CURSOS DE VERANO 2014

APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KNIME, R, HADOOP Y MAHOUT

Introducción a KNIME

María José del Jesus

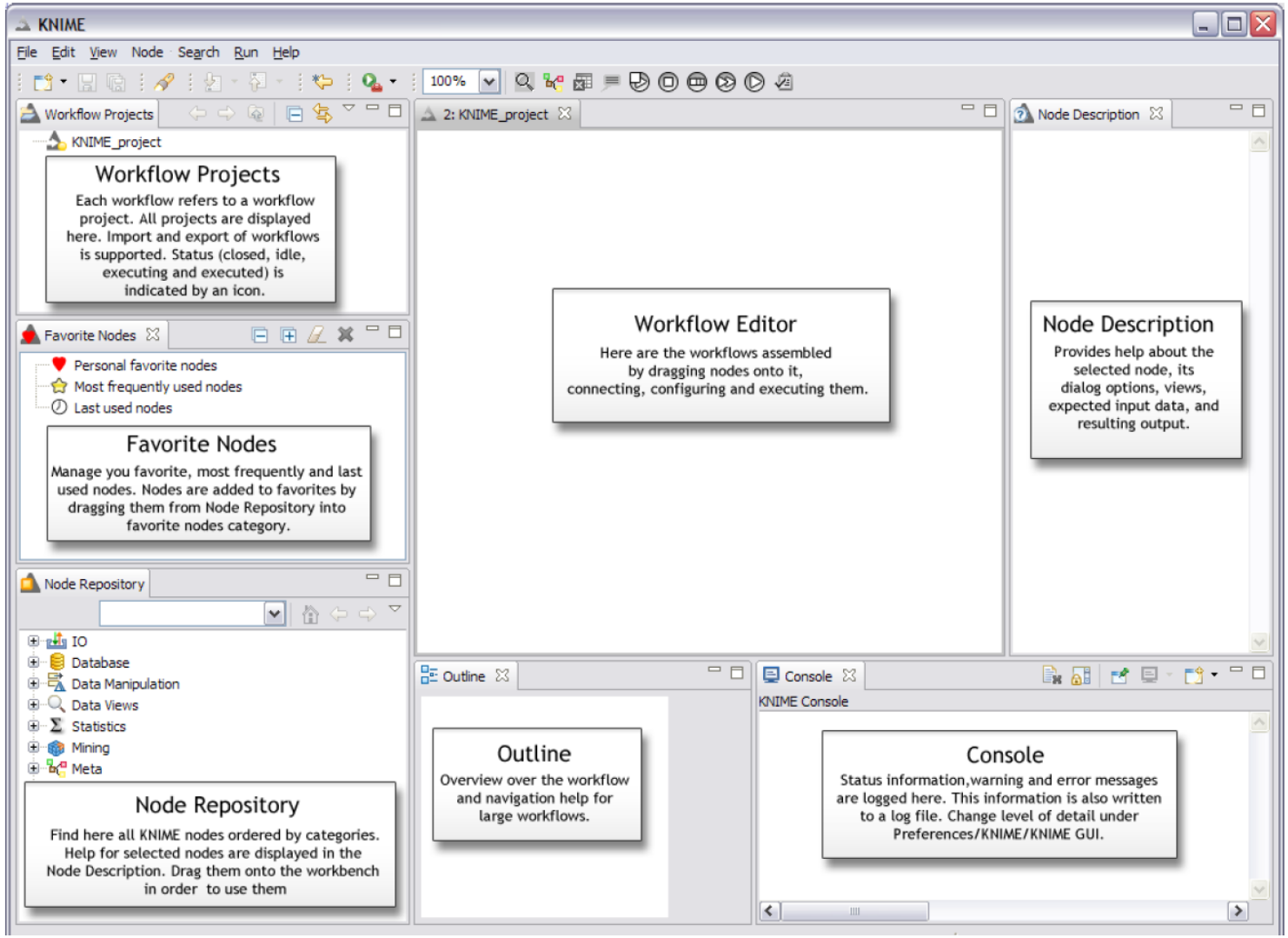
KNIME

- ¿Por qué aprender KNIME?
- El entorno de trabajo de KNIME
- Ejemplos

¿Por qué aprender KNIME?

- KNIME es una herramienta Open Source
- Ofrece el ciclo completo de KDD
  - Visualización de datos
  - Pre-procesado de datos
  - Extracción de modelos mediante algoritmos de Minería de Datos
  - Comparación de modelos
  - Análisis de resultados
- Integración con Weka y R
- Información sobre instalación y uso:
  - https://www.knime.org

El entorno de trabajo en KNIME



El entorno de trabajo en KNIME

Conceptos básicos

- Un proceso de análisis de datos se representa mediante un **flujo de trabajo**
- **Construcción de un flujo de trabajo**:
  - arrastrar nodos del Almacén de nodos y pegar en el editor de flujo de trabajo
  - Conectar nodos
- Estado de un **nodo**
  - Rojo → hay que configurar antes de ejecutar
  - Amarillo → el nodo está preparado para ejecutar
  - Verde → el nodo se ha ejecutado
- **Puertos** (de entrada o salida)
  - Solo se pueden conectar puertos del mismo tipo
    - Datos (triángulo amarillo): transfieren tablas de datos entre nodos
    - Bases de datos (cuadrado marrón)
    - PMML: transfieren modelos ya aprendidos
    - Otros puertos

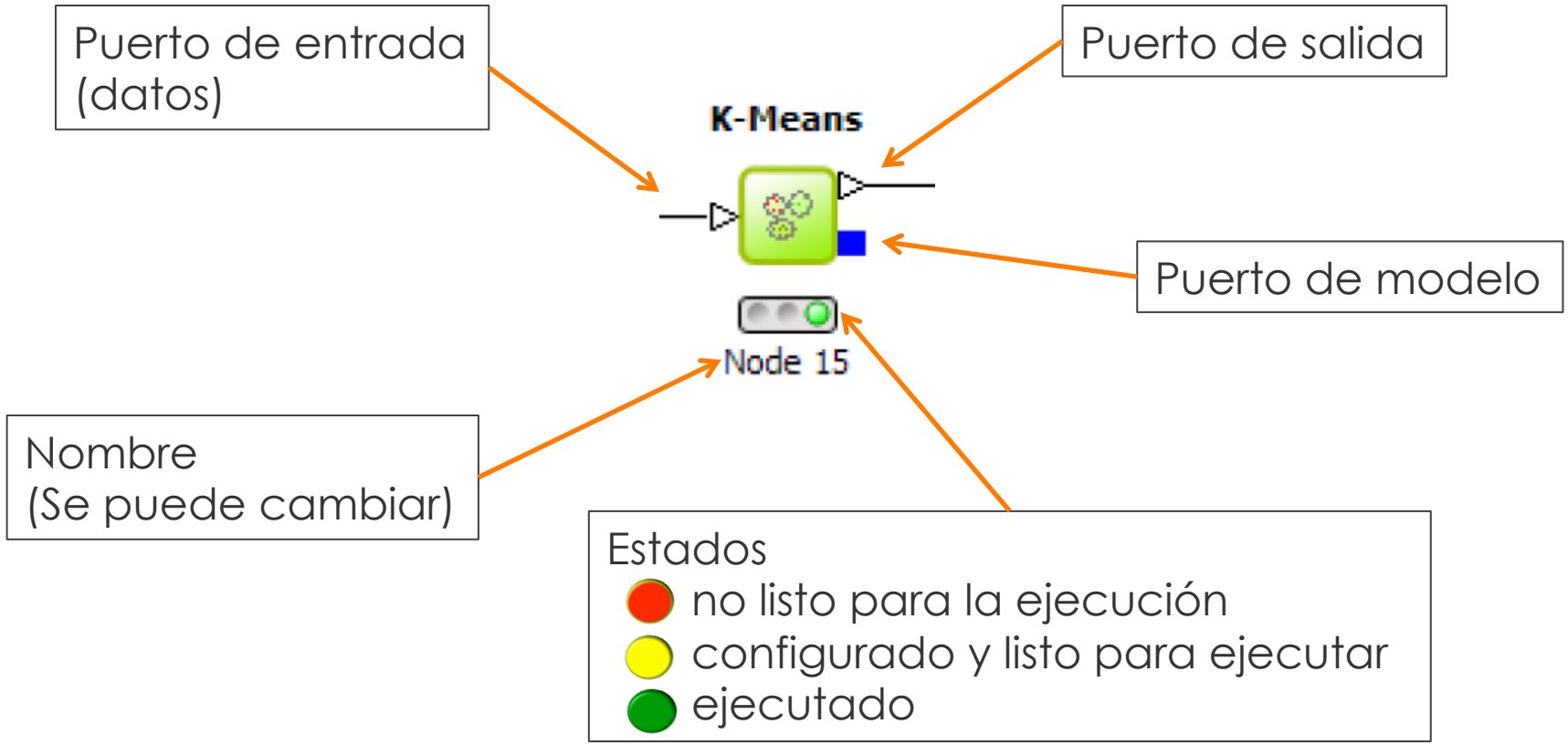
El entorno de trabajo en KNIME

Acciones básicas:

- Crear un proyecto
- Utilizar nodos
- Construir un flujo de datos
- Nodo color manager
- Configuración de nodos
- Ejecución de flujo de datos
- Resultados HiLiting

El entorno de trabajo en KNIME ► nodos

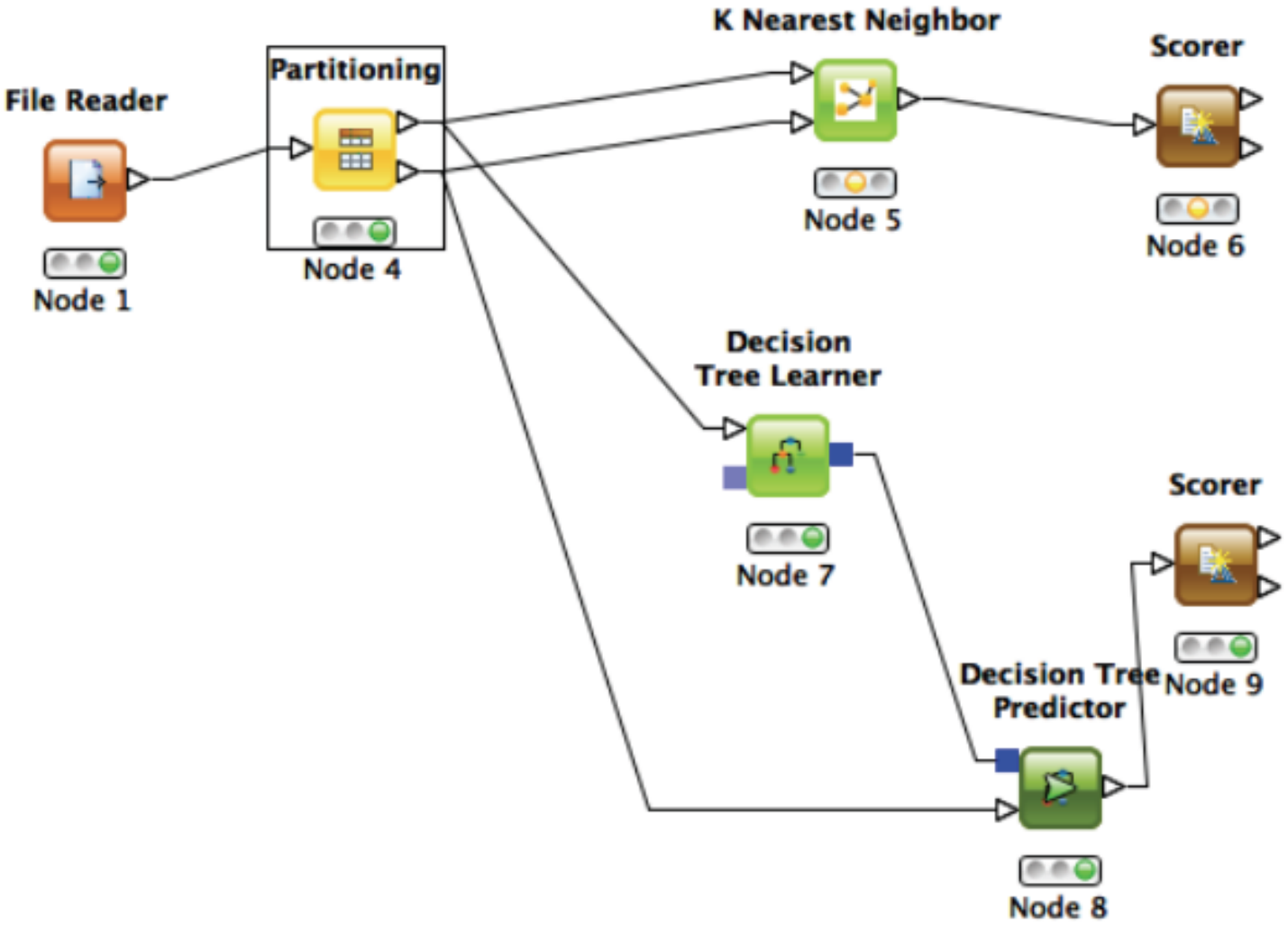
- Son unidades de procesamiento de un workflow



El entorno de trabajo en KNIME ► Construir un flujo de datos

- Se construye un flujo arrastrando y soltando los nodos desde el almacén de nodos al editor de proyectos y conectándolos entre ellos
- Los datos se transportan entre nodos a través de los puertos
- Es necesario, una vez colocados los nodos en el editor, conectar la salida de cada nodo con el predecesor

El entorno de trabajo en KNIME ► Construir un flujo de datos



El entorno de trabajo en KNIME ► nodo Color Manager

- Permite colorear los resultados generados a partir de los datos de entrada
- El coloreo afecta a muchas vistas y ayuda a diferenciar los datos
- Si se inserta este nodo en el flujo de trabajo, los datos se codifican según los colores determinados por el Color Manager

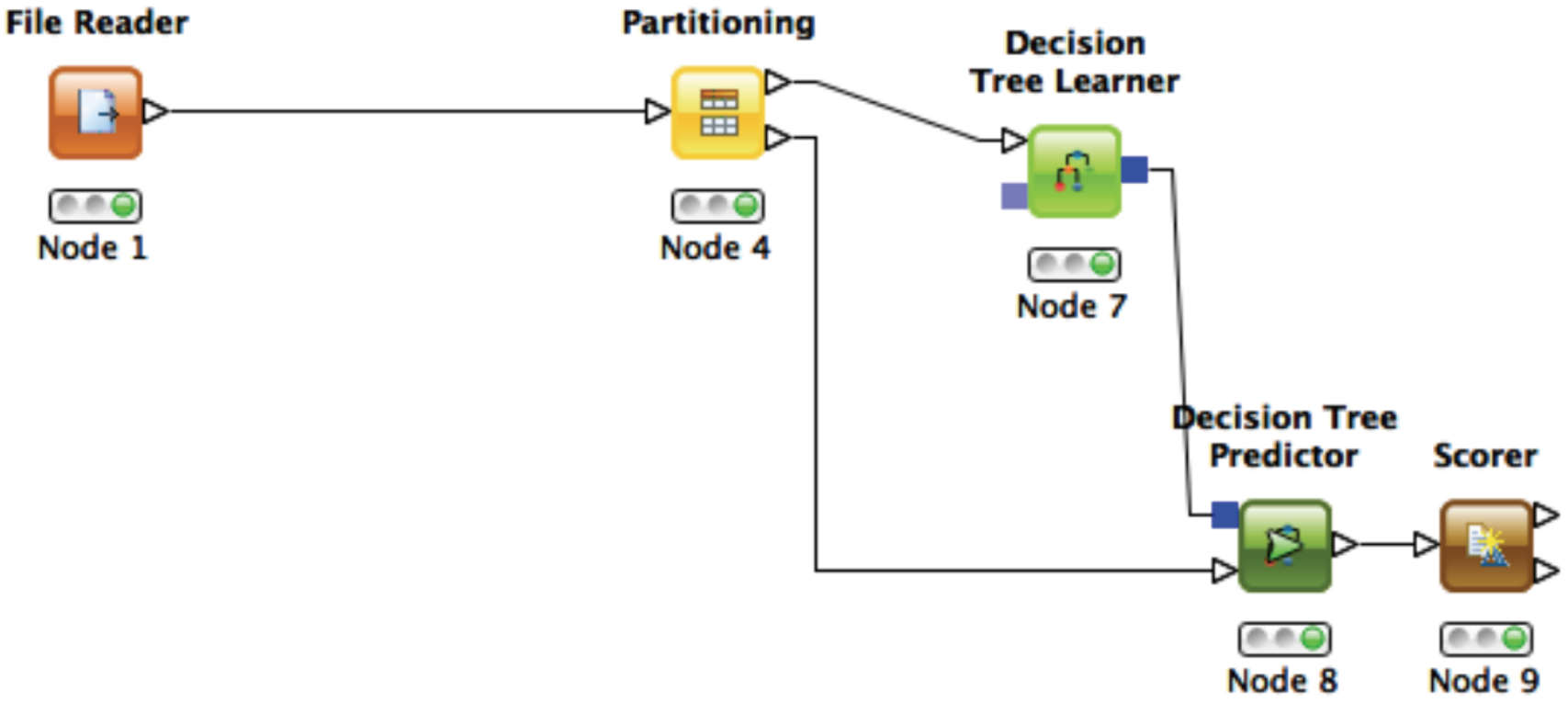
El entorno de trabajo en KNIME ► ejecución del flujo de datos

- Cuando los nodos del flujo tienen color amarillo, se puede ejecutar
- Los nodos se ejecutan de izquierda a derecha
  - Un nodo puede ejecutarse cuando todos los nodos predecesores han terminado su ejecución
- Formas de ejecución:
  - Por nodo (con la opción Execute)
  - Ejecutando el último nodo del flujo (KNIME ejecuta los predecesores)
  - Seleccionando varios nodos y disparando la ejecución (KNIME determina el orden y ejecuta nodos en paralelo, si es posible)

