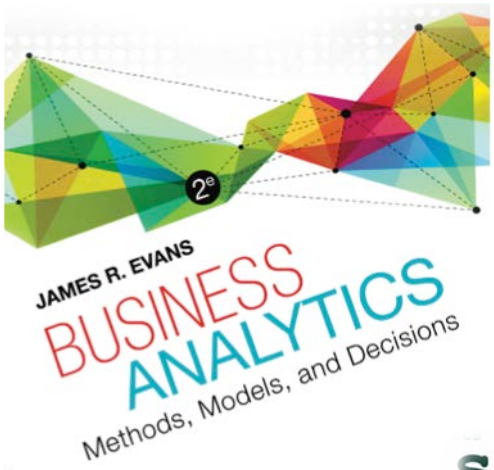


Chapter 10
Introduction to Data Mining

with 

(C) Pearson Education
Adapted by Michael Hahsler

Data Mining

- **Data mining** is focused on better understanding of characteristics and patterns among variables in large databases using a variety of statistical and analytical tools.

- It is used to identify **relationships** among variables in **large data sets** and understand **hidden patterns** that they may contain

The Scope of Data Mining

- *Clustering*
  - Identify groups with elements that are in some way similar to each other.
- *Classification*
  - Analyze data to predict how to classify a new data element.
- *Association Analysis*
  - Analyze databases to identify natural associations among variables and create rules for target marketing or buying recommendations.

Dirty Data

- Real data sets that have missing values or errors. Such data sets are called “dirty” and need to be “cleaned” prior to analyzing them.

- Approaches for handling missing data.
  - Eliminate the records that contain missing data
  - Estimate reasonable values for missing observations, such as the mean or median value

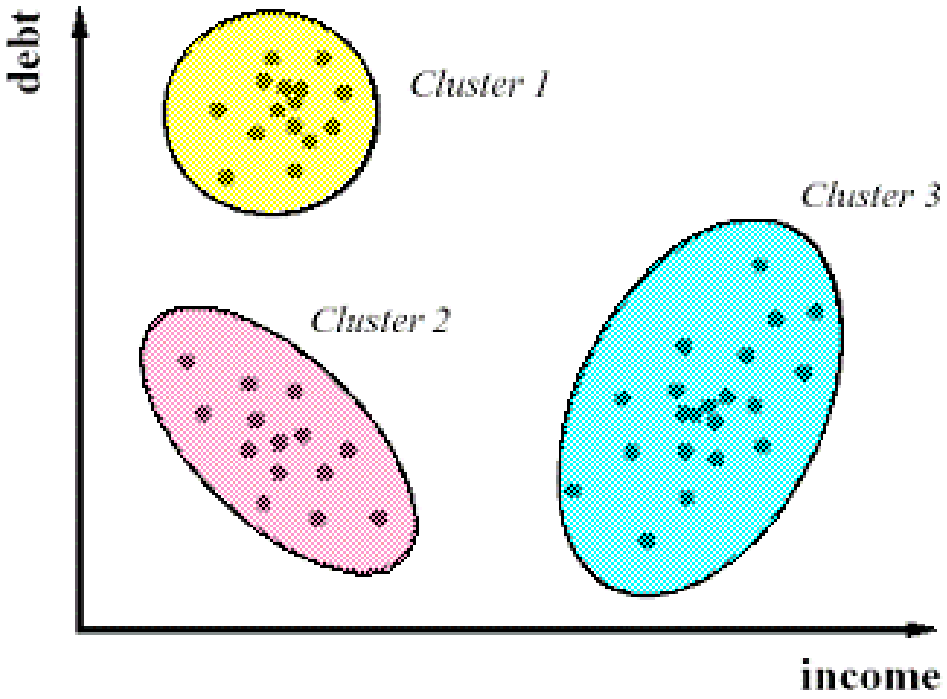
- Try to understand whether missing data are simply random events or if there is a logical reason. Eliminating sample data indiscriminately could result in misleading information and conclusions about the data.

Rapidminer:
- **Blending (e.g., sampling)**
- **Cleansing**

Cluster Analysis

Descriptive Analytics

- **Cluster analysis** (data segmentation) tries to group or segment a collection of objects into clusters, such that those within each cluster are more closely related to one another than to objects assigned to different clusters. The true grouping is typically not known (=*unsupervised learning*).



Distance Measures

- How do we measure similarity?
- Example: **Euclidean distance** is the straight-line distance between two points.
- The Euclidean distance measure between two points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ is

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad (10.1)$$

There exist many other distance measures! E.g., for categorical and mixed data.

Data should be normalized (scaled) before calculating distances.

