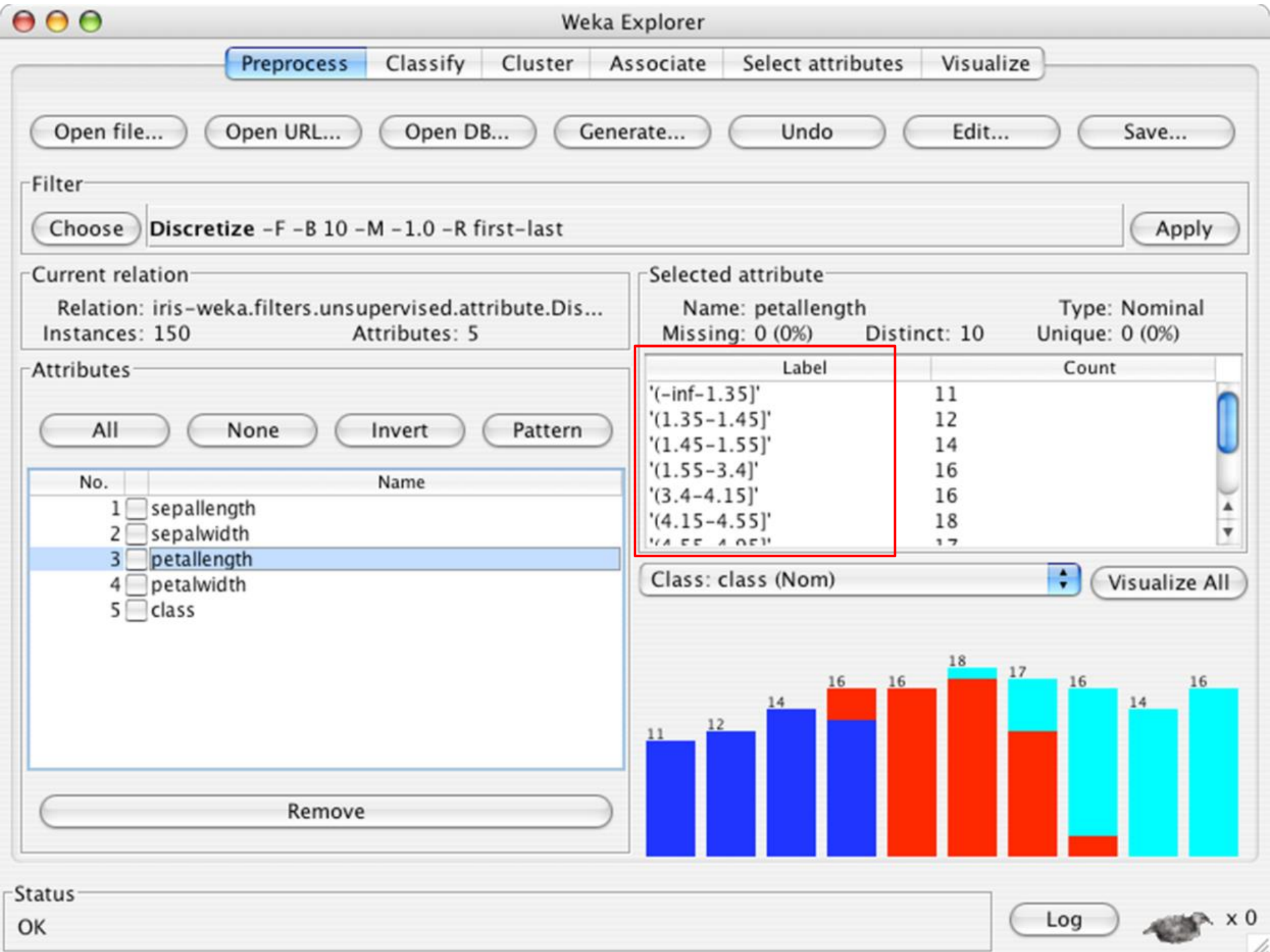
## 4. Use WEKA to preprocess the data



# 4. Data Preprocessing







## 5. Build a decision tree (with preprocessing)

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

- weka
  - filters
    - AllFilter
    - MultiFilter
    - supervised
      - attribute
        - Add
        - AddCluster
        - AddExpression
        - AddID
        - AddNoise
        - AddValues
        - Center
        - ChangeDateFormat
        - ClassAssigner
        - ClusterMembership
        - Copy
        - Discretize
        - FirstOrder
        - InterquartileRange
        - KernelFilter
        - MakeIndicator
        - MathExpression
        - MergeTwoValues