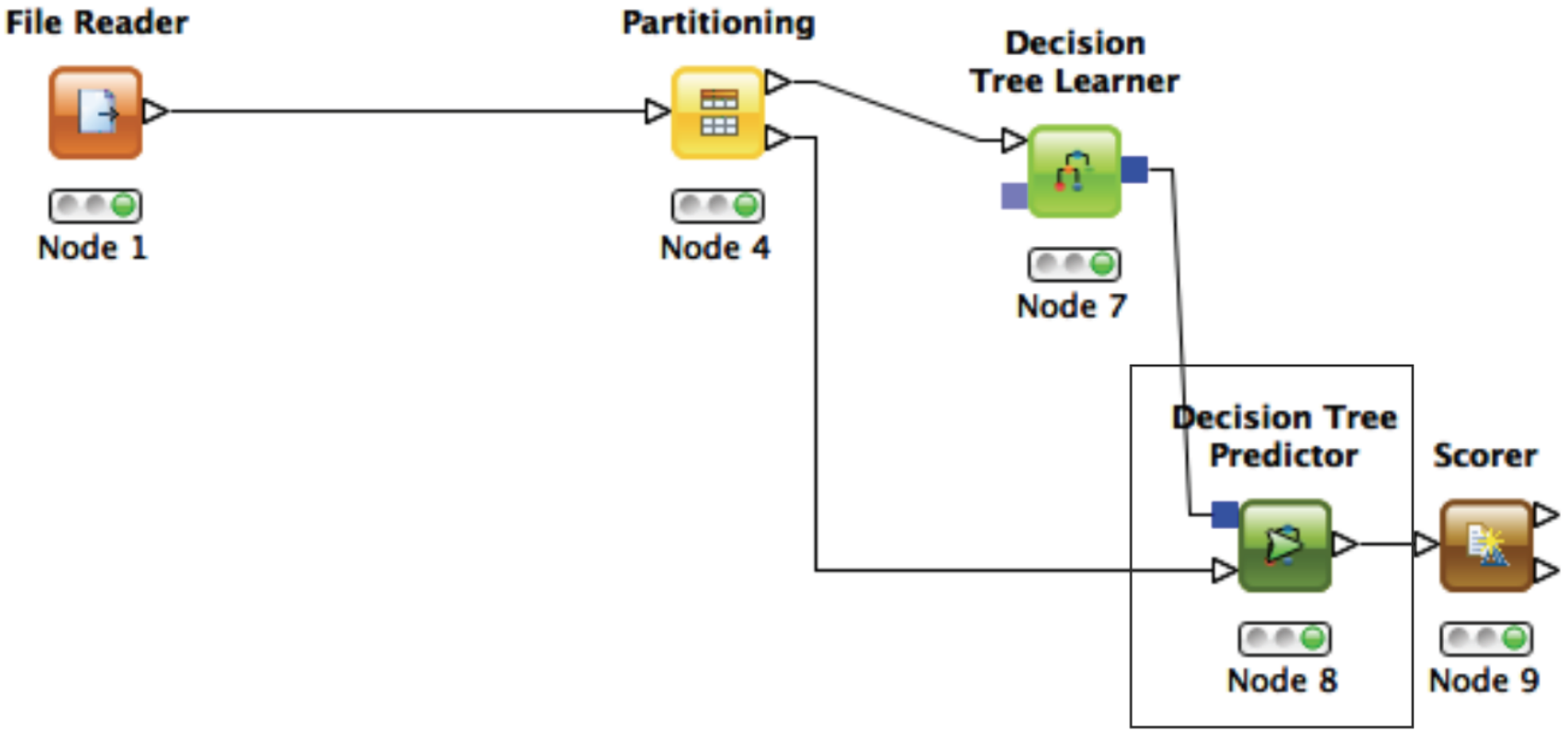
El entorno de trabajo en KNIME ► HiLiting

- Si se seleccionan datos en una vista y se aplica “Hilite” sobre ellos, se podrá ver el efecto “Hilite” sobre los datos en el resto de vistas del flujo que soporten esta opción
- Los datos seleccionados se resaltarán en color naranja

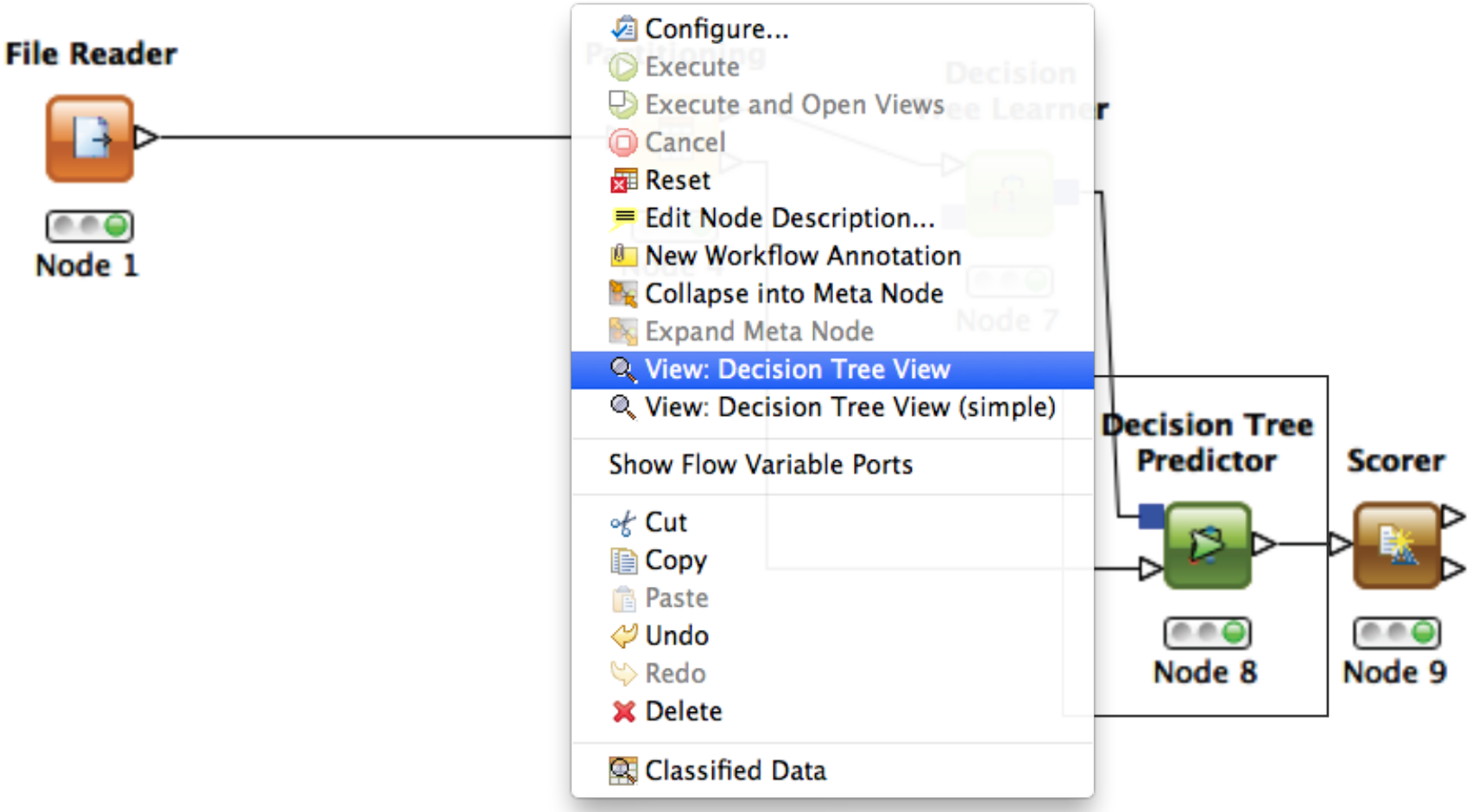
El entorno de trabajo en KNIME ► HiLiting



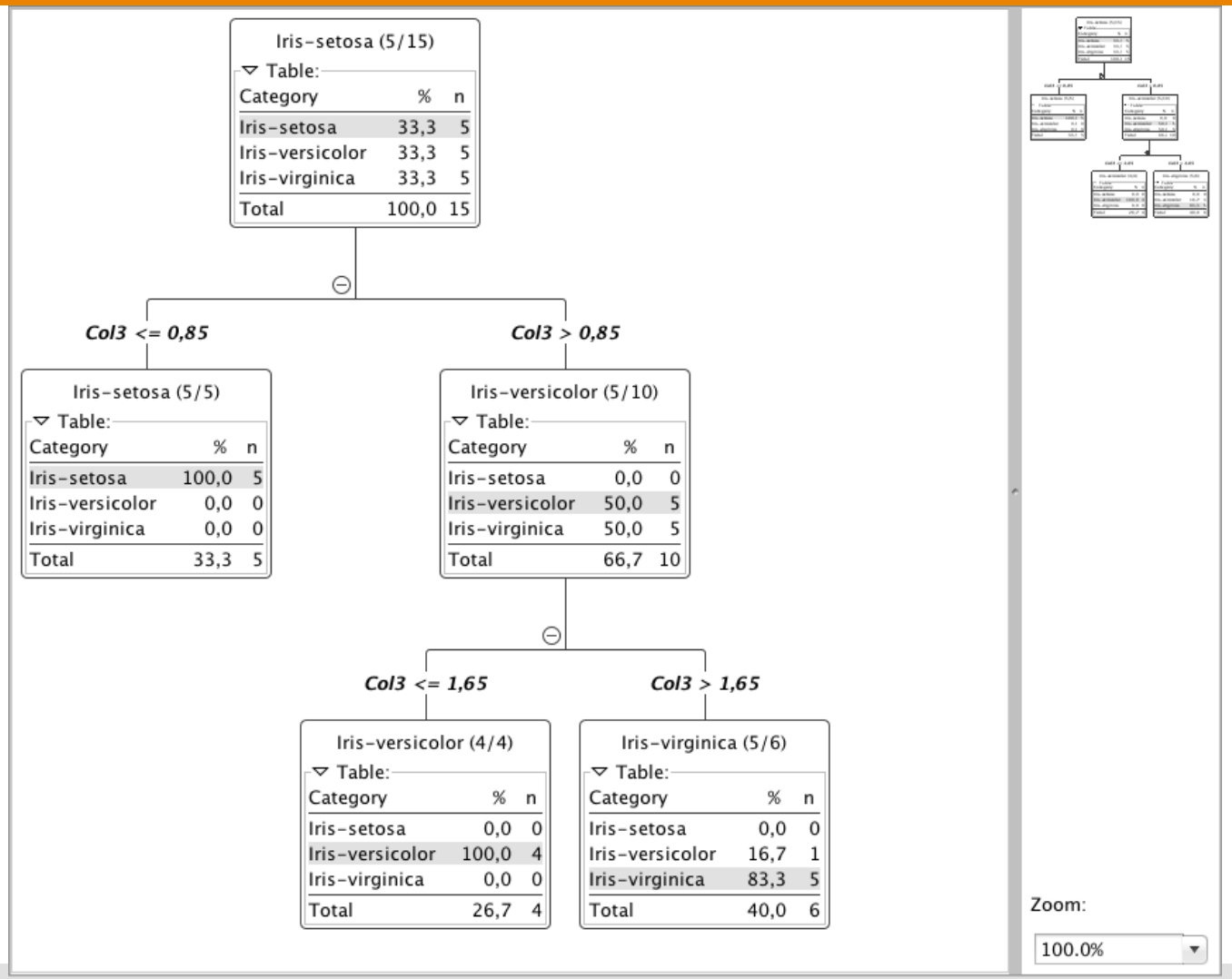
El entorno de trabajo en KNIME ► HiLiting