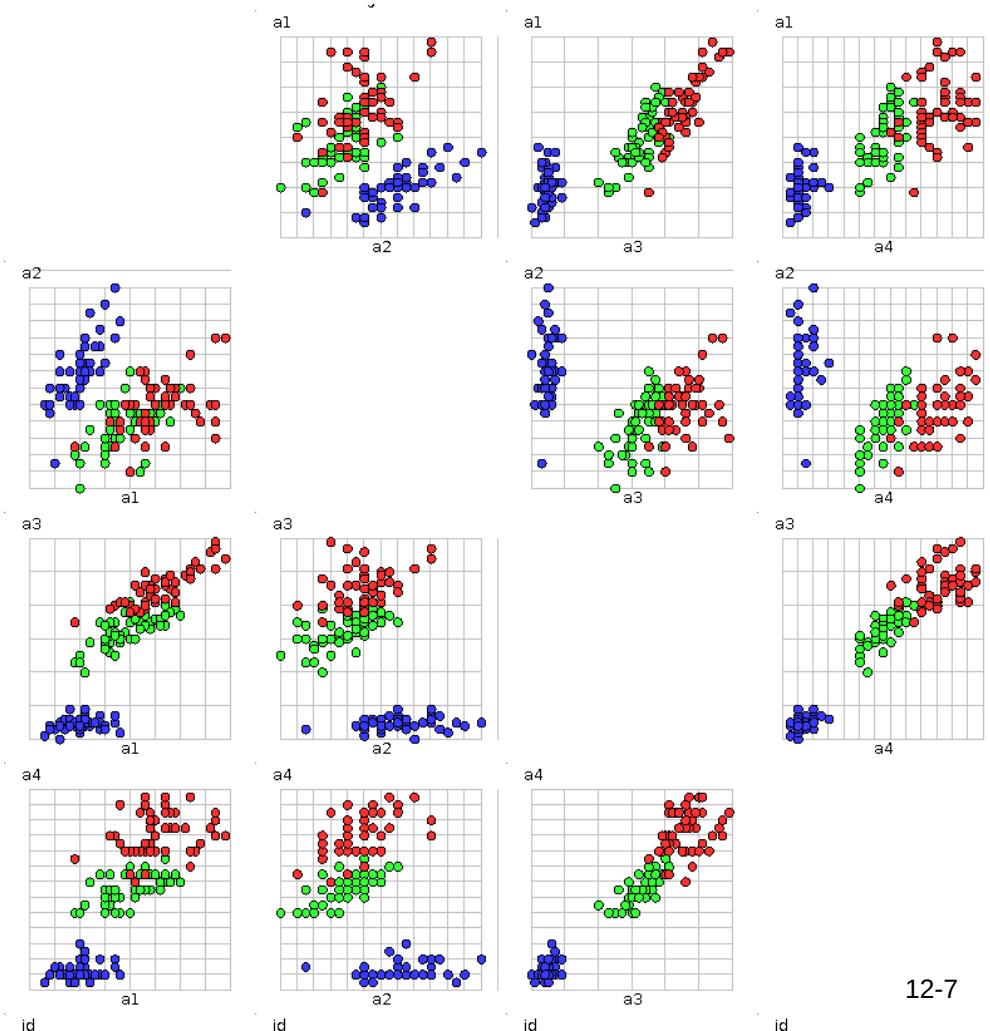
Rapidminer:
Cleansing - Normalization

Iris Data Set

- 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).
- 4 features were measured from each sample: the length and the width of the sepals and petals (in cm)

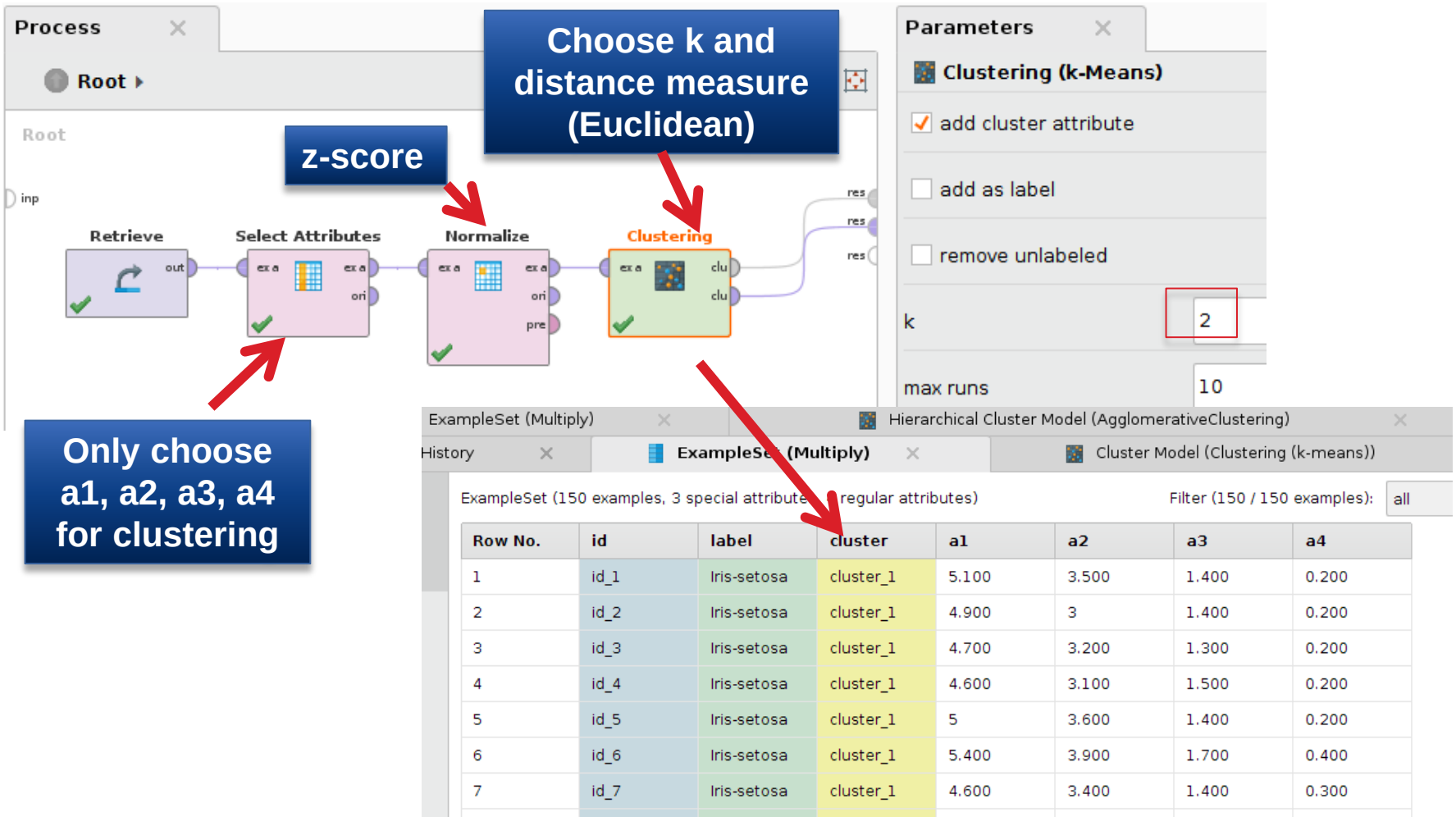
Examples:

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	I. setosa
4.9	3	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
4.6	3.1	1.5	0.2	I. setosa



Partitional Clustering

k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



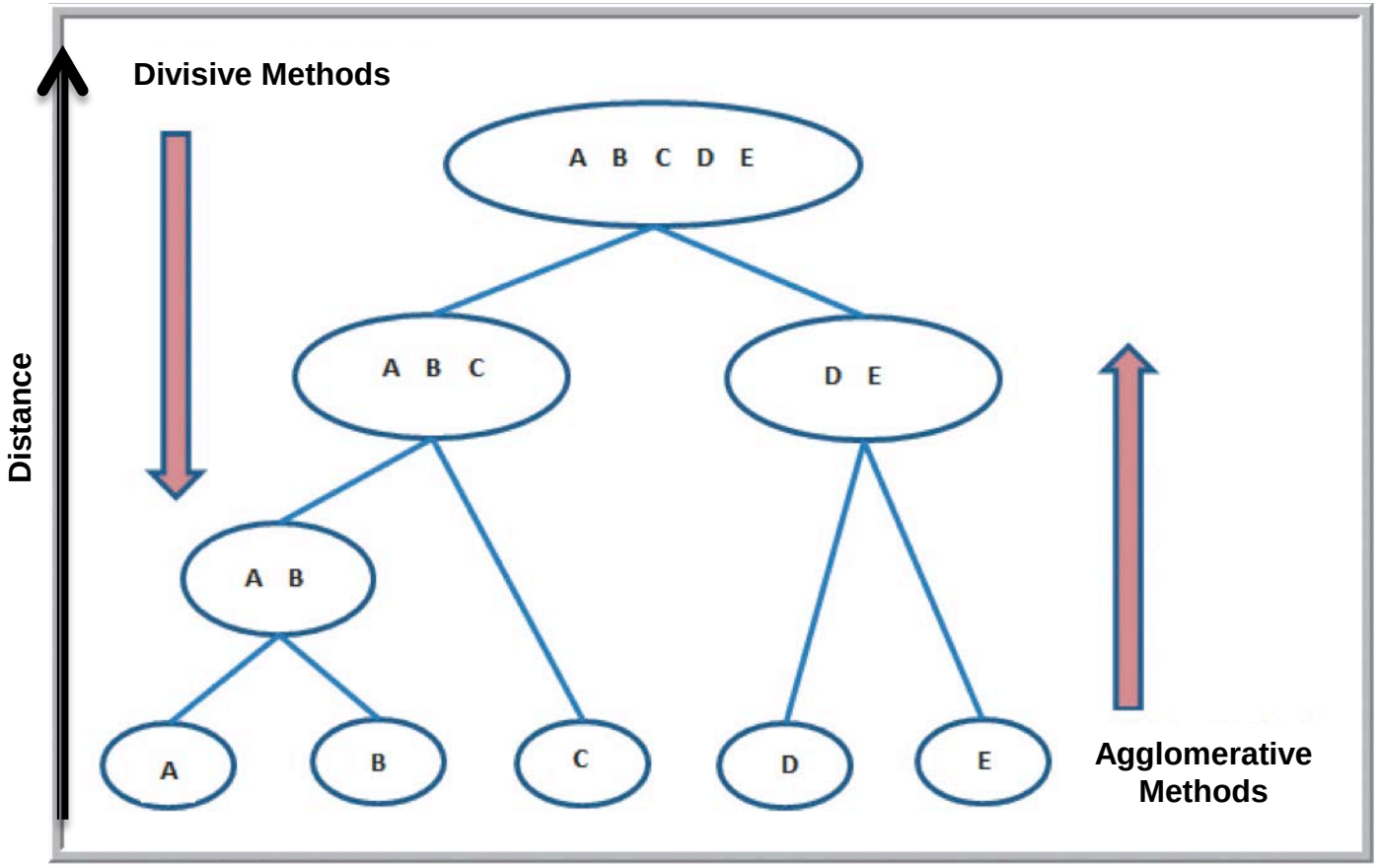
Row No.	id	label	cluster	a1	a2	a3	a4
1	id_1	Iris-setosa	cluster_1	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	cluster_1	4.900	3	1.400	0.200
3	id_3	Iris-setosa	cluster_1	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	cluster_1	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	cluster_1	5	3.600	1.400	0.200
6	id_6	Iris-setosa	cluster_1	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	cluster_1	4.600	3.400	1.400	0.300

Hierarchical Clustering

- **Hierarchical clustering** is a method of cluster analysis which seeks to build a hierarchy of clusters.
- Strategies for hierarchical clustering generally fall into two types:
  - **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
  - **Divisive:** This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