Selected attribute

Name: sepal.length  
Missing: 0 (0%)  
Distinct: 35  
Type: Numeric  
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

Visualize All

16 30 34 28 25 10 7

4.3 6.1 7.9

Status  
OK

Log x 0

## 5. Build a decision tree (with preprocessing)

The screenshot shows the Weka Explorer interface with the 'Discretize' filter configuration dialog open. The dialog is titled 'weka.gui.GenericObjectEditor' and contains the following settings:

- Attribute Indices: first-last
- Bins: 10
- desiredWeightOfInstancesPerInterval: -1.0
- findNumBins: False
- ignoreClass: False
- invertSelection: False
- makeBinary: False
- useEqualFrequency: False

A red arrow points to the 'useEqualFrequency' dropdown menu, which is currently set to 'False'. A text box at the bottom of the dialog states: 'If set to true, equal-frequency binning will be used instead of equal-width binning'.

In the background, a histogram is visible with the following data series:

Value	Frequency
4.3	28
7.9	25
5.843	10
0.828	7

Set useEqualFrequency to False

## 5. Build a decision tree (with preprocessing)

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter  
Choose **Discretize -B 10 -M -1.0 -R first-last** **Not Equal Any More** Apply

Current relation  
Relation: classification-sample-weka.filters.unsupervised.attribute.Dis...  
Instances: 150 Attributes: 5

Attributes  
All None Invert Pattern

No.	Name
1	<input checked="" type="checkbox"/> sepallength
2	<input type="checkbox"/> sepalwidth
3	<input type="checkbox"/> petallength
4	<input type="checkbox"/> petalwidth
5	<input type="checkbox"/> class

Remove

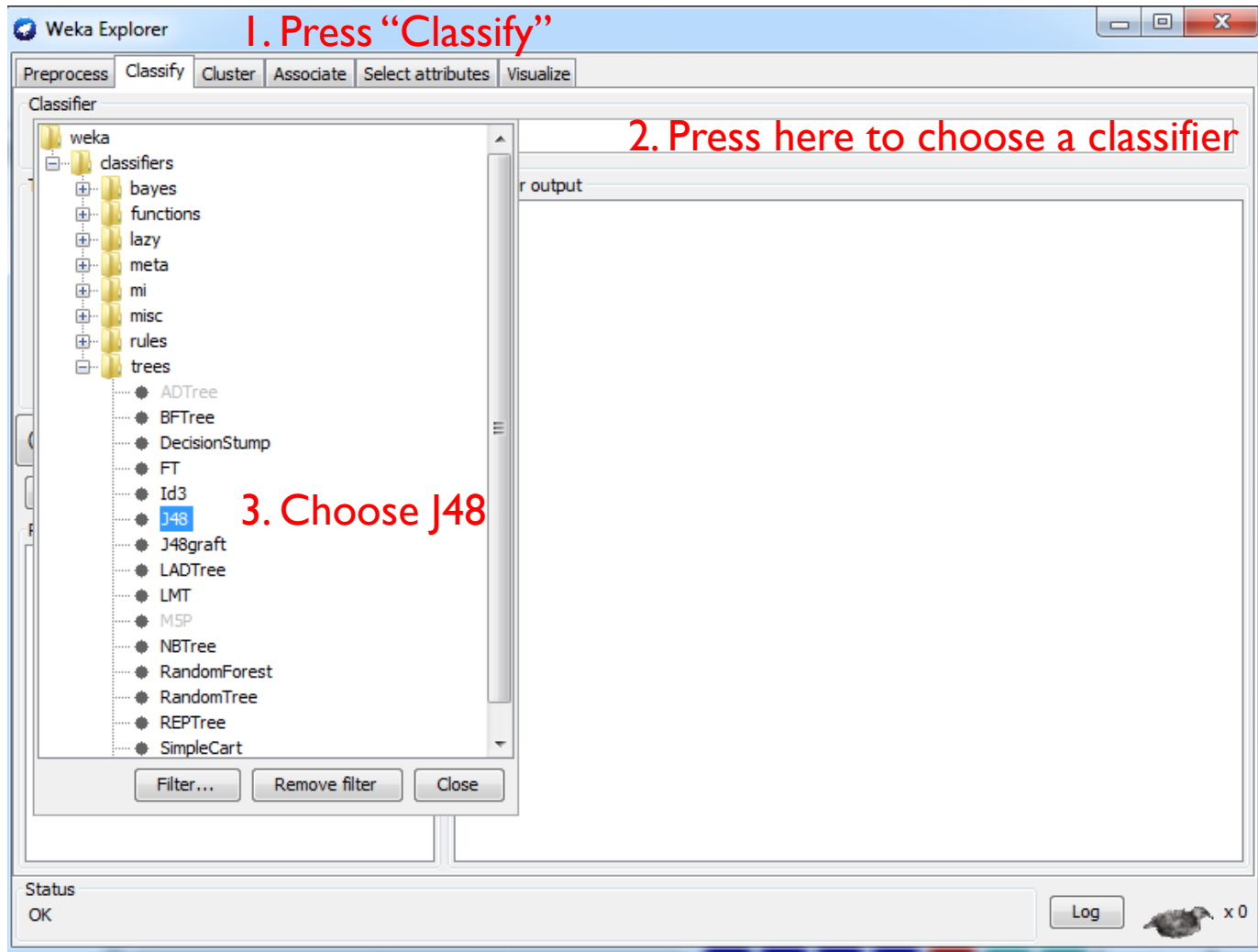
Selected attribute  
Name: sepallength  
Missing: 0 (0%) Distinct: 10 Type: Nominal  
Unique: 0 (0%)