El entorno de trabajo en KNIME ► HiLiting

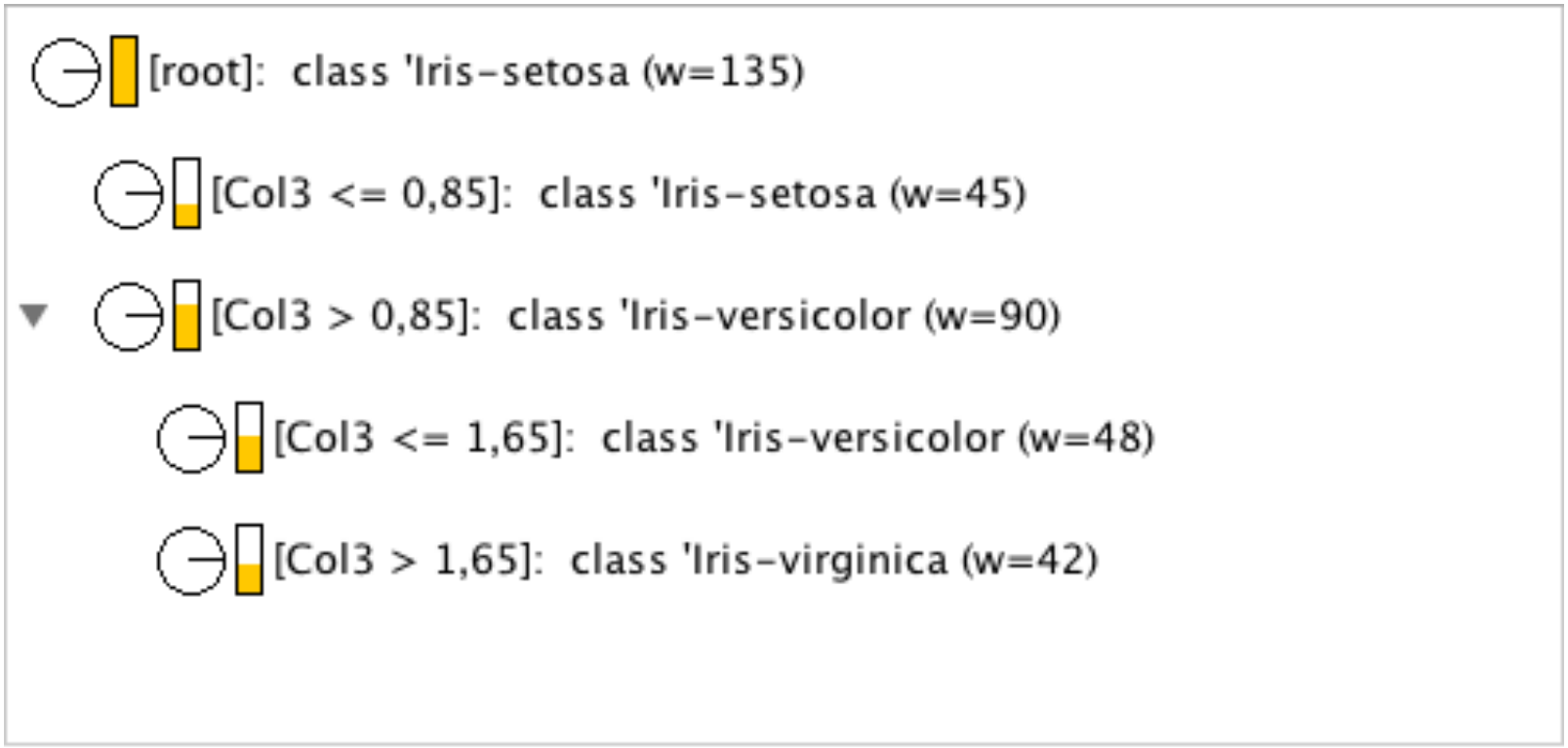


El entorno de trabajo en KNIME ► HiLiting

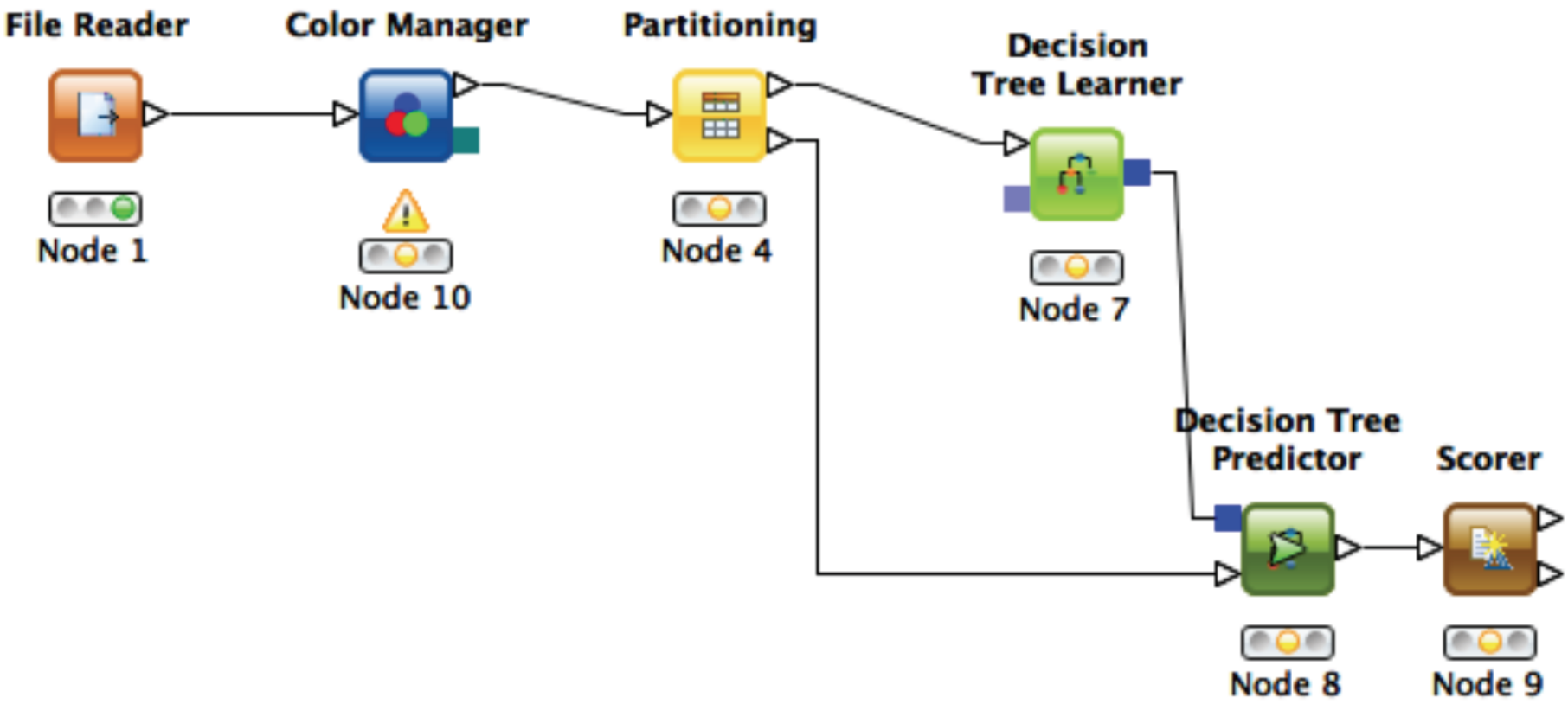


El entorno de trabajo en KNIME ► HiLiting

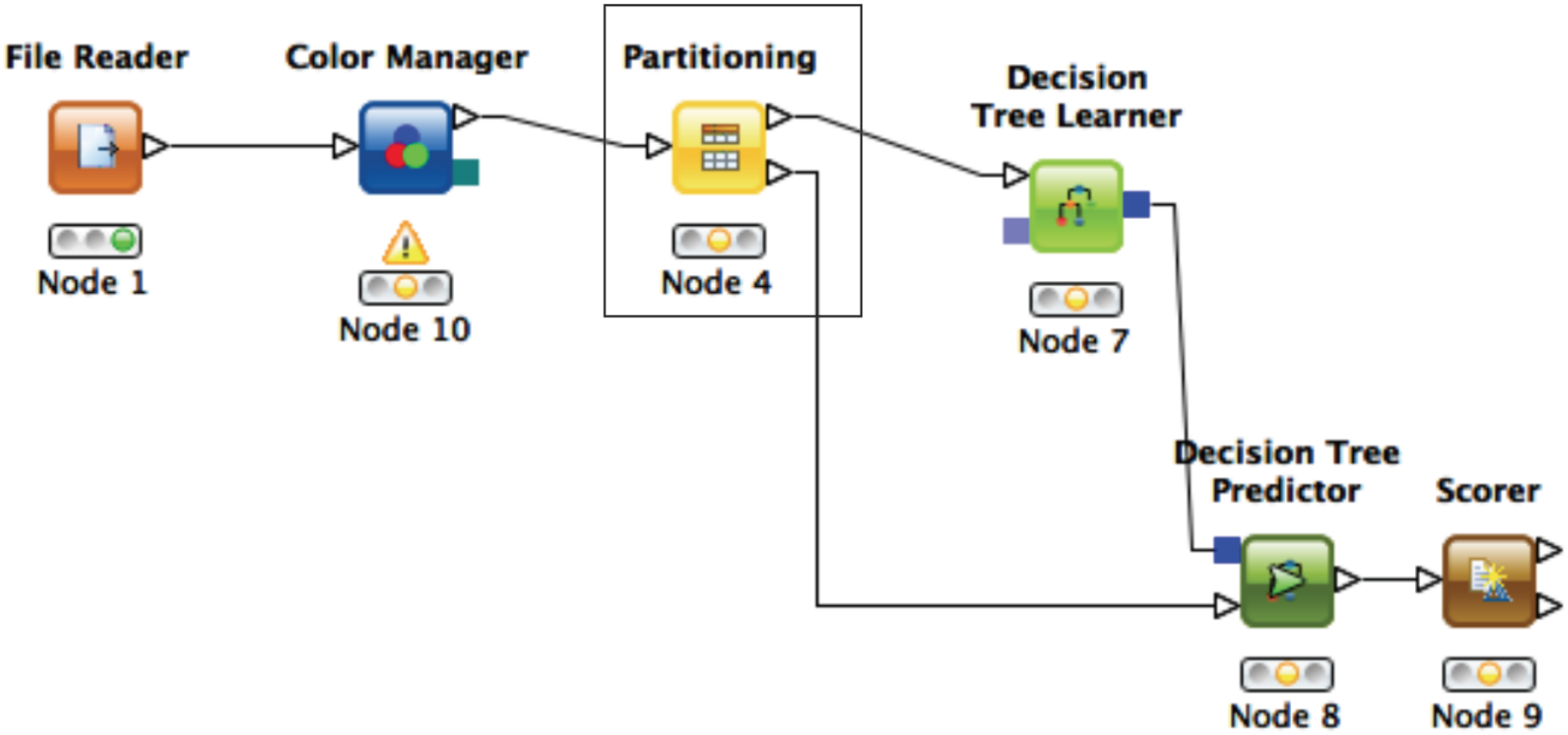
Árbol simple



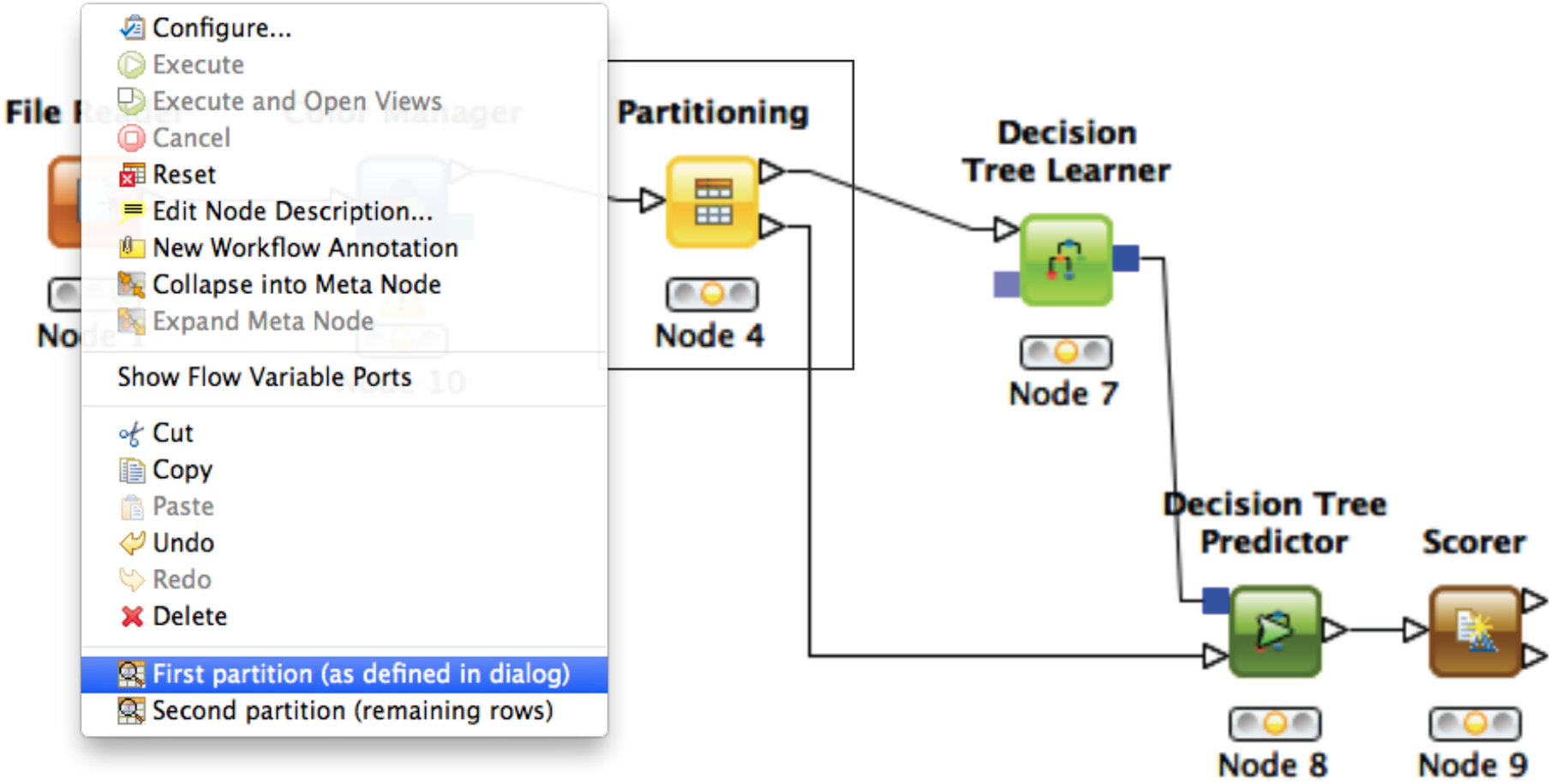
El entorno de trabajo en KNIME ► HiLiting



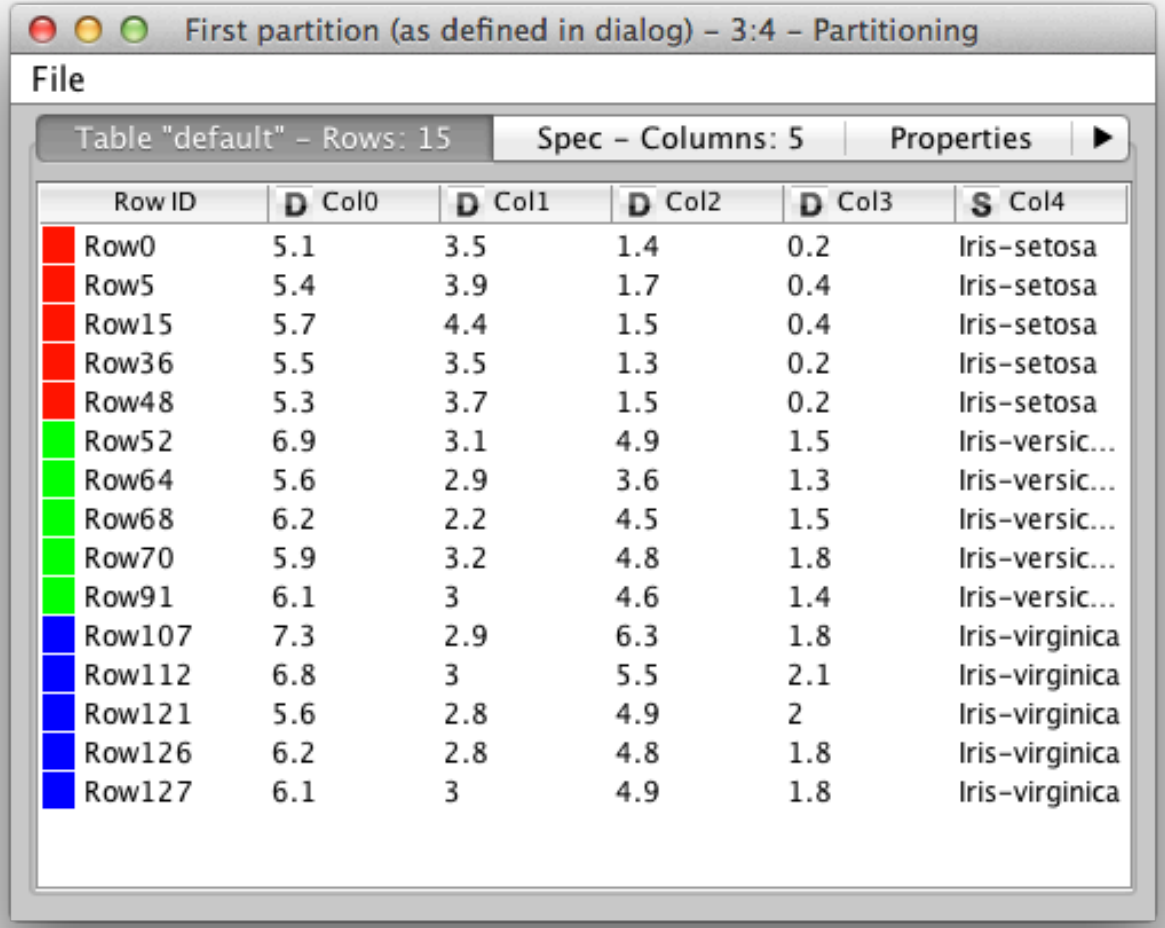
El entorno de trabajo en KNIME ► HiLiting



El entorno de trabajo en KNIME ► HiLiting

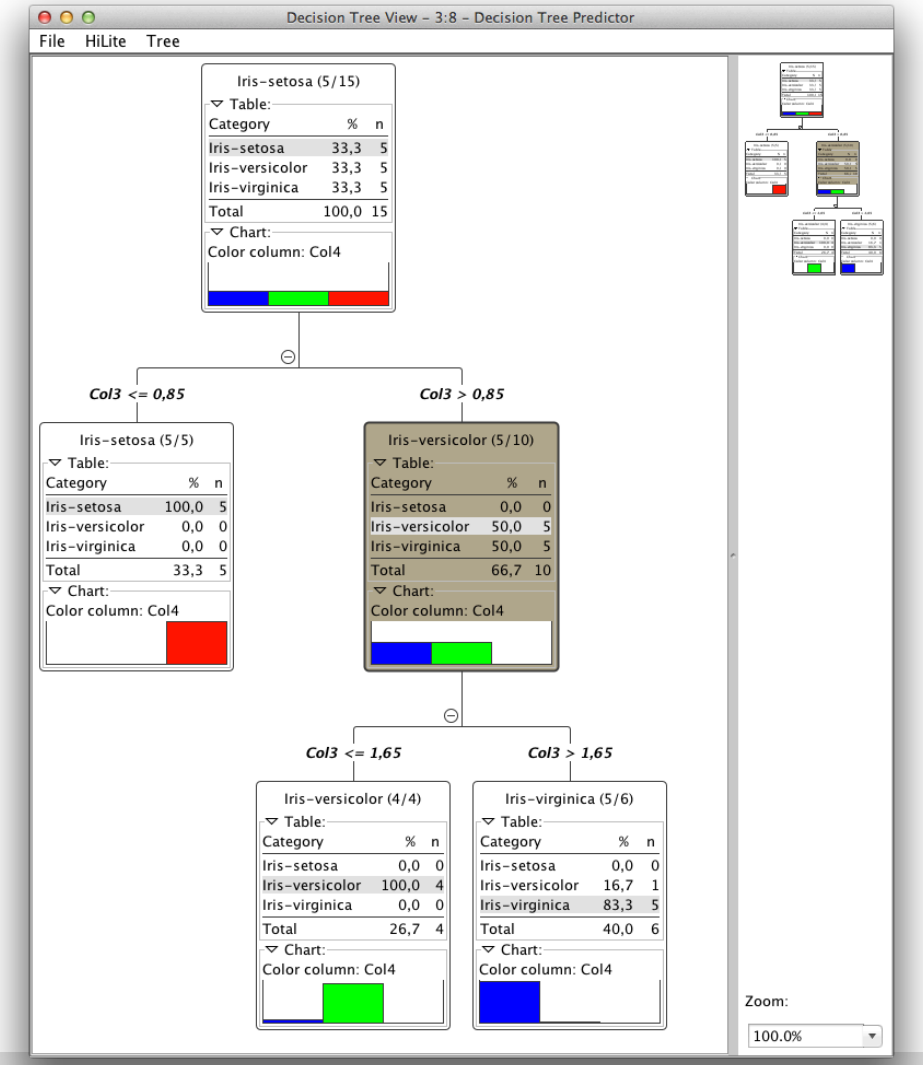


El entorno de trabajo en KNIME ► HiLiting

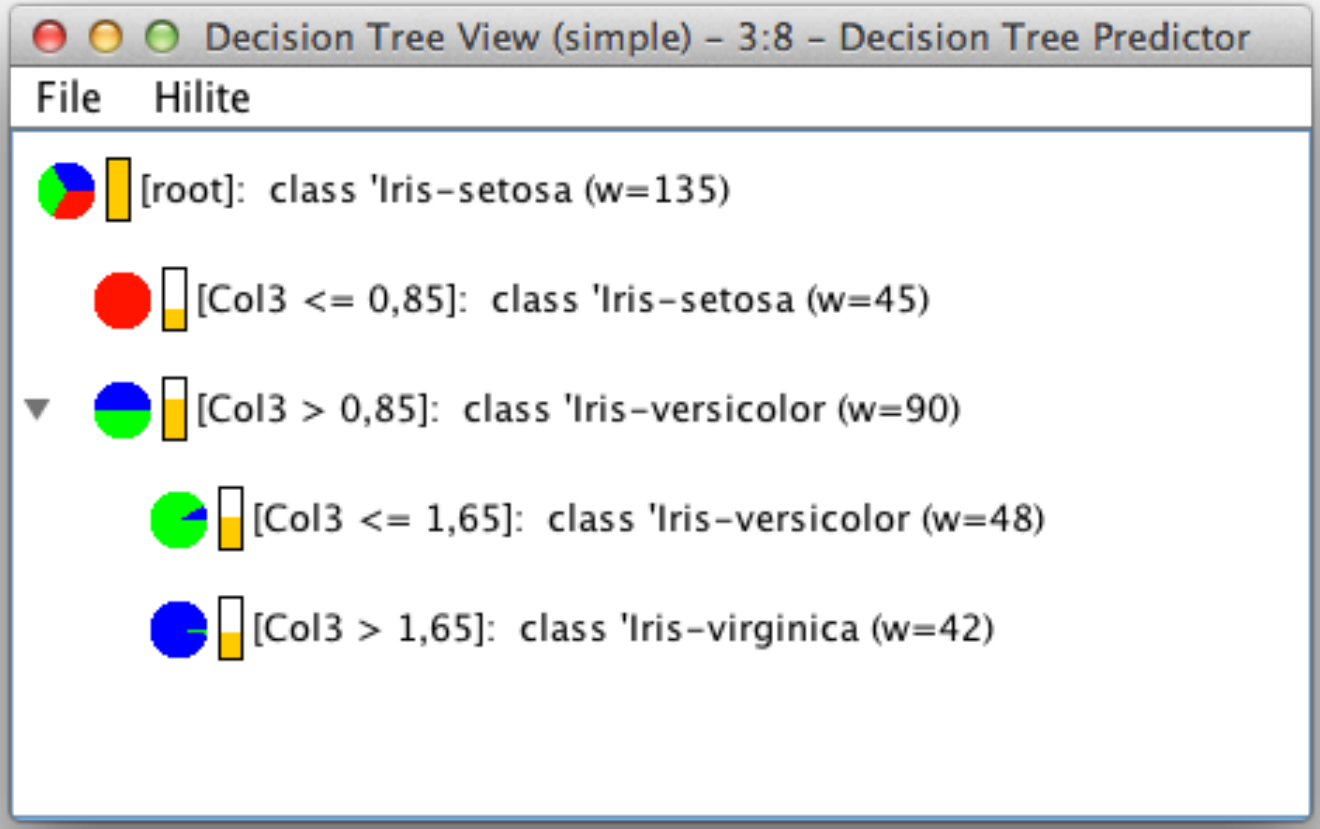


Row ID	D Col0	D Col1	D Col2	D Col3	S Col4
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row5	5.4	3.9	1.7	0.4	Iris-setosa
Row15	5.7	4.4	1.5	0.4	Iris-setosa
Row36	5.5	3.5	1.3	0.2	Iris-setosa
Row48	5.3	3.7	1.5	0.2	Iris-setosa
Row52	6.9	3.1	4.9	1.5	Iris-versic...
Row64	5.6	2.9	3.6	1.3	Iris-versic...
Row68	6.2	2.2	4.5	1.5	Iris-versic...
Row70	5.9	3.2	4.8	1.8	Iris-versic...
Row91	6.1	3	4.6	1.4	Iris-versic...
Row107	7.3	2.9	6.3	1.8	Iris-virginica
Row112	6.8	3	5.5	2.1	Iris-virginica
Row121	5.6	2.8	4.9	2	Iris-virginica
Row126	6.2	2.8	4.8	1.8	Iris-virginica
Row127	6.1	3	4.9	1.8	Iris-virginica