- Hierarchical clustering can be represented by a **dendrogram.**

Agglomerative vs. Divisive Clustering



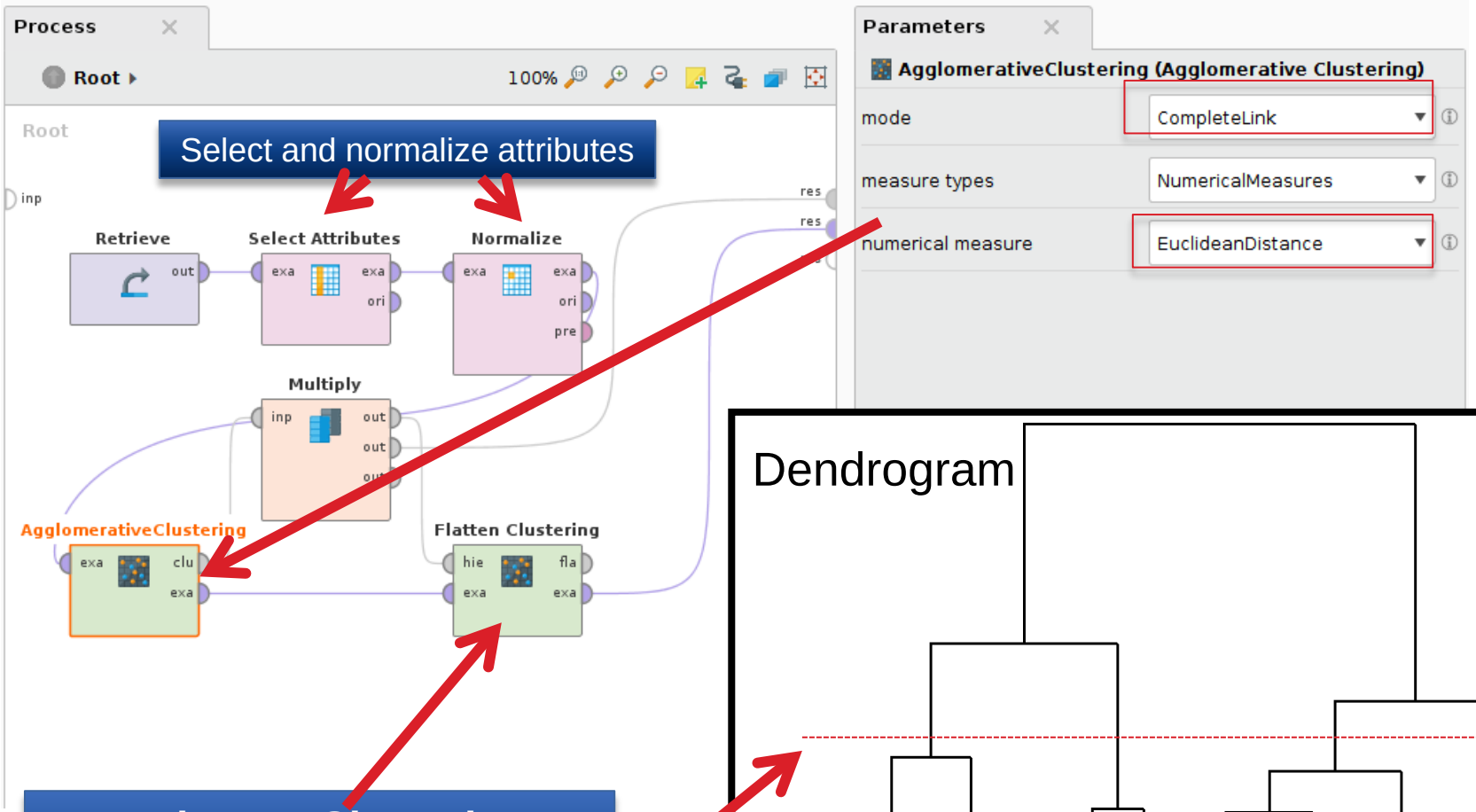
Dendrogram

Agglomerative Clustering Methods

How do we measure distance between groups?

- Single linkage clustering
  - The distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered.
- Complete linkage clustering
  - The distance between groups is the distance between the most distant pair of objects, one from each group.
- Average linkage clustering
  - Uses the mean of all pairwise distances between the objects of two clusters.
- Ward's hierarchical clustering
  - Uses a sum of squares criterion.

Example: Hierarchical Clustering



How to Use Clustering Results

- Analyze each cluster separately (e.g., group-wise means, bar charts)
- Give each cluster a label depending on the objects in the cluster (e.g., large flowers for the iris data set)
- Use the cluster group as an input for other models (e.g., regression or classification)