No.	Label	Count
1	'(-inf-4.66]'	9
2	'(4.66-5.02]'	23
3	'(5.02-5.38]'	14
4	'(5.38-5.74]'	27
5	'(5.74-6.1]'	22
6	'(6.1-6.46]'	20
7	'(6.46-6.82]'	18
8	'(6.82-7.18]'	6
9	'(7.18-7.54]'	5
10	'(7.54-7.9]'	6

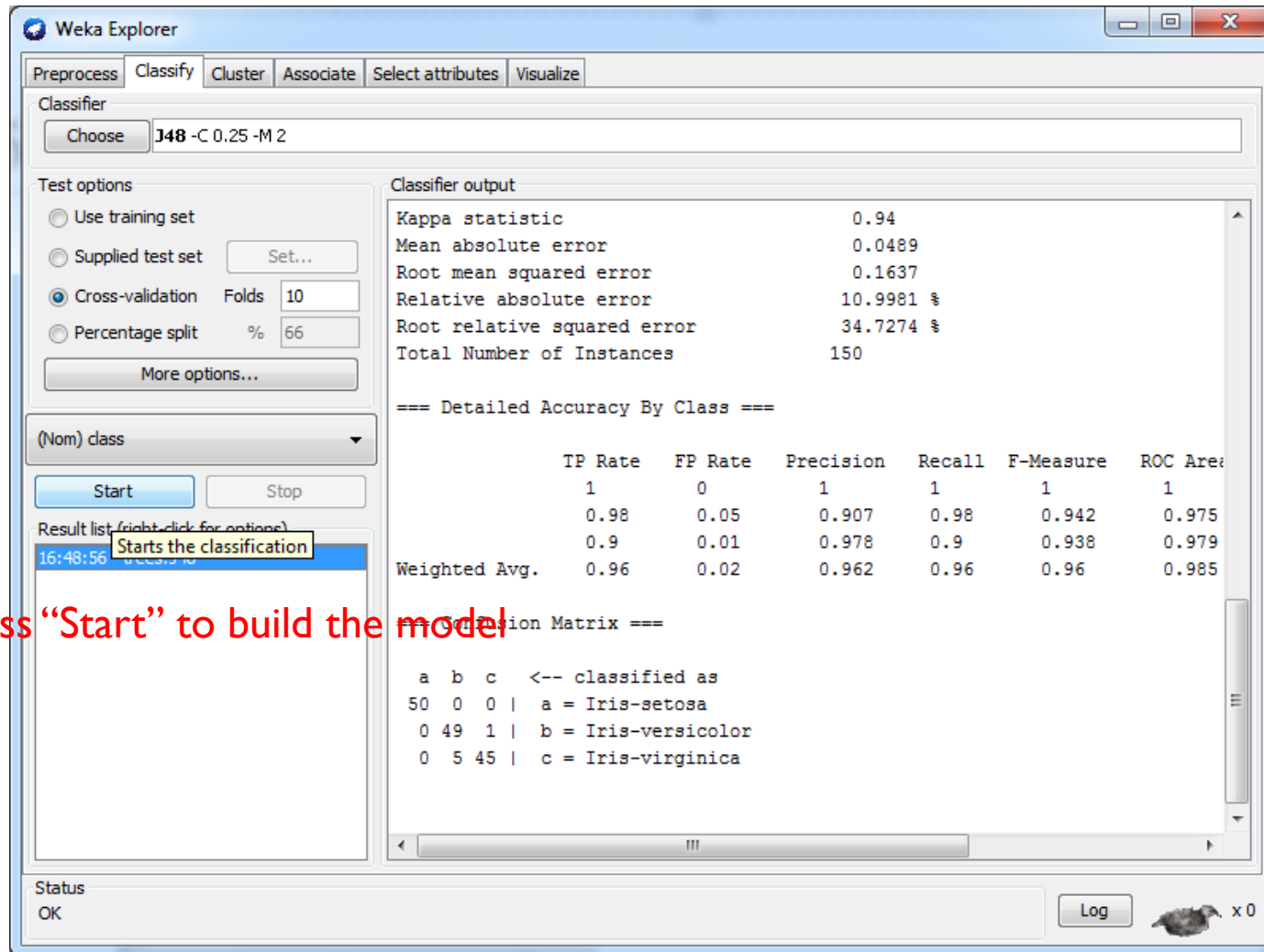
Class: class (Nom) Visualize All

Status  
OK Log x 0

## 5. Build a decision tree (with preprocessing)



## 5. Build a decision tree (with preprocessing)



The screenshot shows the Weka Explorer window with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'J48 -C 0.25 -M 2'. Under 'Test options', 'Cross-validation' is selected with 'Folds' set to 10. The 'Start' button is highlighted with a tooltip that says 'Starts the classification'. The 'Classifier output' pane displays the following metrics:

Metric	Value
Kappa statistic	0.94
Mean absolute error	0.0489
Root mean squared error	0.1637
Relative absolute error	10.9981 %
Root relative squared error	34.7274 %
Total Number of Instances	150

Below the metrics, a 'Detailed Accuracy By Class' table is shown:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
1	1	0	1	1	1	1
0.98	0.05	0.907	0.98	0.942	0.975	
0.9	0.01	0.978	0.9	0.938	0.979	
Weighted Avg.	0.96	0.02	0.962	0.96	0.96	0.985

The 'Confusion Matrix' section shows the following data:

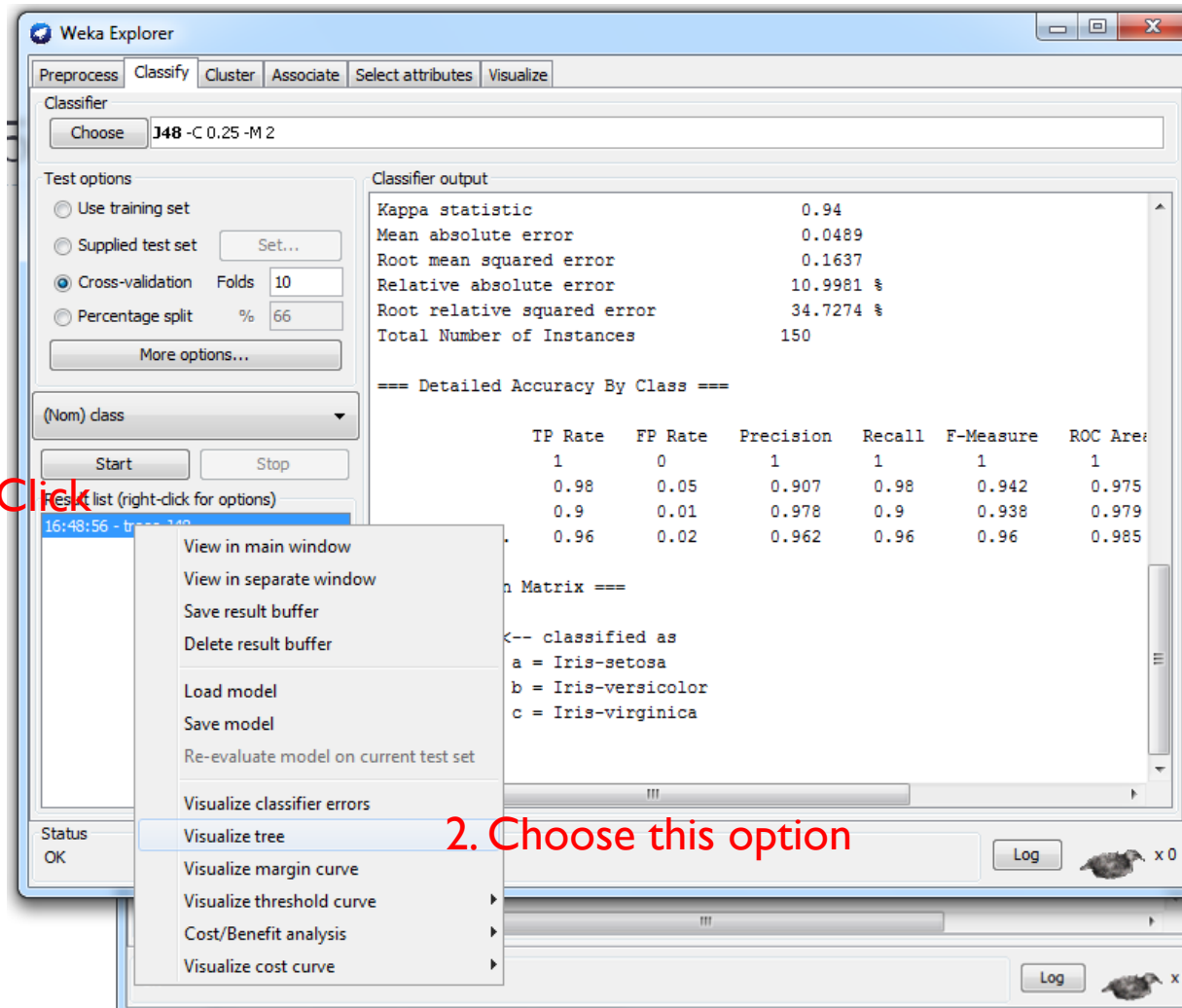
a	b	c	<-- classified as
50	0	0	a = Iris-setosa
0	49	1	b = Iris-versicolor
0	5	45	c = Iris-virginica