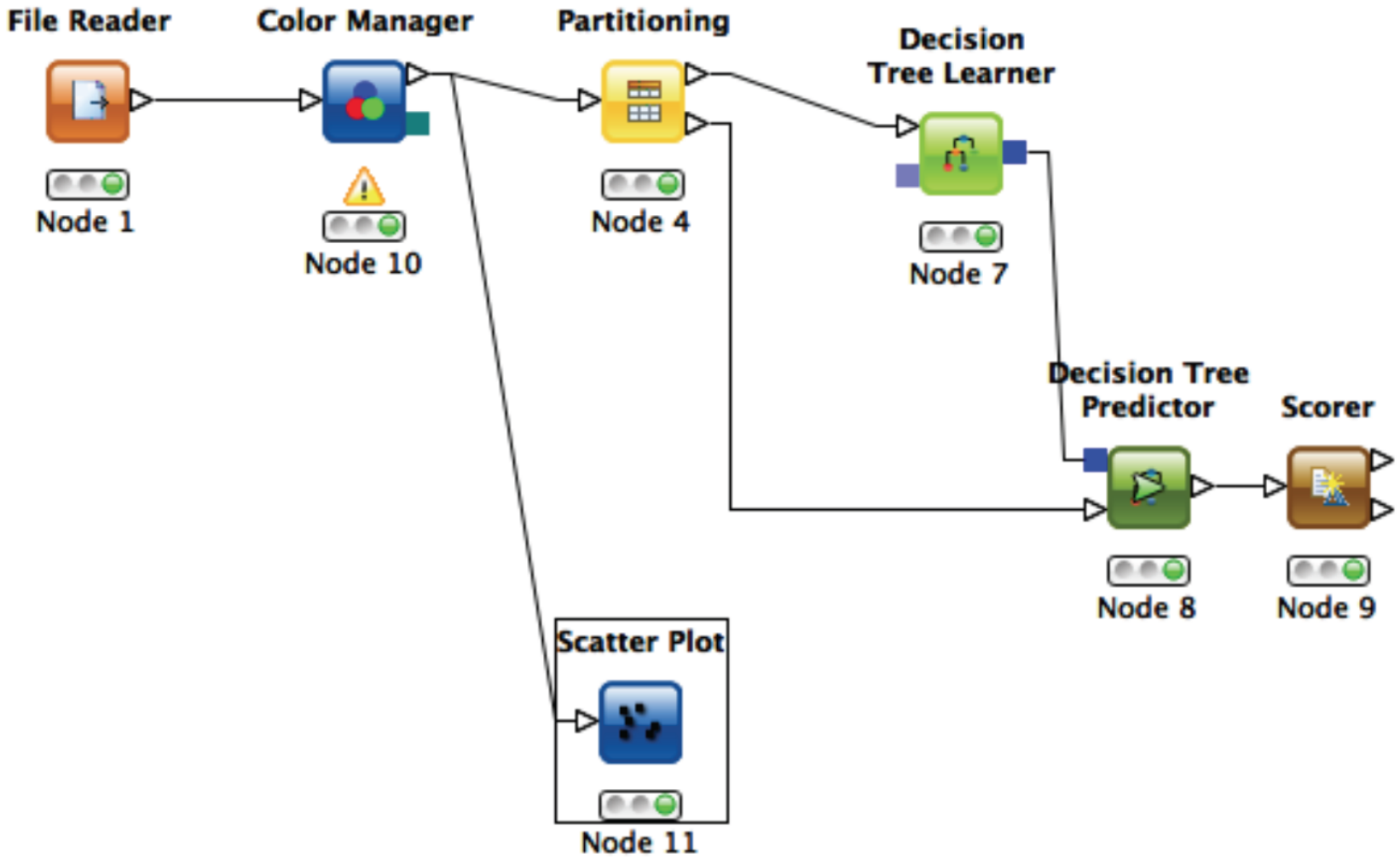
El entorno de trabajo en KNIME ► HiLiting



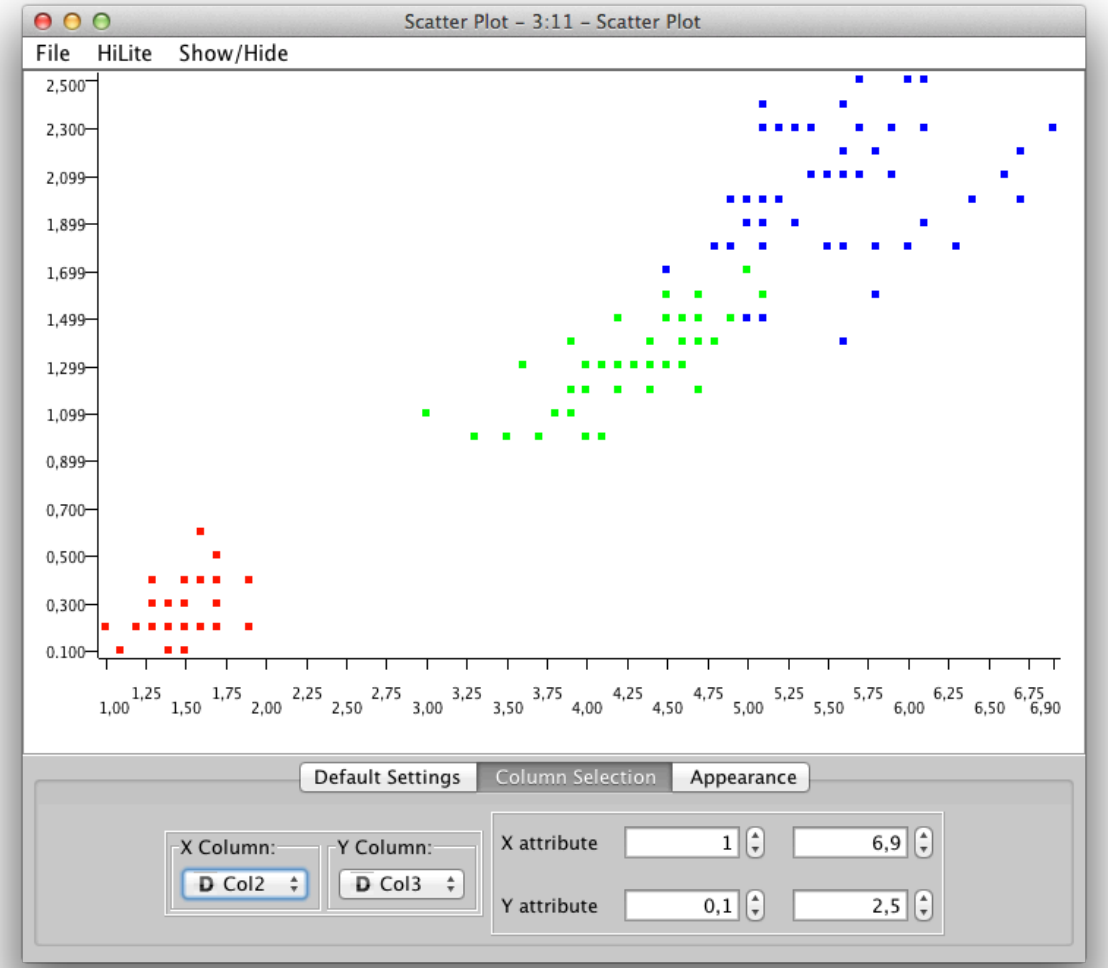
El entorno de trabajo en KNIME ► HiLiting



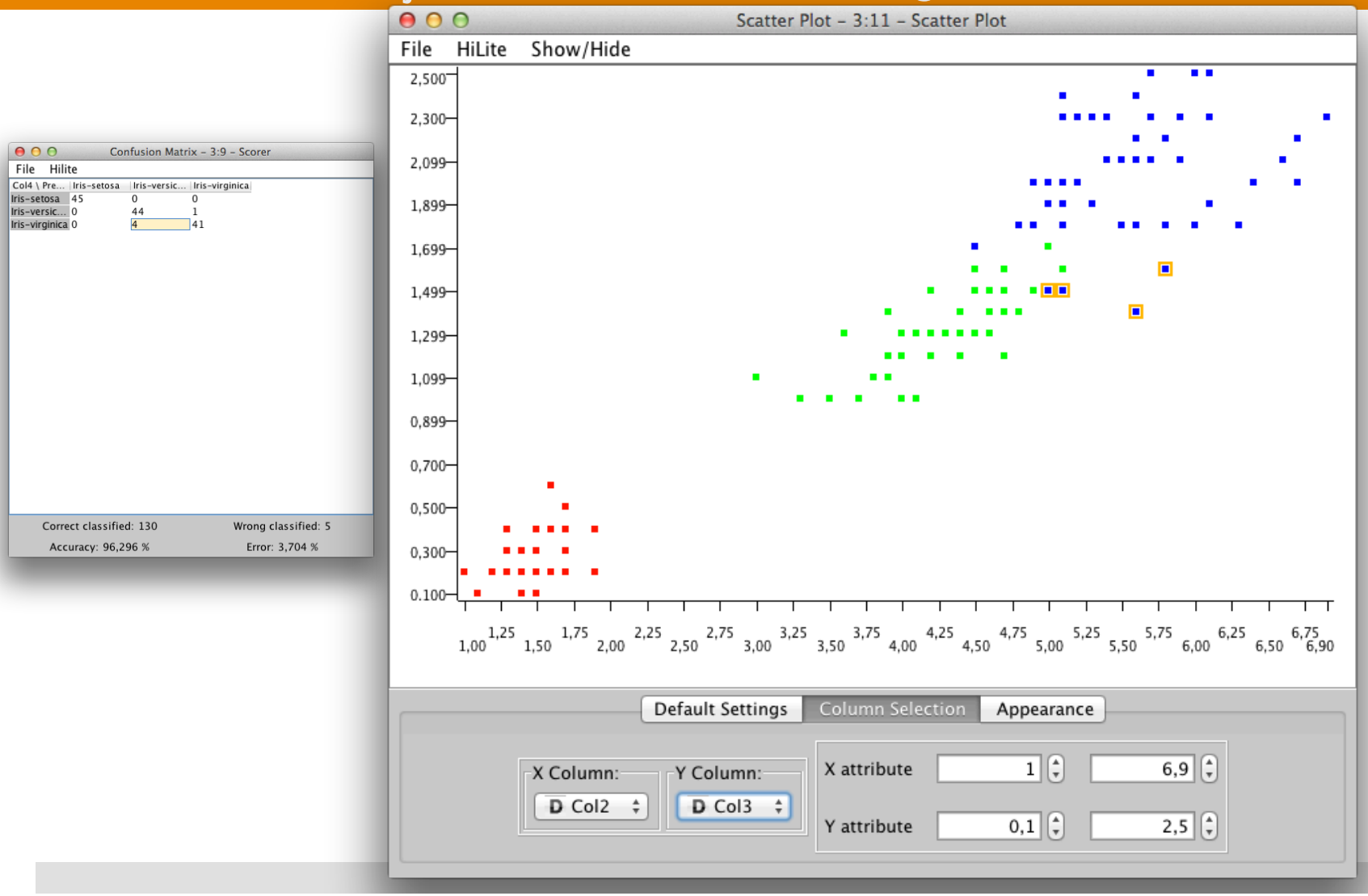
El entorno de trabajo en KNIME ► HiLiting



El entorno de trabajo en KNIME ► HiLiting



El entorno de trabajo en KNIME ▶ HiLiting



El entorno de trabajo en KNIME ► Hotkeys

Task	Hotkey	Description
Node Configuration	F6	opens the configuration dialog of a node
	F7	executes selected nodes
Node Execution	Shift + F7	executes all configured nodes
	Shift + F10	executes configured nodes and opens all views
	F9	cancels selected running nodes
	Shift + F9	cancels all running nodes
Move Nodes and Annotations	Ctrl + Shift + Arrow	moves a selected node in the workflow editor
	Ctrl + Shift + PgUp/PgDown	Moves the selected up or down in z order
	F8	resets selected nodes
Workflow Operations	Ctrl + S	Saves the workflow
	Ctrl + Shift + S	Saves all open workflows
	Ctrl + Shift + W	Closes all open workflows
Meta-node	Shift + F12	Opens meta-node wizard

Ejemplo: Iris

- Carga de datos
- Visualización
- Análisis predictivo
- Análisis descriptivo

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

- Nº ejemplos: 150
- Nº variables: 4
- Nº clases: 3 (50/50/50)

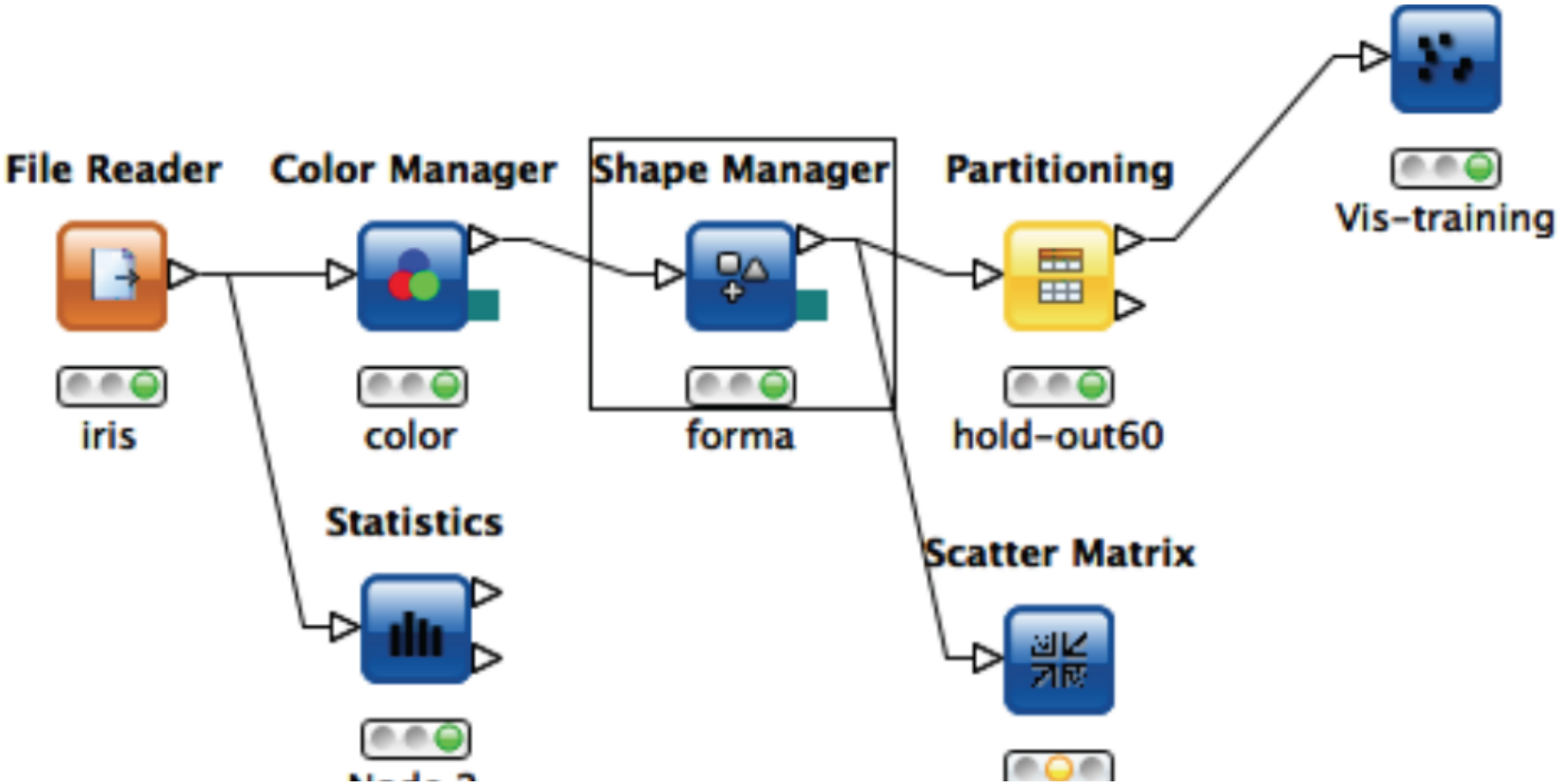
Ejemplo: Iris

1. Cargar el fichero de ejemplos iris.dat
2. Obtener medidas estadísticas
3. Asignar a cada ejemplo un color en función de la clase a la que pertenece
4. Visualizar el conjunto de ejemplos en base a pares de variables
   - Determinar el par de variables "más relevantes"
5. Realizar una partición con hold-out al 60% estratificada
6. Visualizar el conjunto de test en base a las dos variables seleccionadas en el paso 4