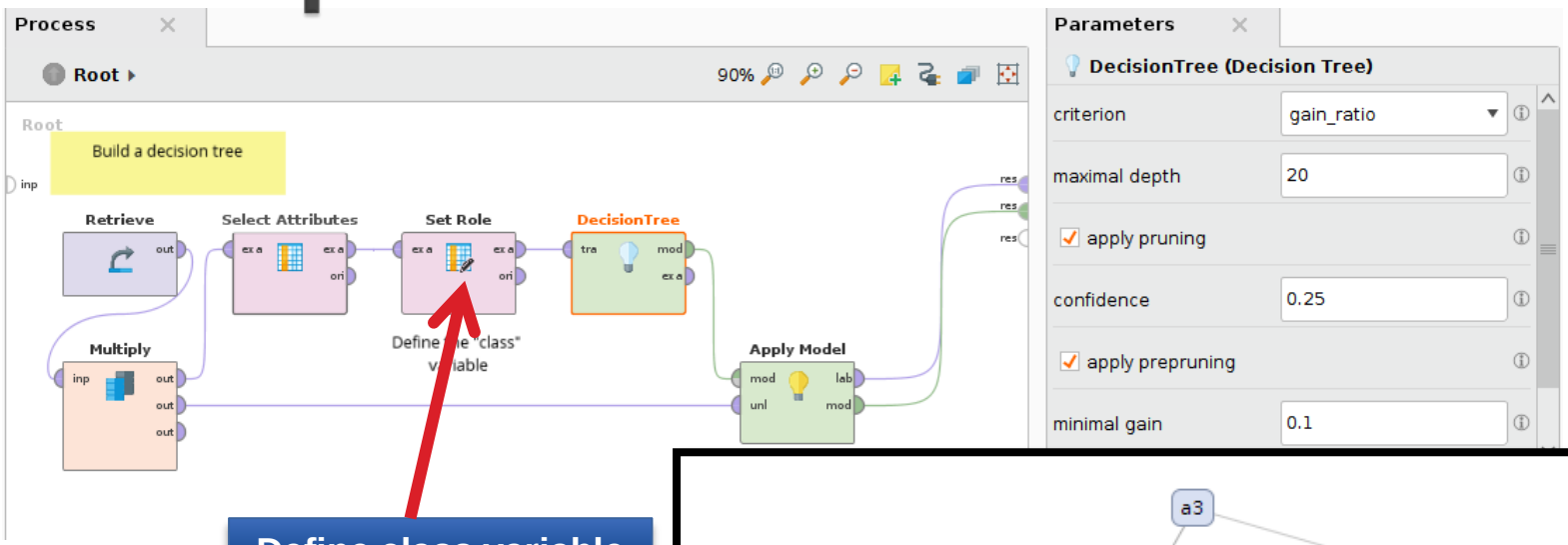
Classification

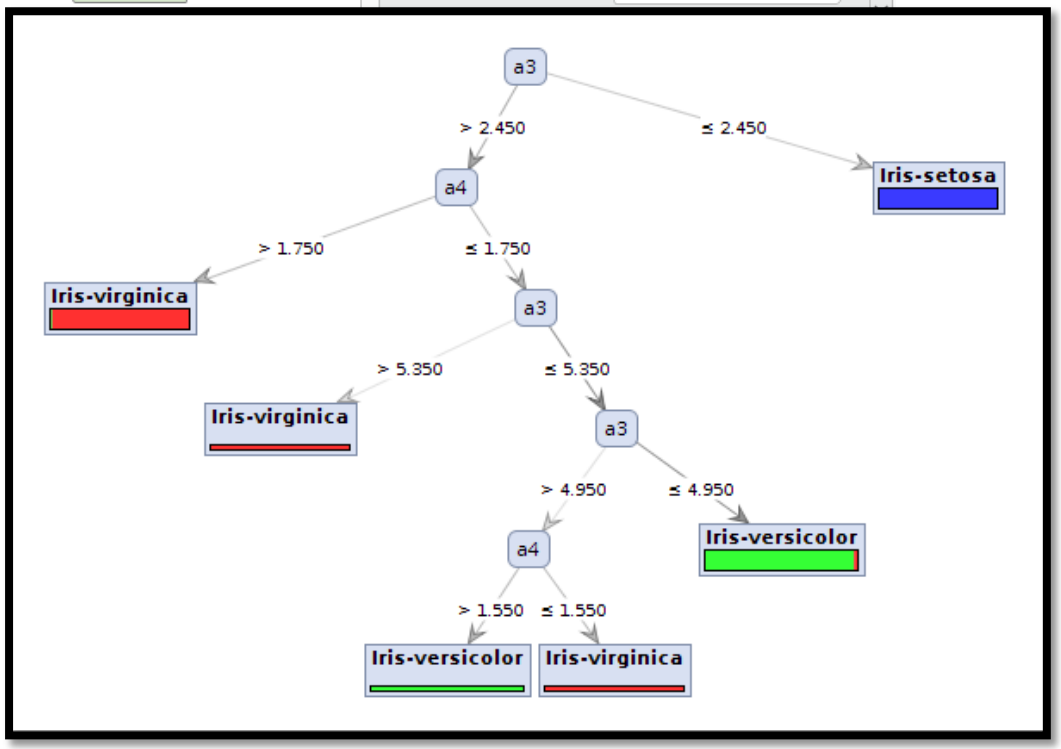
Predictive Analytics

- **Classification** is the problem of predicting to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known (=*supervised learning*).

- Similar to regression, but the outcome is categorical (often yes/no).

Example: Decision Tree

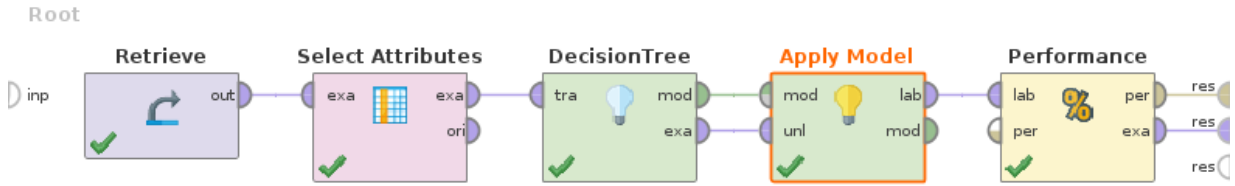




Measuring Classification Performance

- Find the probability of making a misclassification error.
- Represent the results in a **confusion matrix**, which shows the number of cases that were classified either correctly or incorrectly.
- Summarize the error rate into a single value. For example **accuracy** or **kappa**. Both measure the chance of making a correct prediction.