The status bar at the bottom indicates 'OK' and 'Log'.

Press "Start" to build the model

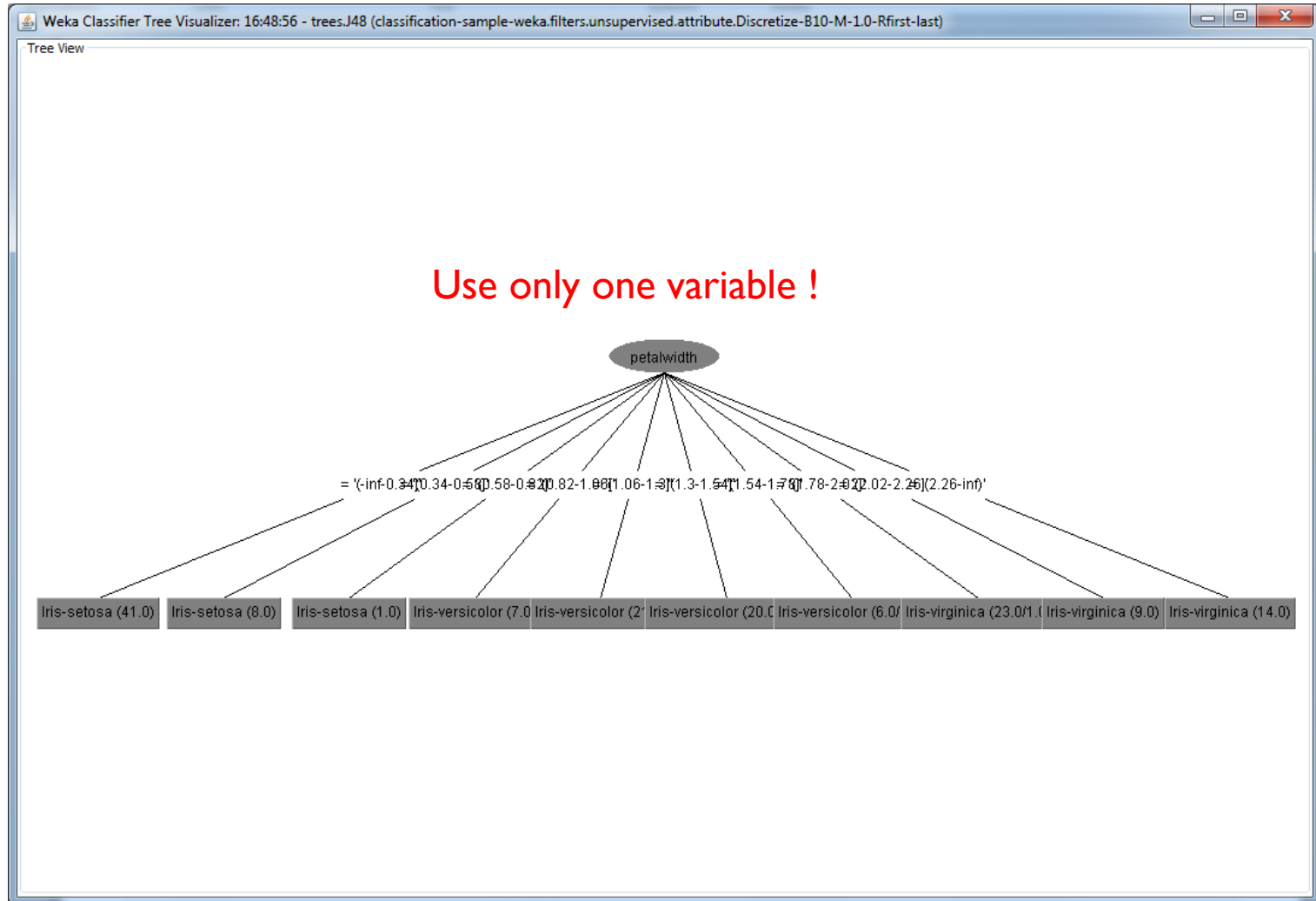
## 5. Build a decision tree (with preprocessing)