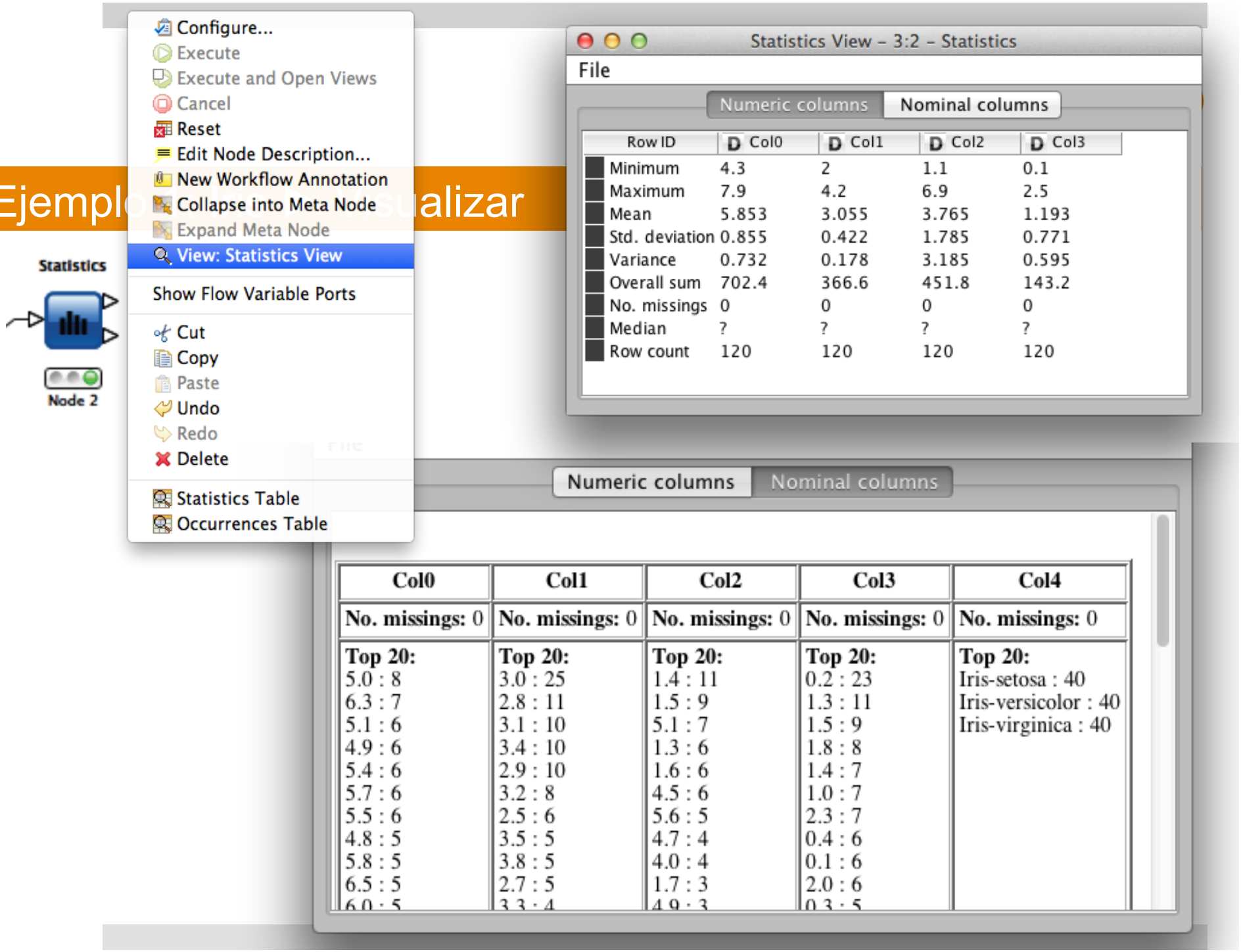
Ejemplo: Iris

1. File Reader: IO → Read
2. Statistics: Statistics
   - Calcula y muestra estadísticas
3. Color Manager: Data Views → Property
   - Asigna colores a las clases
4. Scatter Matrix : Data Views → Utility
   - Visualiza los ejemplos según pares de variables (scatter plots)
5. Partitioning: Data Manipulation → Row → Transform
   - Hold-out
6. Scatter Plot: Data Views → Utility

Ejemplo 1: Iris ► Carga de datos y visualización

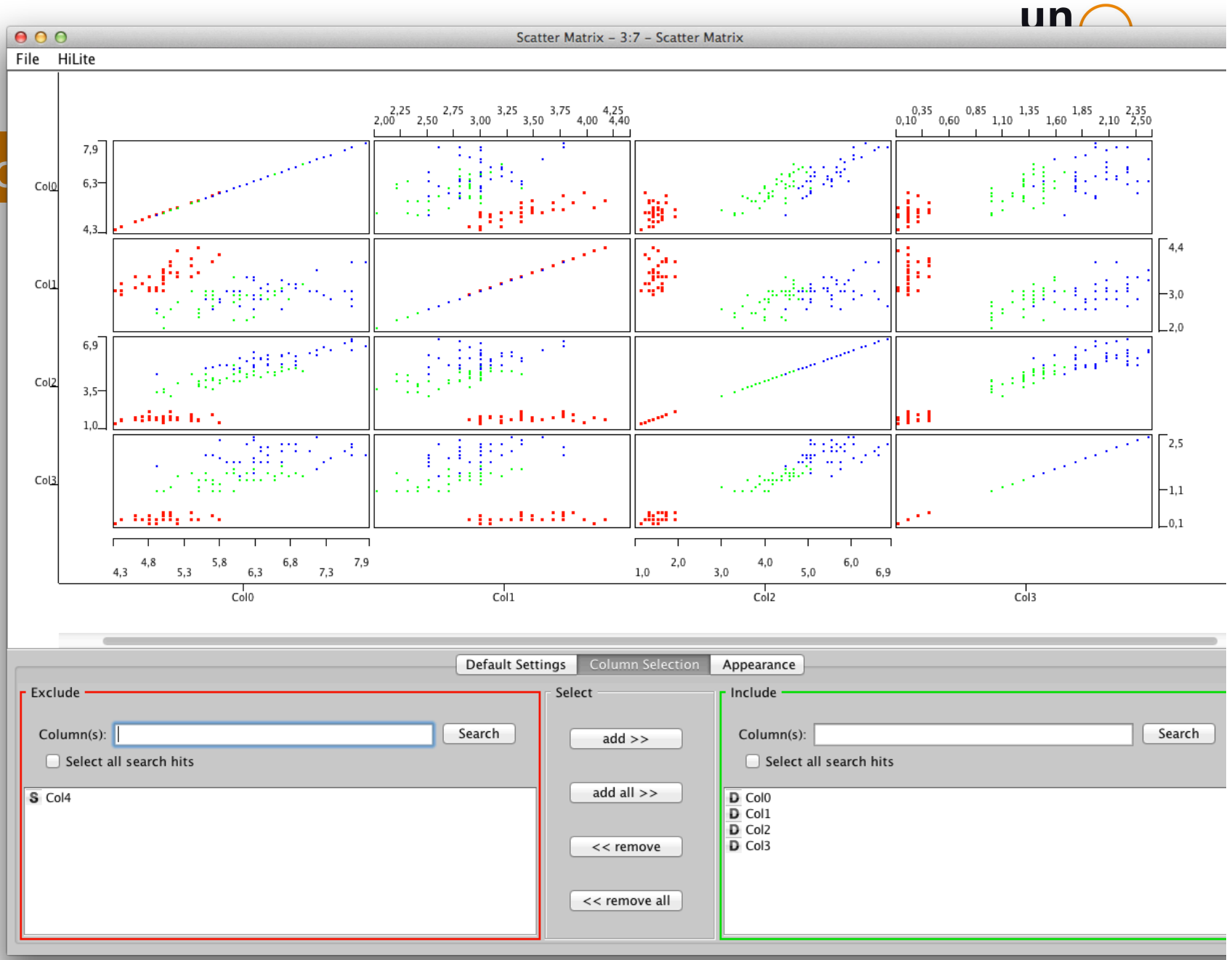


Ejemplo de cómo visualizar

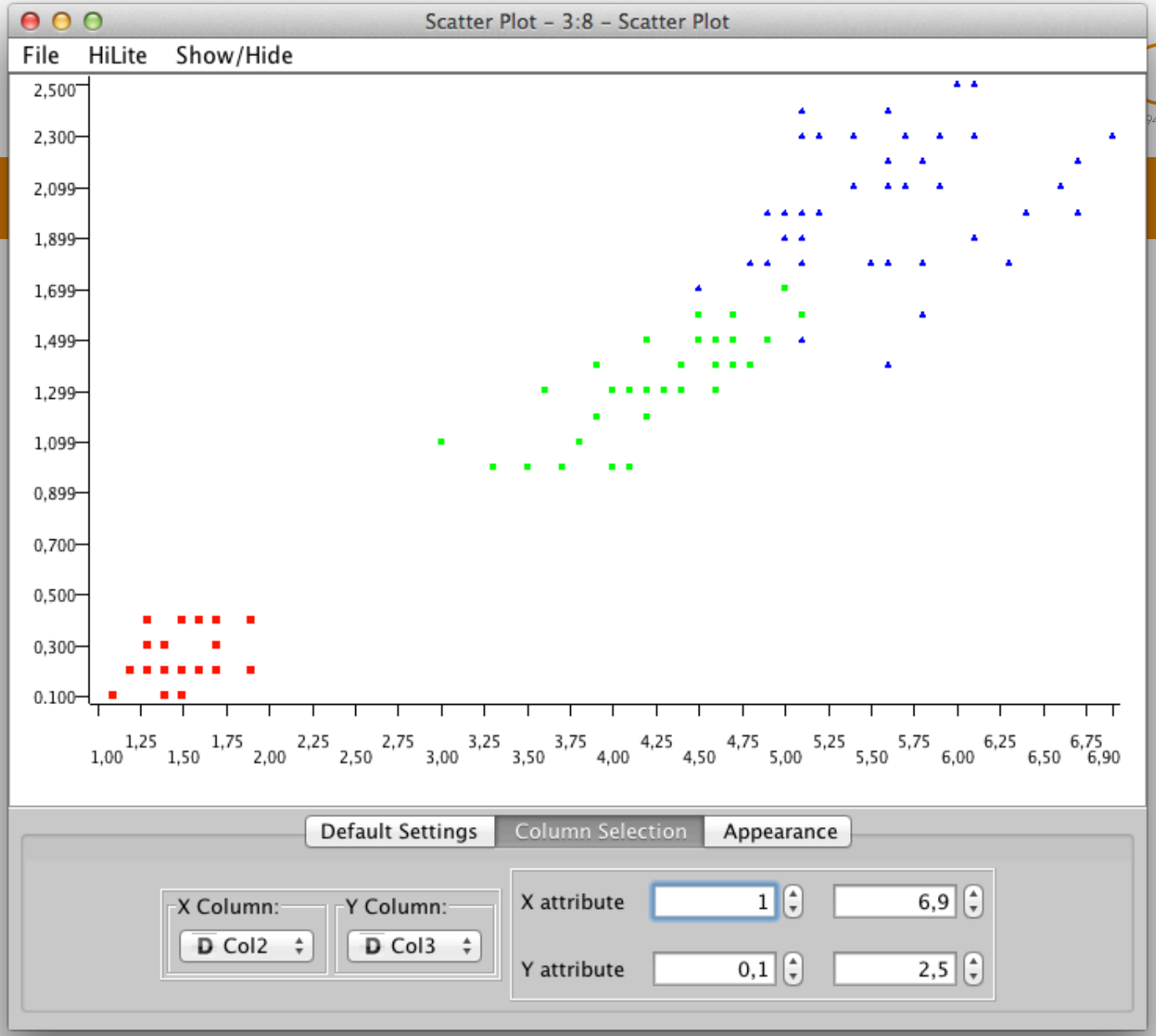


Ejemplo


Scatter Matrix
Node 7


Scatter Matrix – 3:7 – Scatter Matrix
File
HiLite
Col0
Col1
Col2
Col3
Default Settings
Column Selection
Appearance
Exclude
Column(s):
Search
Select all search hits
S Col4
Select
add >>
add all >>
<< remove
<< remove all
Include
Column(s):
Search
Select all search hits
D Col0
D Col1
D Col2
D Col3