Example Evaluation (In-Sample Testing)



Predicted labels (and confidence of prediction)

ExampleSet (150 examples, 6 special attributes, 4 regular attributes) Filter (150 / 150

Row No.	id	label	predictio...	confiden...	confiden...	confiden...
1	id_1	Iris-setosa	Iris-setosa	1	0	0
2	id_2	Iris-setosa	Iris-setosa	1	0	0
3	id_3	Iris-setosa	Iris-setosa	1	0	0
4	id_4	Iris-setosa	Iris-setosa	1	0	0
5	id_5	Iris-setosa	Iris-setosa	1	0	0
6	id_6	Iris-setosa	Iris-setosa	1	0	0
7	id_7	Iris-setosa	Iris-setosa	1	0	0
8	id_8	Iris-setosa	Iris-setosa	1	0	0
9	id_9	Iris-set				

Confusion Matrix with Accuracy

accuracy: 98.67%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	50	0	0	100.00%
pred. Iris-versicolor	0	49	1	98.00%
pred. Iris-virginica	0	1	49	98.00%
class recall	100.00%	98.00%	98.00%	

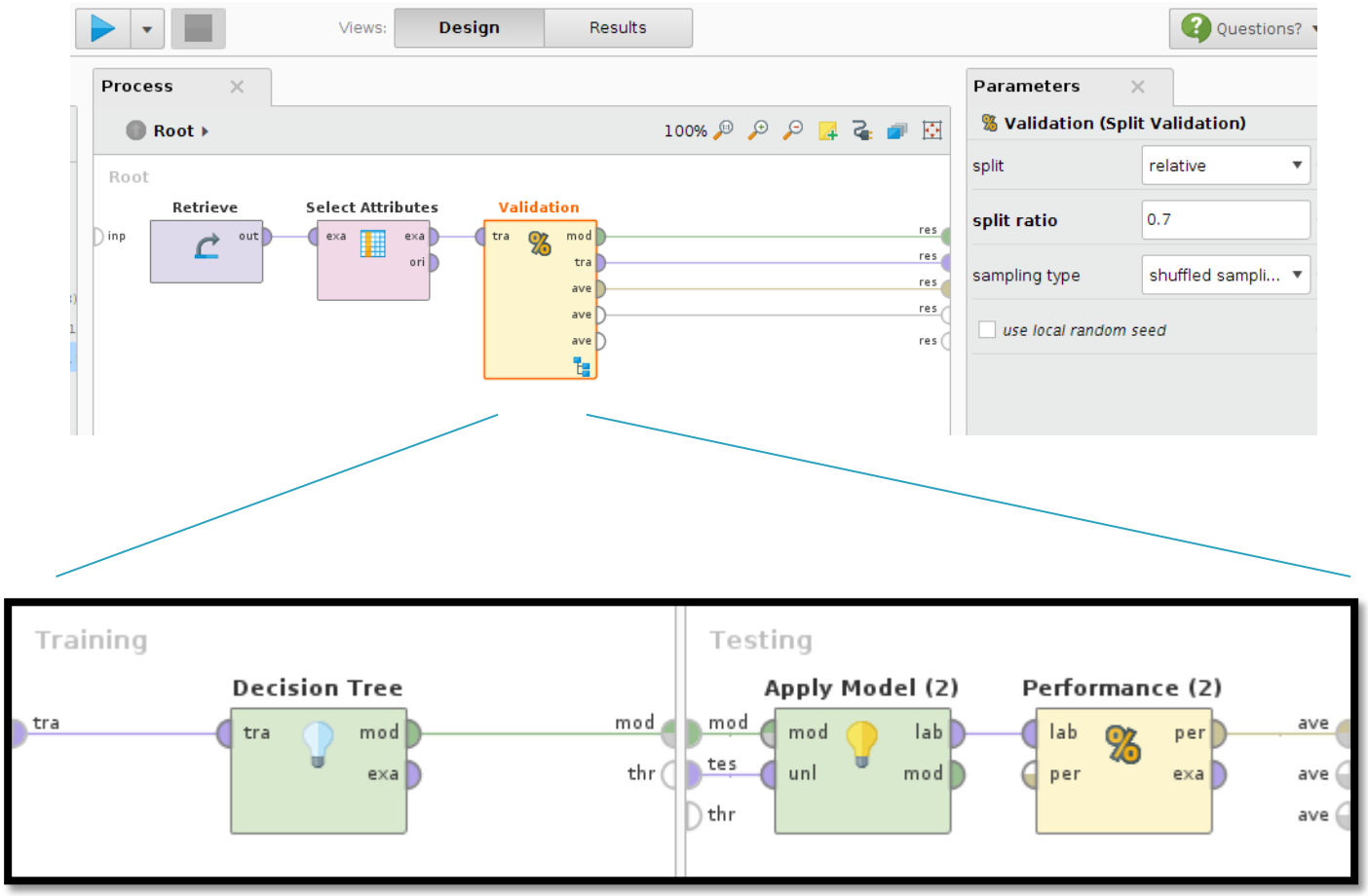
Using Training and Test Data

- Testing on the data used for training is not a good idea. We are more interested in how the model performs on new data!

- The data can be partitioned into:
  - training data set – has known outcomes and is used to "teach" the data-mining algorithm
  - test data set – tests the accuracy of the model

80% training / 20% testing is very common.

Example: Training and Test Data



You will get a confusion matrix for the test data.