1. Right Click



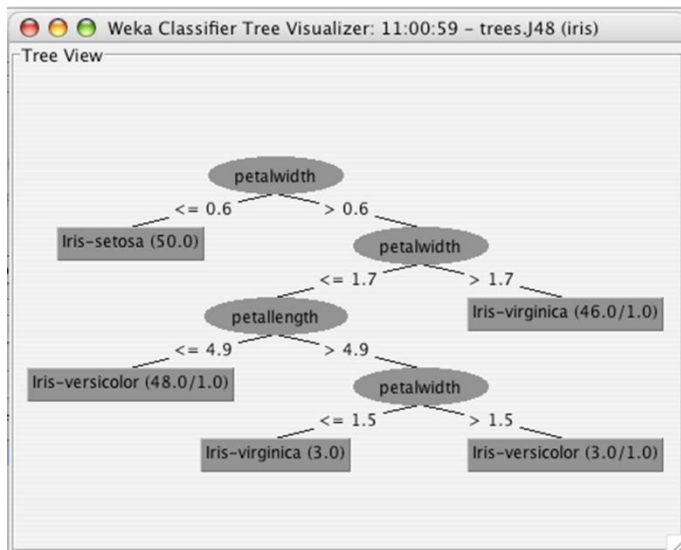
2. Choose this option

## 5. Build a decision tree (with preprocessing)

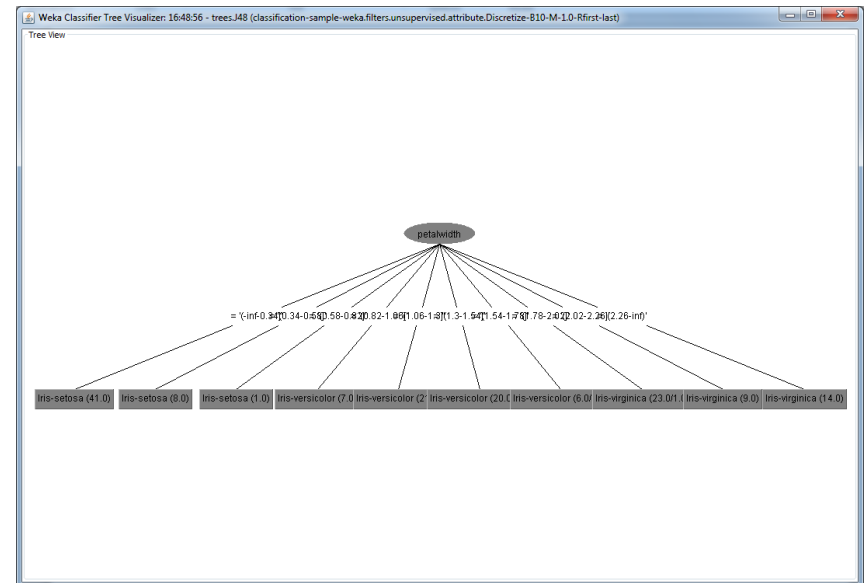




## Without preprocessing



## With Preprocessing



# Model Evaluation

---

## ► Confusion Matrix

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

Actual Class	Predicted 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

---

Actual Class	Predicted 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", Addison Wesley, 2006.

- ▶ **Datasets**

- ▶ [UC Irvine Machine Learning Repository](#)