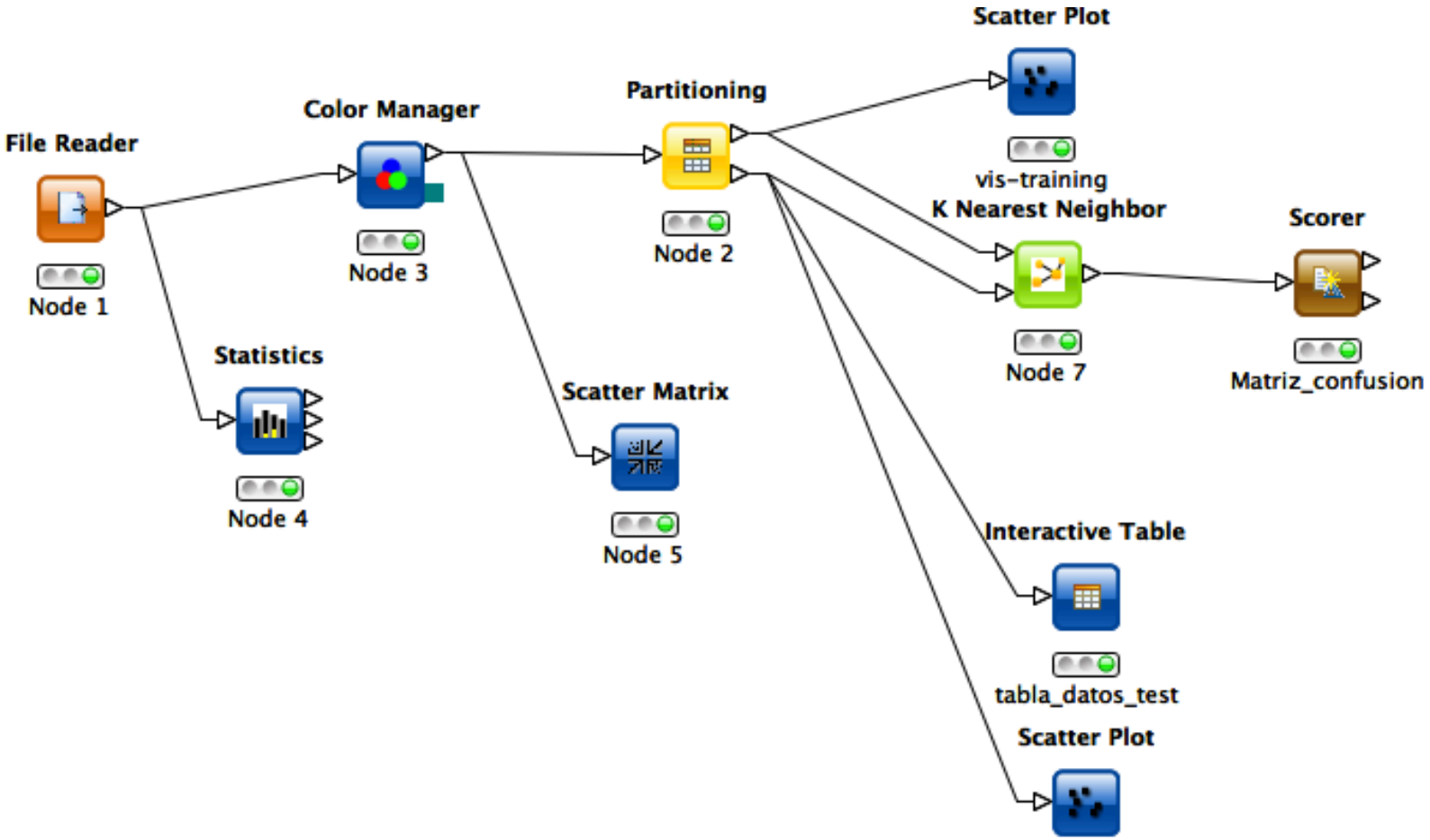
Ejemplo 1: Iris



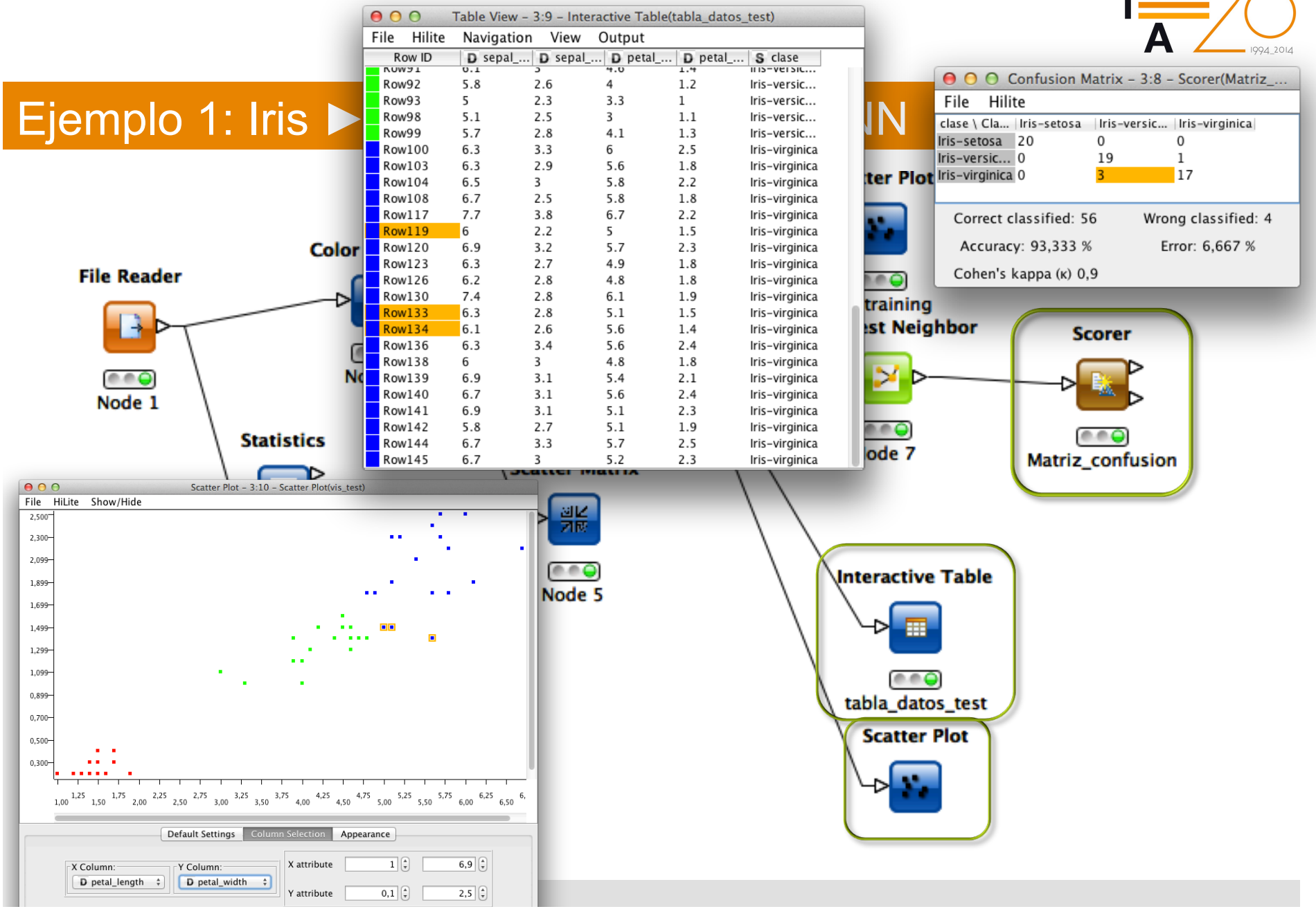
Ejemplo: Iris

- Para agilizar: hold-out 60% estratificado
- 1-NN (k Nearest Neigbor: Mining → Misc Classifiers)
  - Visualizar resultados (Scorer: Mining → Scoring)
  - Marcar y visualizar errores en test
    - Scatter Plot
    - Interactive Table: Data Views → Utility
  - ¿Funcionará mejor con otro valor de k?
  - ¿Funcionará mejor ponderando por distancias?
- Decision trees
- Random forest
- Naïve Bayes

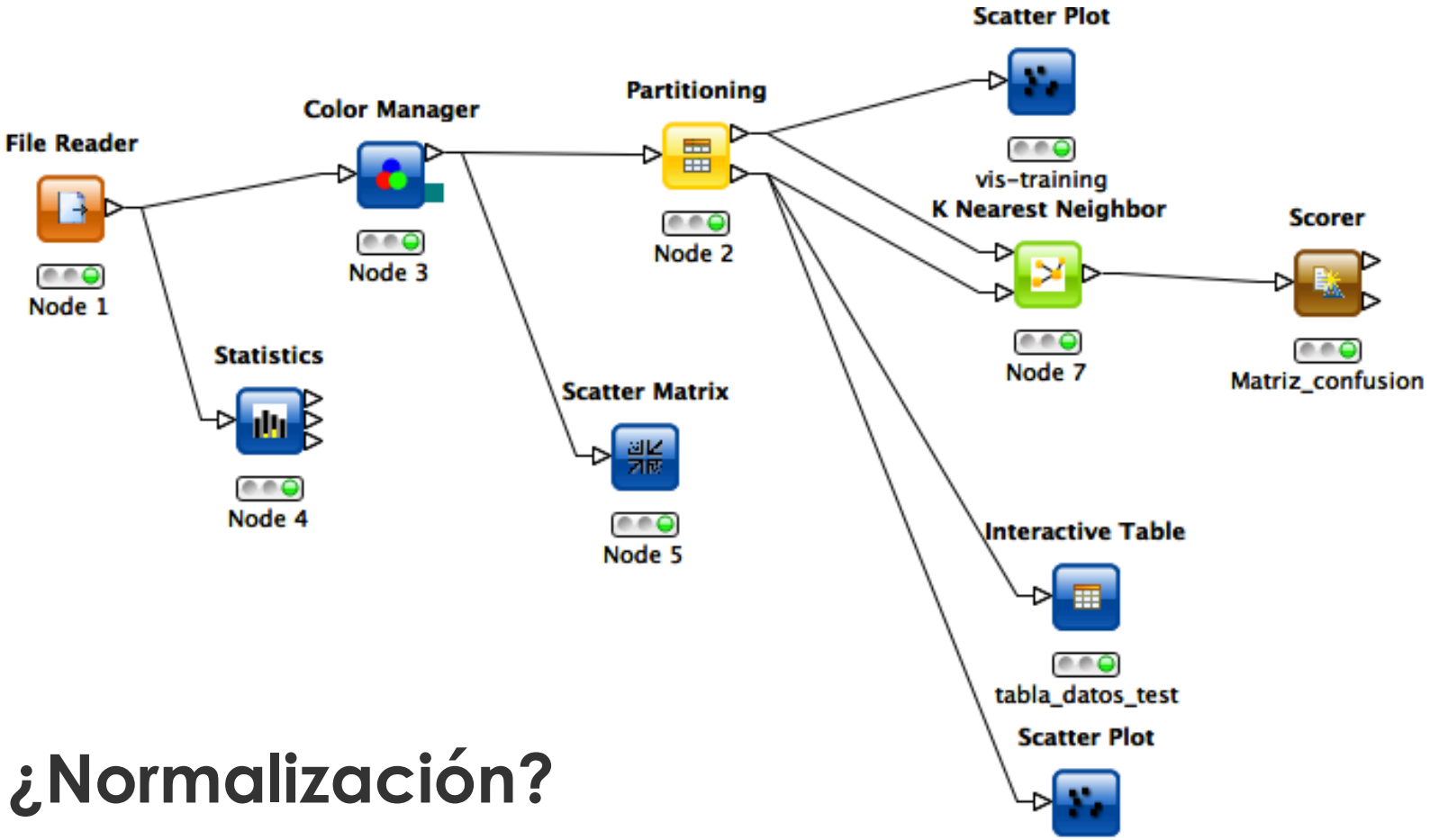
Ejemplo 1: Iris ► Análisis predictivo ► kNN



Ejemplo 1: Iris



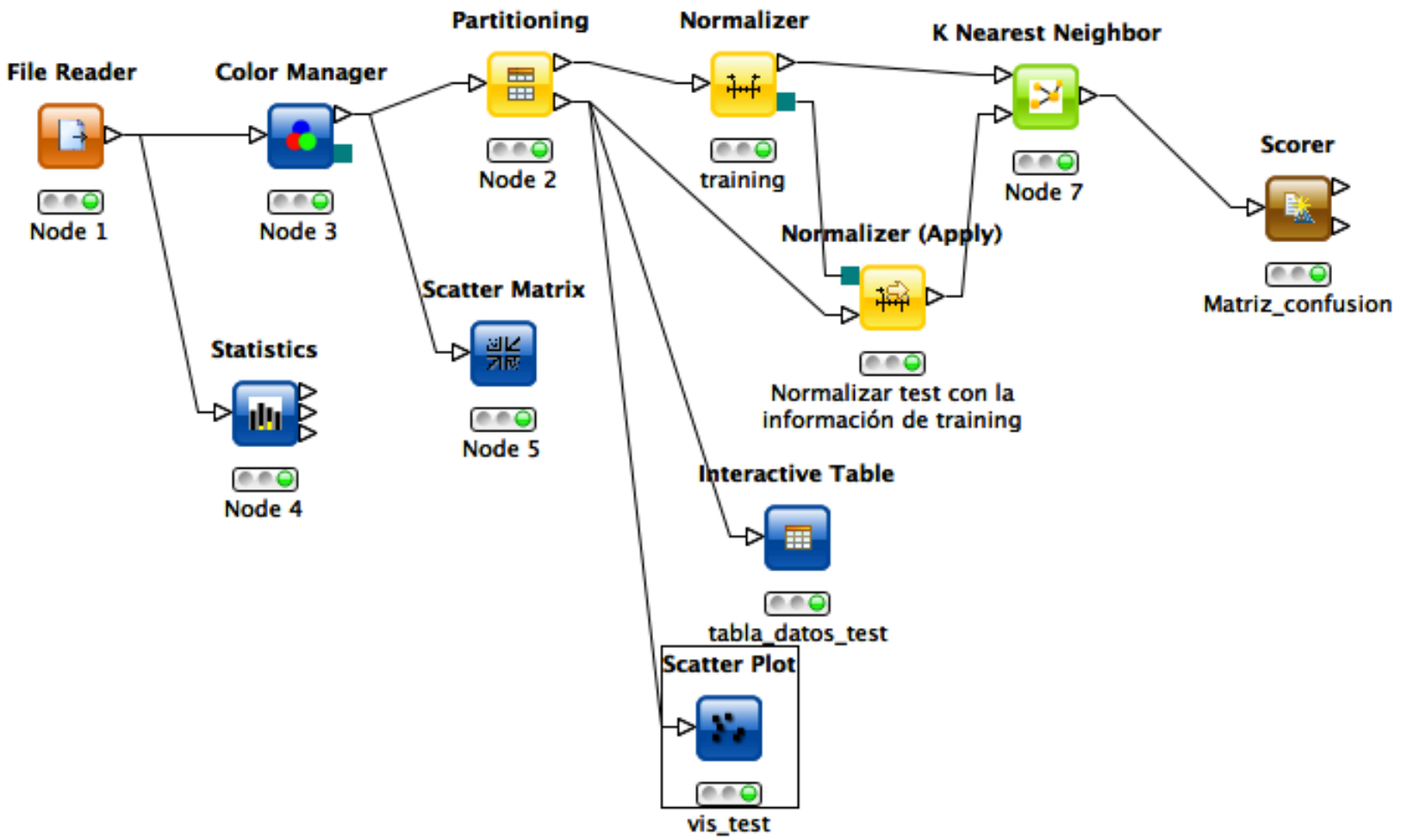
Ejemplo 1: Iris ► Análisis predictivo ► kNN



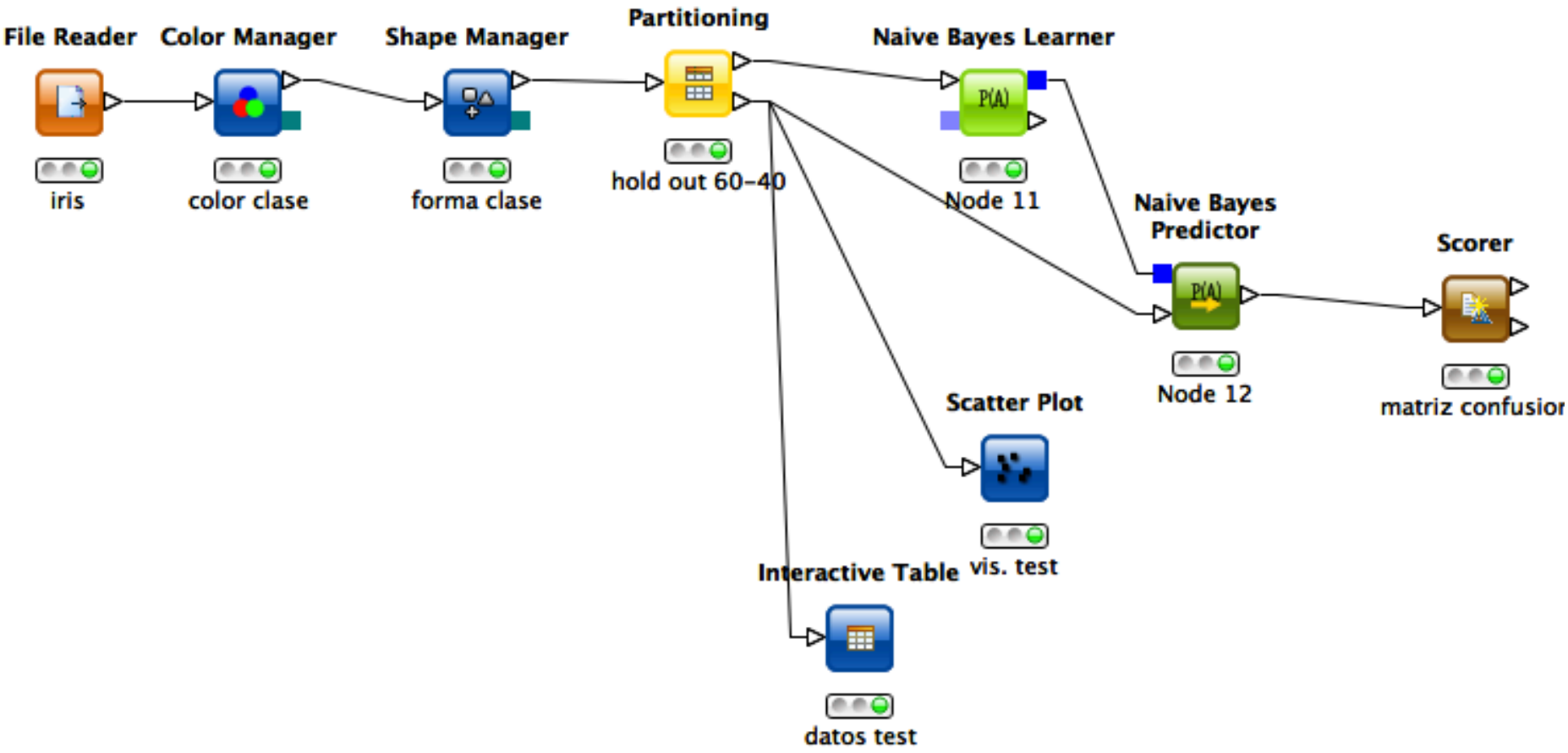
¿Normalización?