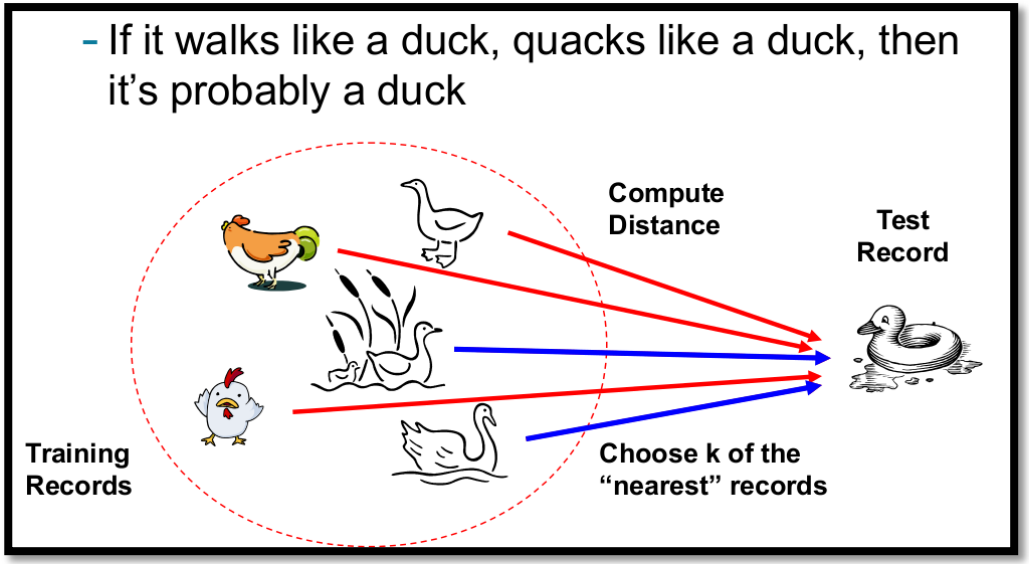
Other Classification Techniques

- *k-Nearest Neighbors (k-NN) Algorithm*
  - Finds records in a database that have similar numerical values of a set of predictor variables

- *Logistic Regression*
  - Estimates the probability of belonging to a category using a regression on the predictor variables.

k-Nearest Neighbors (*k*-NN)

- Measure the Euclidean distance between records in the training data set.
- The nearest neighbor to a record in the training data set is the one that that has the smallest distance from it.
  - If $k = 1$, then the 1-NN rule classifies a record in the same category as its nearest neighbor.
  - *k*-NN rule finds the *k*-Nearest Neighbors in the training data set to each record we want to classify and then assigns the classification as the classification of majority of the *k* nearest neighbors
- Typically, various values of *k* are used and then results inspected to determine which is best.



Logistic Regression

- **Logistic regression** is variation of linear regression in which the dependent variable is binary (0/1 or True/False).

- Predicts probabilities. Usually if the predicted probability for 1 is >50% then class 1 is predicted.

Classification Using Logistic Regression

- Estimate the probability *p* that an observation belongs to category 1, *P*(*Y* = 1), and, consequently, the probability 1 - *p* that it belongs to category 0, *P*(*Y* = 0).
- Then use a *cutoff value*, typically 0.5, with which to compare *p* and classify the observation into one of the two categories.
- The dependent variable is called the **logit**, which is the natural logarithm of *p*/(1 – *p*) – called the **odds** of belonging to category 1.
- The form of a logistic regression model is

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (10.3)$$

- The logit function can be solved for *p*:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}} \qquad (10.4)$$

Example

- Just replace the classification operator in Rapid Miner with whatever model you like.

Association Rule Mining

Descriptive Analytics

- **Association rule mining**, often called *affinity analysis*, seeks to uncover associations and/or correlation relationships in **large binary data sets**
  - Association rules identify attributes that occur together frequently in a given data set.
  - **Market basket analysis**, for example, is used determine groups of items consumers tend to purchase together.
- Association rules provide information in the form of if-then (antecedent-consequent) statements.

Example: Custom Computer Configuration

- *PC Purchase Data*
- We might want to know which components are often ordered together.

items

transactions

	A	B	C	D	E	F	G	H	I	J	K	L
1	PC Purchase Data											
2												
3	Processor			Screen Size			Memory			Hard Drive		
4												
5	Intel Core i3	Intel Core i5	Intel Core i7	10 inch screen	12 inch screen	15 inch screen	2 GB	4 GB	8 GB	320 GB	500 GB	750 GB
6	0	1	0	0	1	0	0	1	0	0	1	0
7	0	1	0	0	0	1	0	0	1	0	0	1
8	0	1	0	0	1	0	0	1	0	1	0	0
9	1	0	0	0	1	0	0	0	1	0	1	0
10	0	0	1	0	0	1	0	0	1	0	0	1
11	0	0	1	0	1	0	0	1	0	0	0	1
12	0	0	1	0	0	1	0	0	1	0	0	1
13	1	0	0	0	1	0	0	1	0	0	1	0
14	0	1	0	1	0	0	1	0	0	0	1	0

Measuring Strength of Association

- **Support for the (association) rule** is the percentage (or number) of transactions that include all items both antecedent and consequent.

$$\text{support} = P(antecedent\ and\ consequent) = \frac{\#\ transactions\ containing\ items}{\#\ transactions}$$

- **Confidence of the (association) rule** is the ratio of the number of transactions that include all items in the rule to the number of transactions that include all items in the antecedent.

$$\text{confidence} = P(\text{consequent} \mid \text{antecedent}) = \frac{P(\text{antecedent and consequent})}{P(\text{antecedent})} \qquad (10.5)$$

- **Lift** is a ratio of confidence to expected confidence.
  - Expected confidence is the number of transactions that include the consequent divided by the total number of transactions.
  - 1.0 means no relationship. Lift >> 1.0 means a strong association rule.