Normalizer: Data Manipulation → Column → Transform

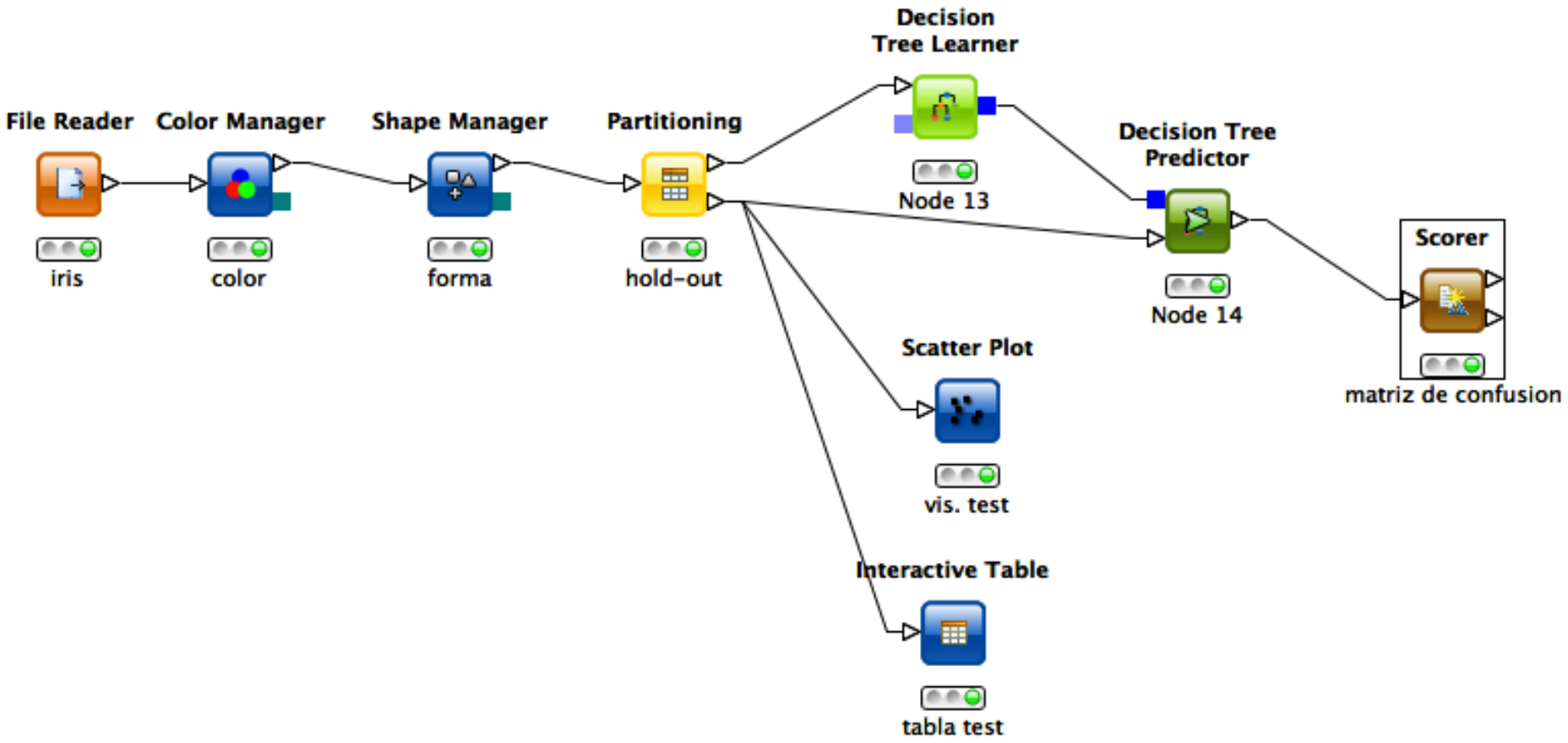
Ejemplo 1: Iris ► Análisis predictivo ► kNN

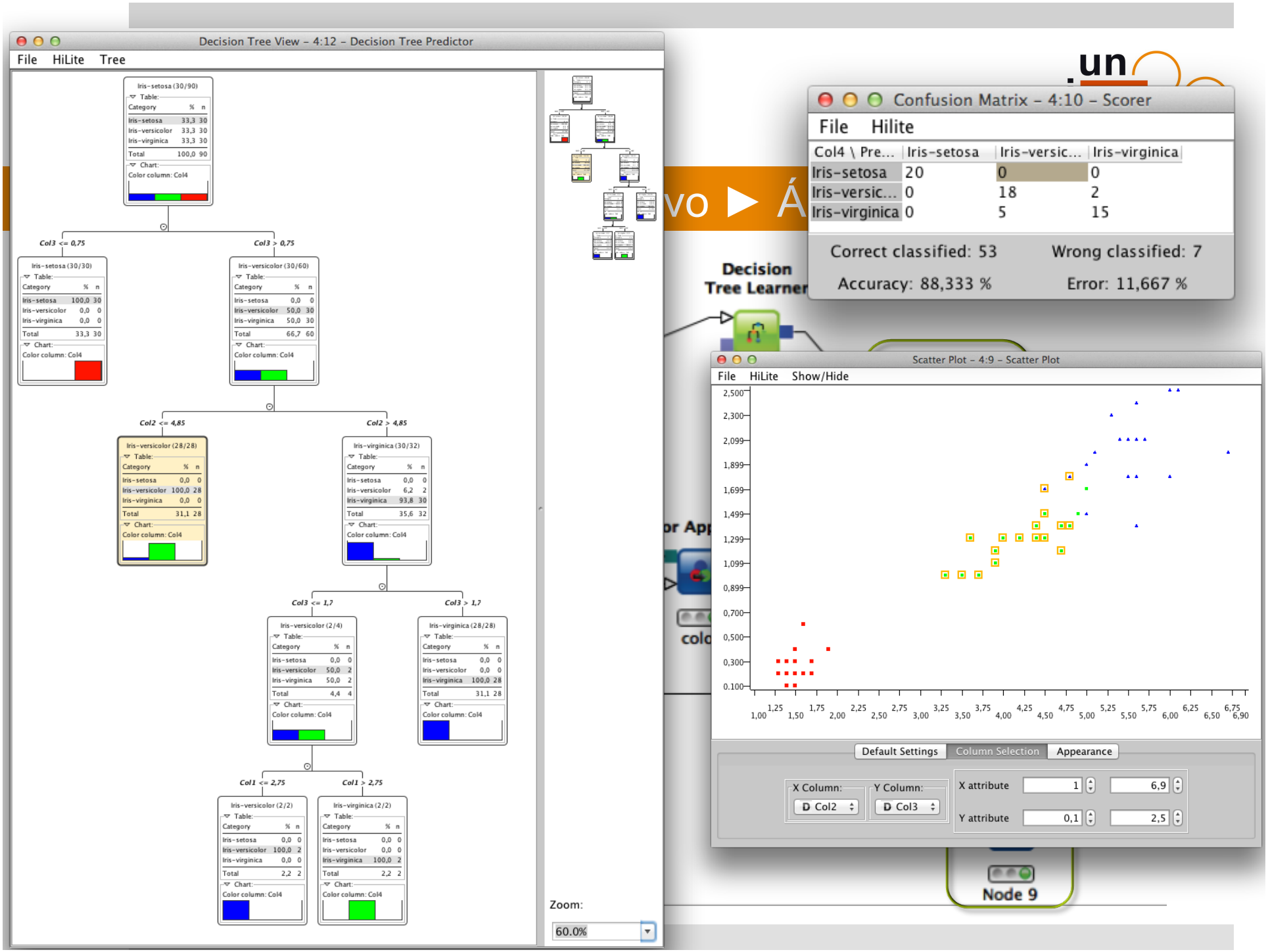


Ejemplo 1: Iris ► Análisis predictivo ► Naïve Bayes



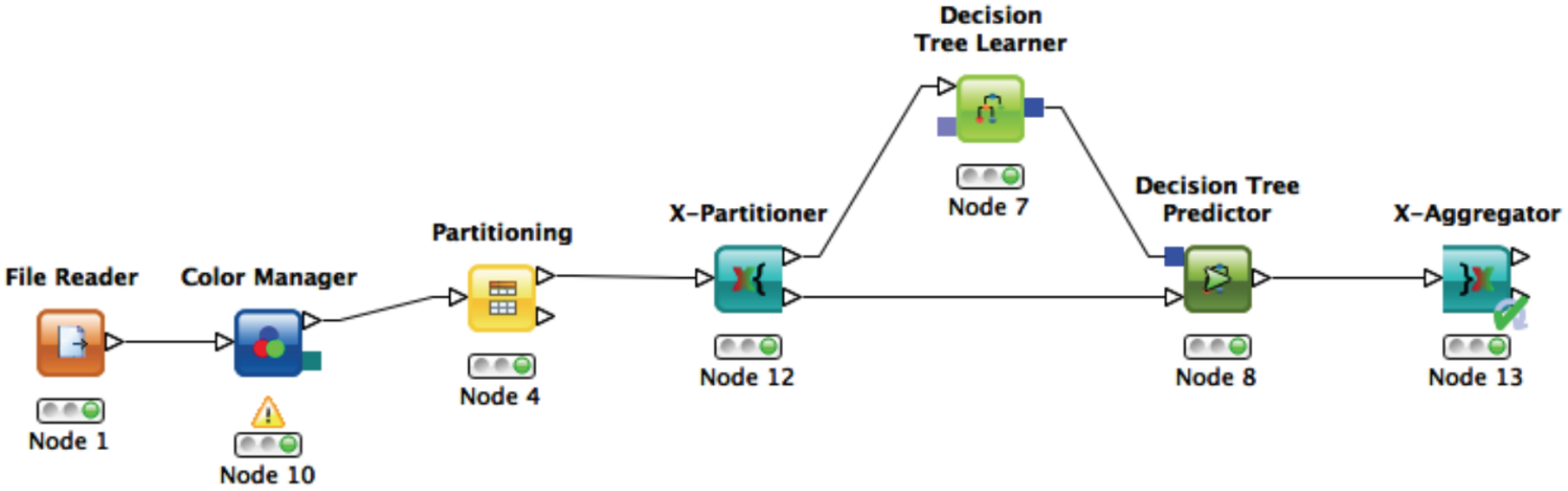
Ejemplo 1: Iris ► Análisis predictivo ► Árboles de decisión



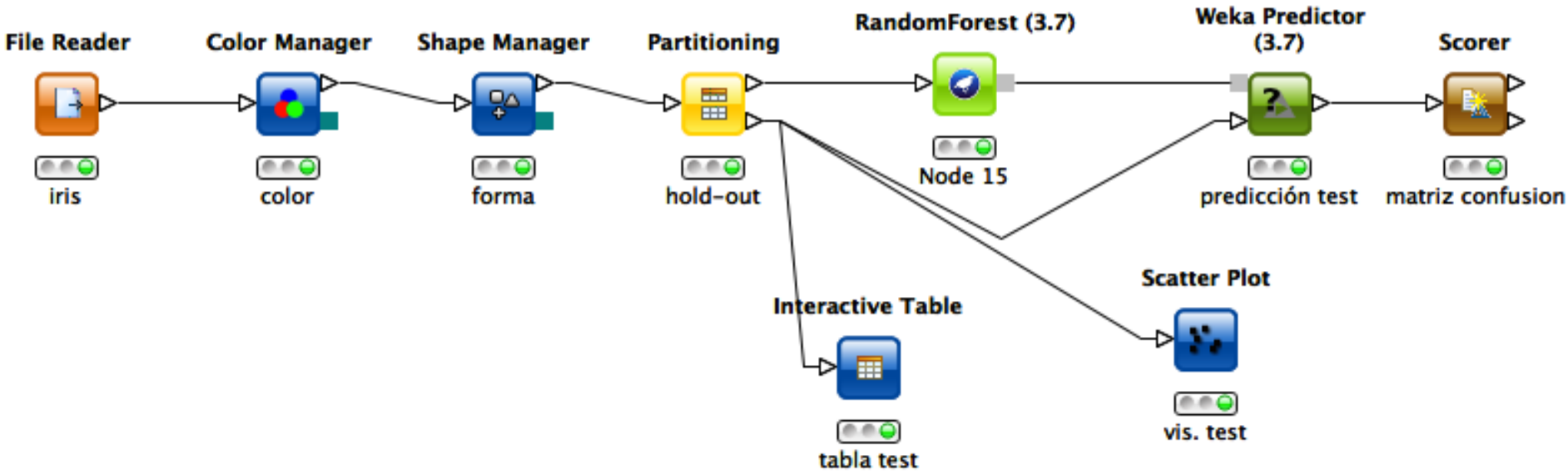

Decision Tree View – 4:12 – Decision Tree Predictor
File HiLite Tree
Iris-setosa (30/90)
Col3 <= 0,75
Col3 > 0,75
Iris-setosa (30/30)
Iris-versicolor (30/60)
Col2 <= 4,85
Col2 > 4,85
Iris-versicolor (28/28)
Iris-virginica (30/32)
Col3 <= 1,7
Col3 > 1,7
Iris-versicolor (2/4)
Iris-virginica (28/28)
Col1 <= 2,75
Col1 > 2,75
Iris-versicolor (2/2)
Iris-virginica (2/2)
Zoom:
60.0%
Confusion Matrix – 4:10 – Scorer
File Hilite
Col4 \ Pre... Iris-setosa Iris-versic... Iris-virginica
Iris-setosa 20 0 0
Iris-versic... 0 18 2
Iris-virginica 0 5 15
Correct classified: 53
Wrong classified: 7
Accuracy: 88,333 %
Error: 11,667 %
Decision Tree Learner
Scatter Plot – 4:9 – Scatter Plot
File HiLite Show/Hide
Default Settings
Column Selection
Appearance
X Column:
Col2
Y Column:
Col3
X attribute
Y attribute
Node 9

Ejemplo 1: Iris ► Análisis predictivo

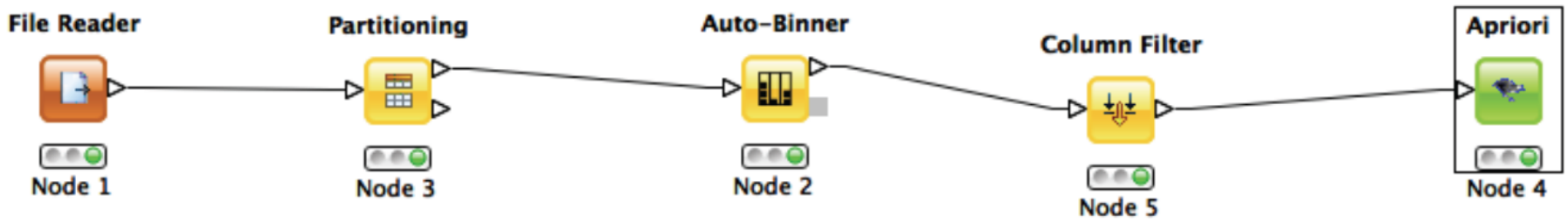
Los ciclos de validación cruzada se verán en los casos de estudio



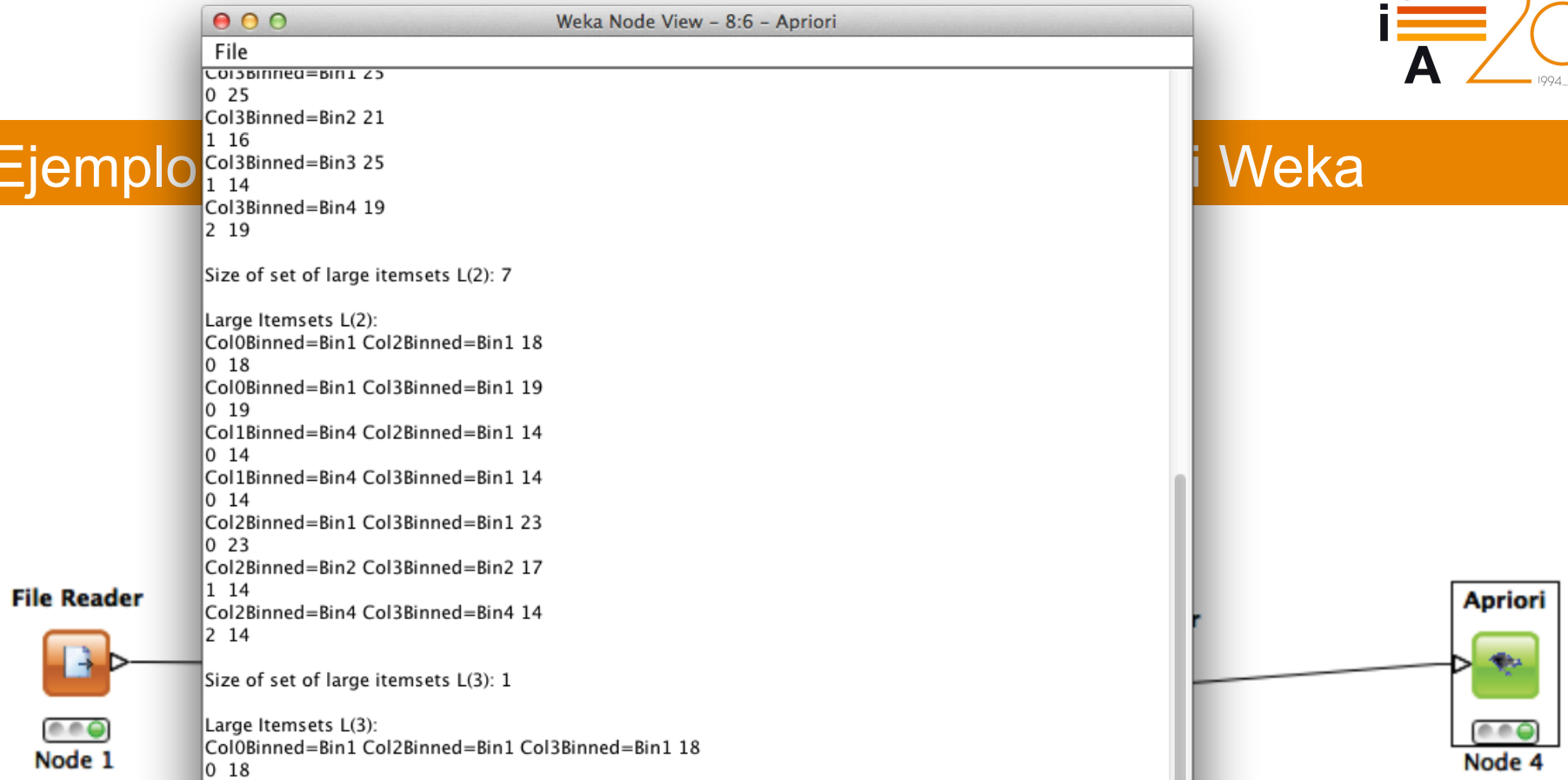
Ejemplo 1: Iris ► Análisis predictivo ► Random Forest