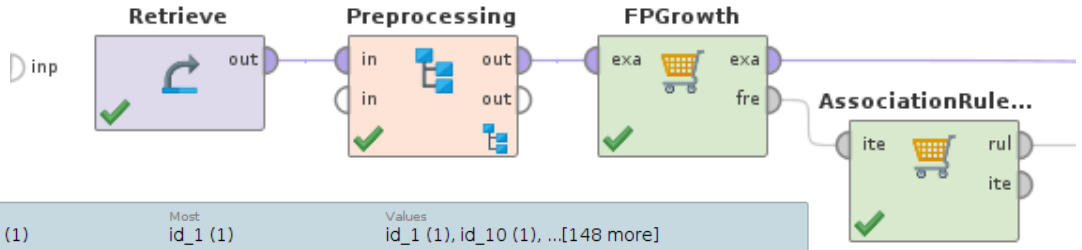
Example: Measuring Strength of Association

- A supermarket database has 100,000 point-of-sale transactions;
  - 2000 include both A and B items;
  - 5000 include C; and
  - 800 include A, B, and C

- Association rule: ***{A , B} => C***
  ("If A and B are purchased, then C is also purchased.")

  - Support = 800/100,000 = 0.008
  - Confidence = 800/2000 = 0.40
  - Expected confidence = 5000/100,000 = 0.05
  - Lift = 0.40/0.05 = 8

Example



Id **id**	Nominal	0	Least id_99 (1)	Most id_1 (1)	Values id_1 (1), id_10 (1), ...[148 more]
Label **label**	Nominal	0	Least Iris-virginica (50)	Most Iris-setosa (50)	Values Iris-setosa (50), Iris-versicolor (50), ...[1 more]
a1 = range1 [-∞ - 5.050]	Binominal	0	Least true (32)	Most false (118)	Values false (118), true (32)
a1 = range2 [5.050 - 5.650]	Binominal	0	Least true (33)	Most false (117)	Values false (117), true (33)
a1 = range3 [5.65					
a1 = range4 [6.15					
a1 = range5 [6.55					

Show rules matching

all of these conclusions:

label = Iris-virginica
label = Iris-versicolor
label = Iris-setosa
a4 = range3 [1.150 - 1.550]
a3 = range1 [-∞ - 1.550]
a4 = range1 [-∞ - 0.250]
a3 = range5 [5.350 - ∞]
a1 = range5 [6.550 - ∞]
a4 = range5 [1.950 - ∞]
a3 = range3 [3.950 - 4.650]

No.	Premises	Conclusion	Su
24	label = Iris-versicolor, a2 = range2 [2.750 -...	a4 = range3 [1.150 - 1.550], a3 = range3 ...	0.
47	label = Iris-virginica, a1 = range5 [6.550 - ∞]	a3 = range5 [5.350 - ∞]	0.
43	a1 = range5 [6.550 - ∞], a4 = range5 [1.9...	a3 = range5 [5.350 - ∞]	0.
44	a1 = range5 [6.550 - ∞], a4 = range5 [1.9...	label = Iris-virginica, a3 = range5 [5.350 - ∞]	0.
45	label = Iris-virginica, a1 = range5 [6.550 - ...	a3 = range5 [5.350 - ∞]	0.
67	a2 = range2 [2.750 - 3.050], a3 = range3 ...	label = Iris-versicolor, a4 = range3 [1.150 -...	0.
35	label = Iris-versicolor, a2 = range2 [2.750 -...	a3 = range3 [3.950 - 4.650]	0.

AssociationRules

Association Rules
[label = Iris-versicolor] --> [a4 = range3 [1.150 - 1.550]] (confidence: 0.700)
[a3 = range5 [5.350 - ∞]] --> [a4 = range5 [1.950 - ∞]] (confidence: 0.700)
[a3 = range5 [5.350 - ∞]] --> [label = Iris-virginica, a4 = range5 [1.950 - ∞]] (confidence: 0.700)
[label = Iris-virginica, a3 = range5 [5.350 - ∞]] --> [a4 = range5 [1.950 - ∞]] (confidence: 0.700)
[label = Iris-versicolor, a4 = range3 [1.150 - 1.550]] --> [a3 = range3 [3.950 - 4.650]] (confidence: 0.714)
[a3 = range5 [5.350 - ∞], a4 = range5 [1.950 - ∞]] --> [a1 = range5 [6.550 - ∞]] (confidence: 0.714)
[a3 = range5 [5.350 - ∞], a4 = range5 [1.950 - ∞]] --> [label = Iris-virginica, a1 = range5 [6.550 - ∞]] (confic
[label = Iris-virginica, a3 = range5 [5.350 - ∞], a4 = range5 [1.950 - ∞]] --> [a1 = range5 [6.550 - ∞]] (confic
[a4 = range5 [1.950 - ∞]] --> [a3 = range5 [5.350 - ∞]] (confidence: 0.724)
[a4 = range5 [1.950 - ∞]] --> [label = Iris-virginica, a3 = range5 [5.350 - ∞]] (confidence: 0.724)
[label = Iris-virginica, a4 = range5 [1.950 - ∞]] --> [a3 = range5 [5.350 - ∞]] (confidence: 0.724)
[a3 = range1 [-∞ - 1.550]] --> [a4 = range1 [-∞ - 0.250]] (confidence: 0.730)
[a3 = range1 [-∞ - 1.550]] --> [label = Iris-setosa, a4 = range1 [-∞ - 0.250]] (confidence: 0.730)
[label = Iris-setosa, a3 = range1 [-∞ - 1.550]] --> [a4 = range1 [-∞ - 0.250]] (confidence: 0.730)

Conclusion

- Data mining offers many methods for descriptive and predictive analytics.
- Different methods work better for different data sets.
- It is often not clear which method to use and most analytics professionals will try and compare several methods.
- **The most critical part is cleaning and preparing the data and to ask the right questions.**