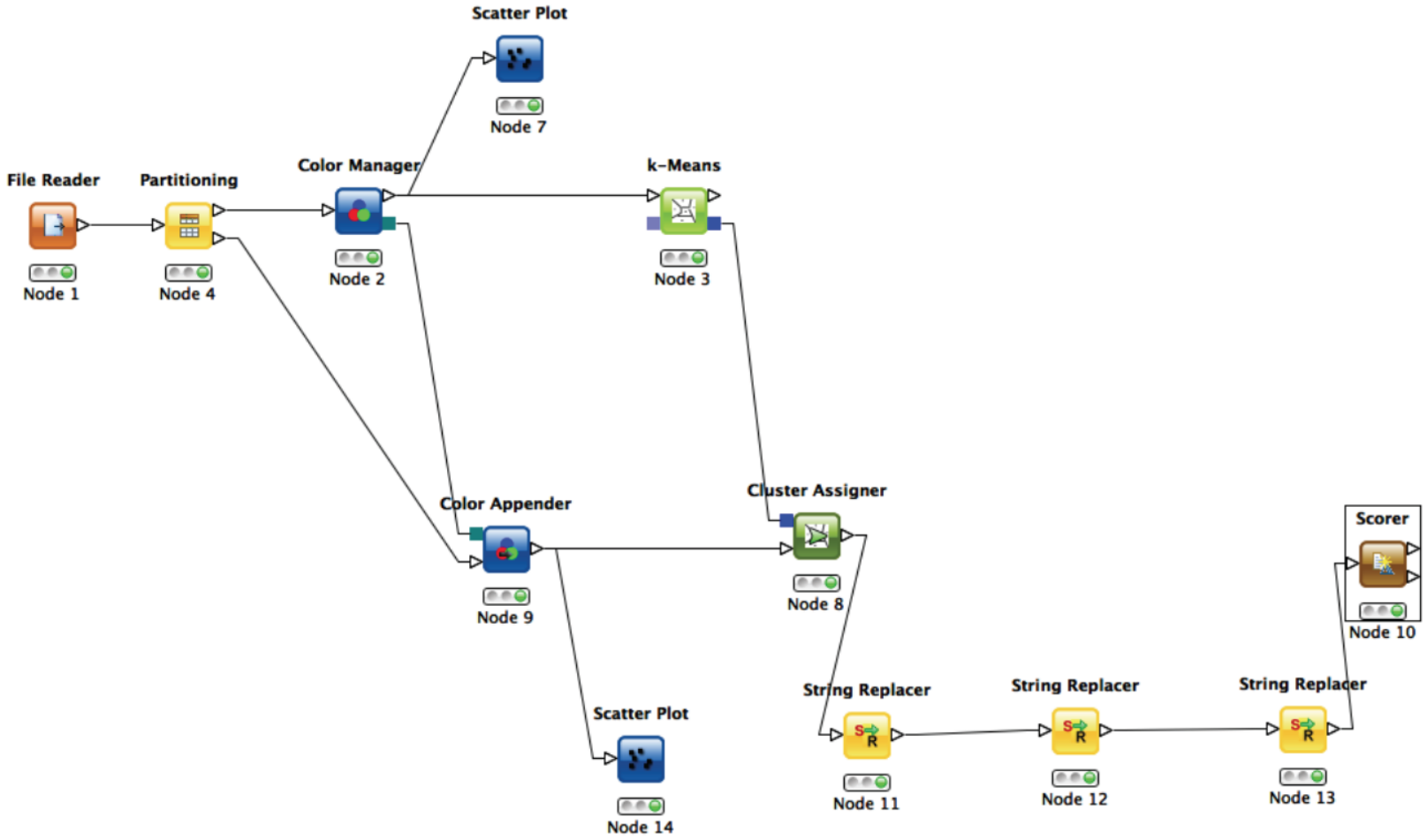
Ejemplo 1: Iris ► Análisis descriptivo ► A priori Weka



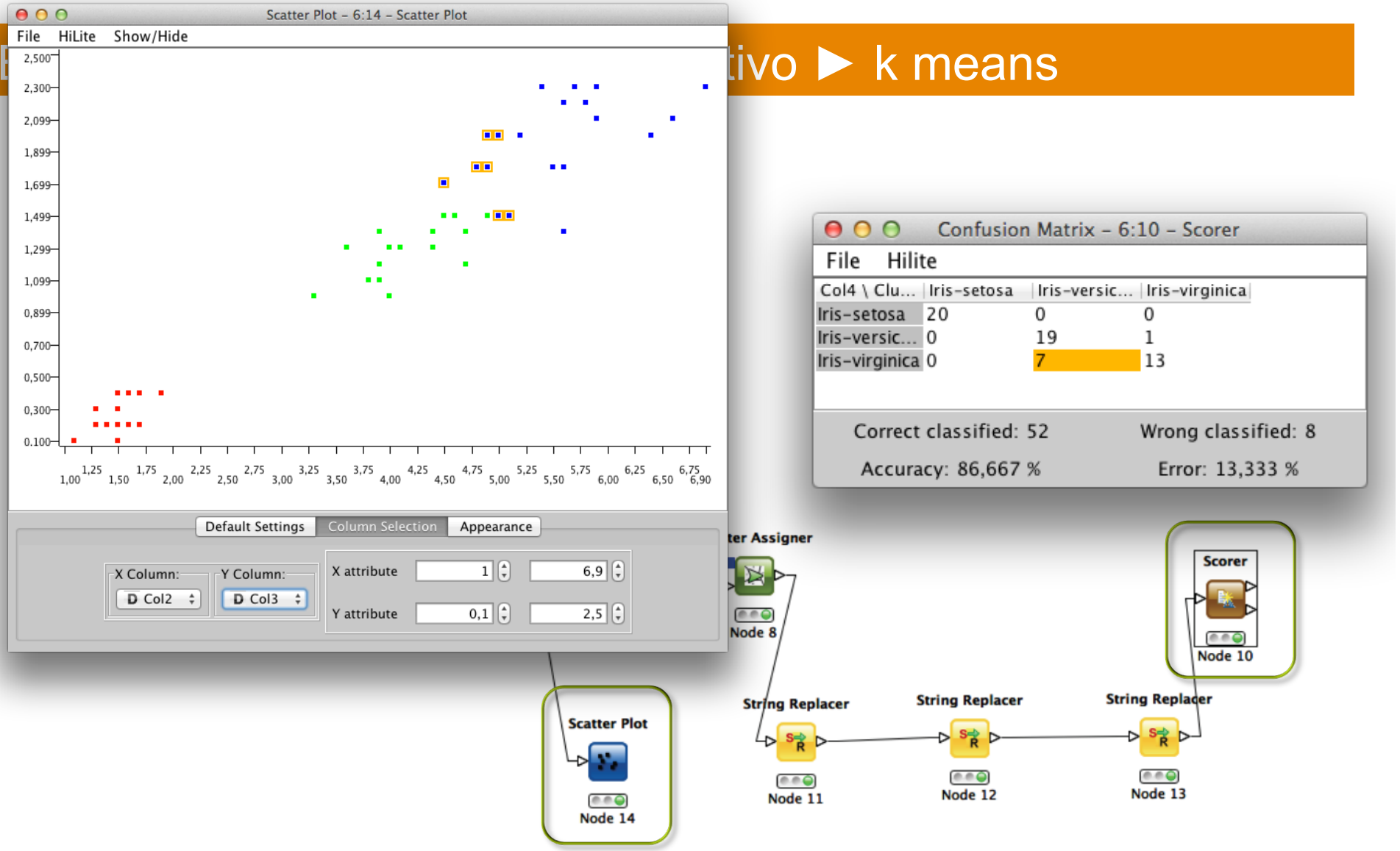
Ejemplo … i Weka



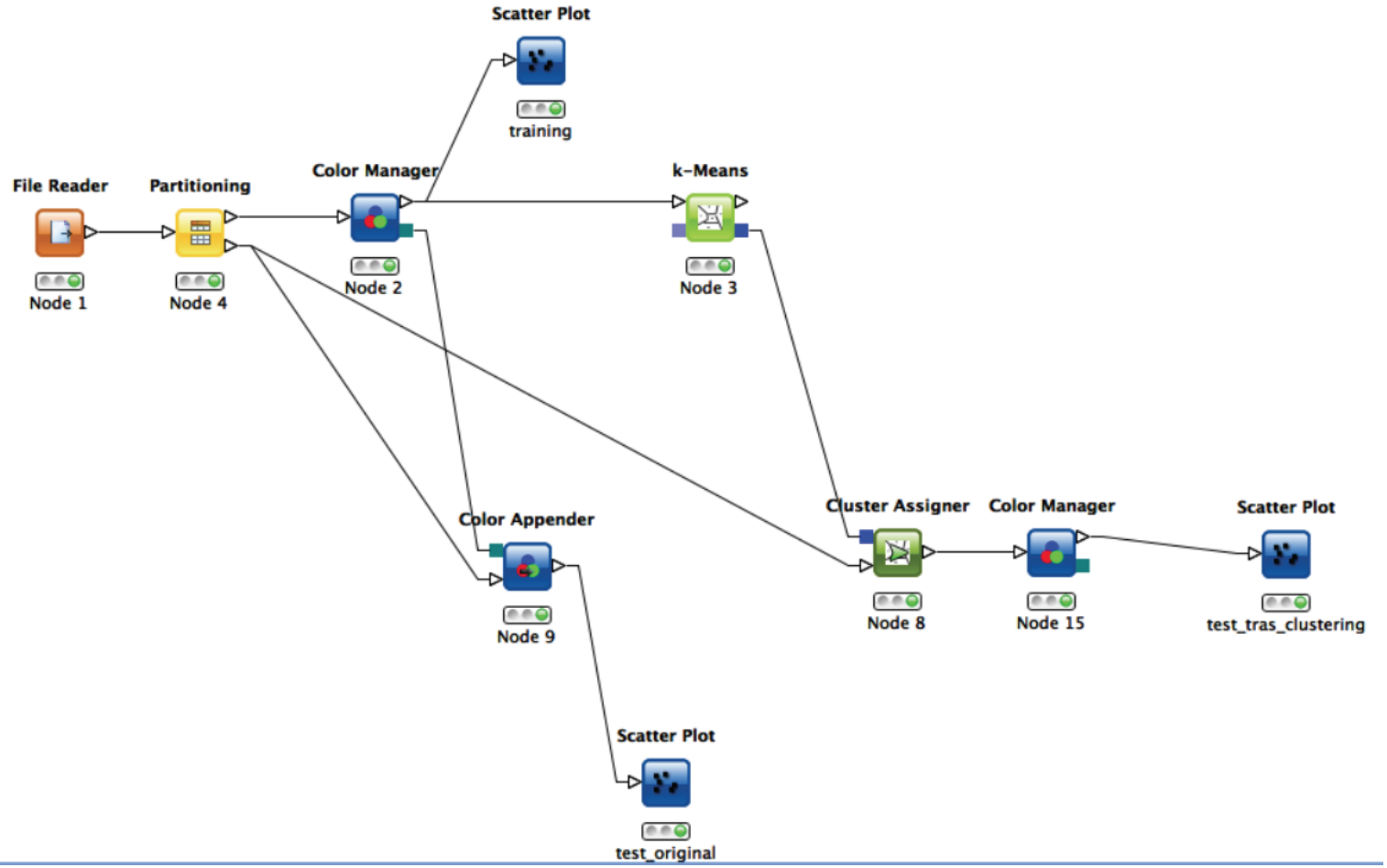
Ejemplo 1: Iris ► Análisis descriptivo ► k means

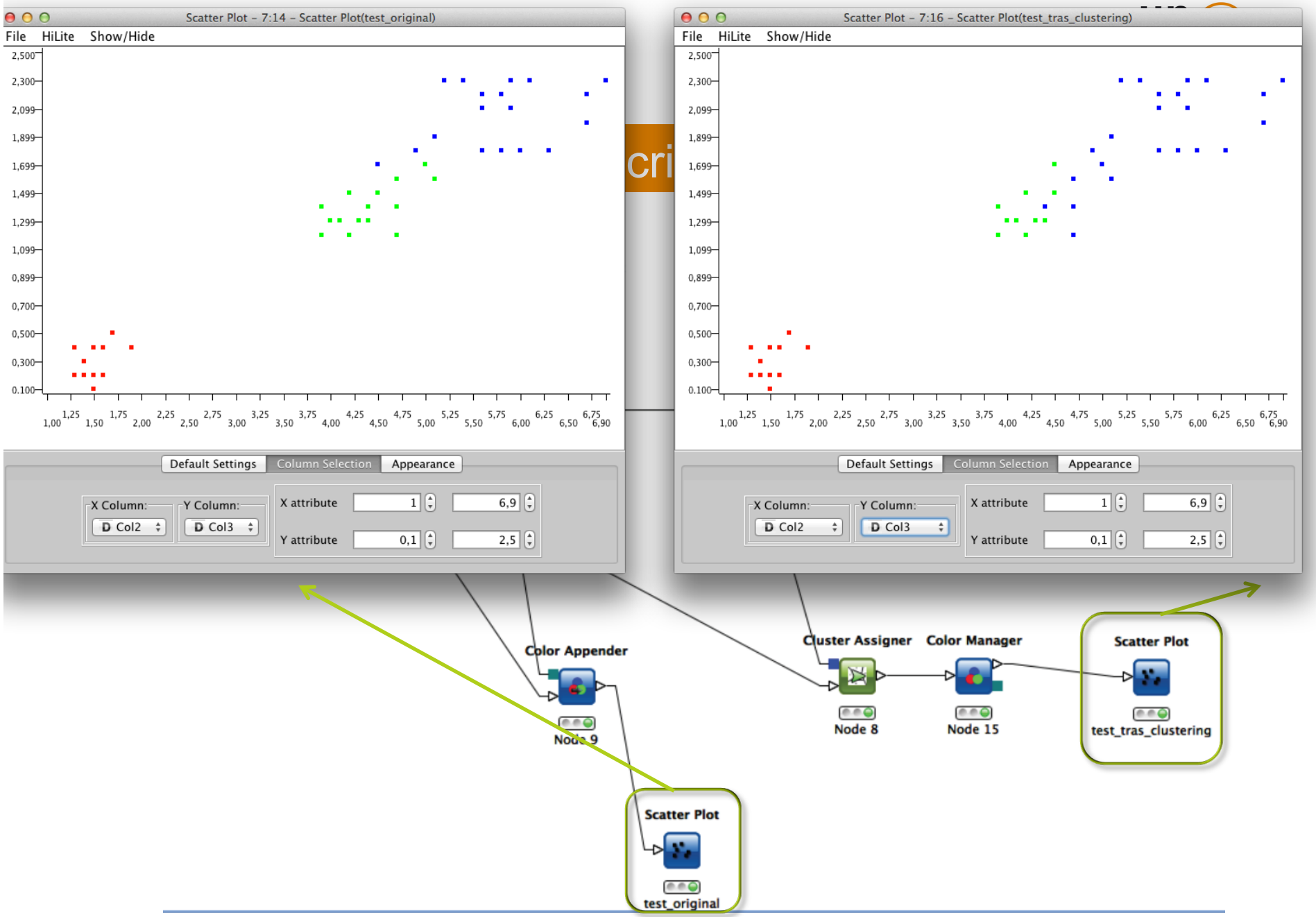


tivo ► k means


Scatter Plot – 6:14 – Scatter Plot
File HiLite Show/Hide
Default Settings
Column Selection
Appearance
X Column:
Col2
Y Column:
Col3
X attribute
1
6,9
Y attribute
0,1
2,5
Confusion Matrix – 6:10 – Scorer
File Hilite
Col4 \ Clu... Iris-setosa Iris-versic... Iris-virginica
Iris-setosa 20 0 0
Iris-versic... 0 19 1
Iris-virginica 0 7 13
Correct classified: 52
Wrong classified: 8
Accuracy: 86,667 %
Error: 13,333 %
Scatter Plot
Node 14
Node 8
String Replacer
Node 11
String Replacer
Node 12
String Replacer
Node 13
Scorer
Node 10

Ejemplo 1: Iris ► Análisis descriptivo ► k means




Scatter Plot – 7:14 – Scatter Plot(test_original)
File HiLite Show/Hide
Default Settings Column Selection Appearance
X Column: Col2
Y Column: Col3
X attribute 1 6,9
Y attribute 0,1 2,5
Scatter Plot – 7:16 – Scatter Plot(test_tras_clustering)
File HiLite Show/Hide
Default Settings Column Selection Appearance
X Column: Col2
Y Column: Col3
X attribute 1 6,9
Y attribute 0,1 2,5
Color Appender
Node 9
Cluster Assigner
Node 8
Color Manager
Node 15
Scatter Plot
test_tras_clustering
Scatter Plot
test_original

Ejemplo 2: Pima

- Carga de datos
- Visualización
- Análisis predictivo
- Análisis descriptivo

https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

- 8 variables predictoras (continuas)
- 768 ejemplos
- 2 clases (500/268)

- ¿Valores perdidos?

Ejemplo 3: Wine

- Carga de datos
- Visualización
- Análisis predictivo
- Análisis descriptivo

https://archive.ics.uci.edu/ml/datasets/Wine

- 12 variables predictoras (continuas)
- 178 ejemplos
- 3 clases (59/71/48)

Ejemplo 4: Spam

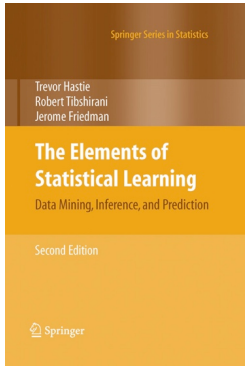
- Carga de datos
- Visualización
- Preprocesamiento
- Análisis predictivo
- Análisis descriptivo

https://archive.ics.uci.edu/ml/datasets/Spambase

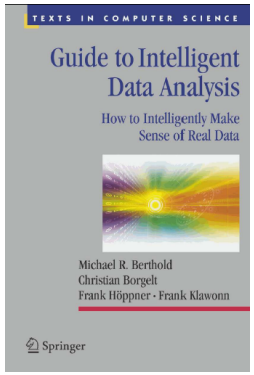
- 57 variables predictoras (55 continuas, 2 enteras)
- 4601 ejemplos
- 2 clases (2788/1813)

UCI Machine Learning Repository

Bibliografía



- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, by Trevor Hastie, Robert Tibshirani and Jerome Friedman. 2009



- Guide to Intelligent Data Analysis. How to Intelligently Make Sense of Real Data, by M. R. Berthold, C. Borgelt, F. Höppner, F. Klawonn. Springer 2010

Más información

- KNIME pages (www.knime.org)
- KNIME Tech (tech.knime.org)
  - KNIME Quickstart Guide
    https://tech.knime.org/files/KNIME_quickstart.pdf
- KNIME TV Channel on YouTube
- 100 best KNIME Videos
  http://meta-guide.com/videography/100-best-knime-videos

CURSOS DE VERANO 2014

APROXIMACIÓN PRÁCTICA A LA CIENCIA DE DATOS Y BIG DATA: HERRAMIENTAS KNIME, R, HADOOP Y MAHOUT

Introducción a KNIME

María José del Jesus