



SCIENZA DEI DATI

Roberto Pellungrini
Università di Pisa

KDD LAB – Knowledge Discovery and Data Mining Lab.
<http://kdd.isti.cnr.it>



Cosa vedremo oggi

- Introduzione a KNIME
- Principali componenti
- Aprire ed esplorare i dati
- Clusterizzazione
- Classificazione

Cosa è KNIME?

- KNIME = Konstanz Information Miner
- Sviluppato inizialmente presso l'università di Konstanz in Germania
- Versione completamente gratuita per desktop
- Piattaforma modulare per la data science basata su **workflows a nodi**.
- Mette a disposizione funzioni standard per **data mining, analisi e manipolazione dei dati**
- Si possono installare varie estensioni per integrare nuove funzionalità

Risorse per KNIME

- Pagina web principale e documentazione

<https://www.knime.com>

- Downloads

<https://www.knime.com/downloads/download-knime>

- Apprendimento

<https://www.knime.com/learning>

Download the latest KNIME Analytics Platform for Windows, Linux, and macOS: **4.4.2**. This version is intended for end users and provides everything needed to immediately begin using KNIME as well as extend KNIME with extension packages developed by others.

Windows

KNIME Analytics Platform for Windows (installer)

The installer adds an icon to the desktop and suggests suitable memory settings

Download (548 MB)

KNIME Analytics Platform for Windows (self-extracting archive)

The self-extracting archive only creates a folder holding the KNIME installation

Download (553 MB)

KNIME Analytics Platform for Windows (zip archive)

Download (677 MB)



Linux

KNIME Analytics Platform for Linux

Download (712 MB)

Mac

KNIME Analytics Platform for macOS (10.13 and above)

Download (556 MB)

Find out what's new in the latest KNIME 4.4 release [here](#).

KNIME Explorer

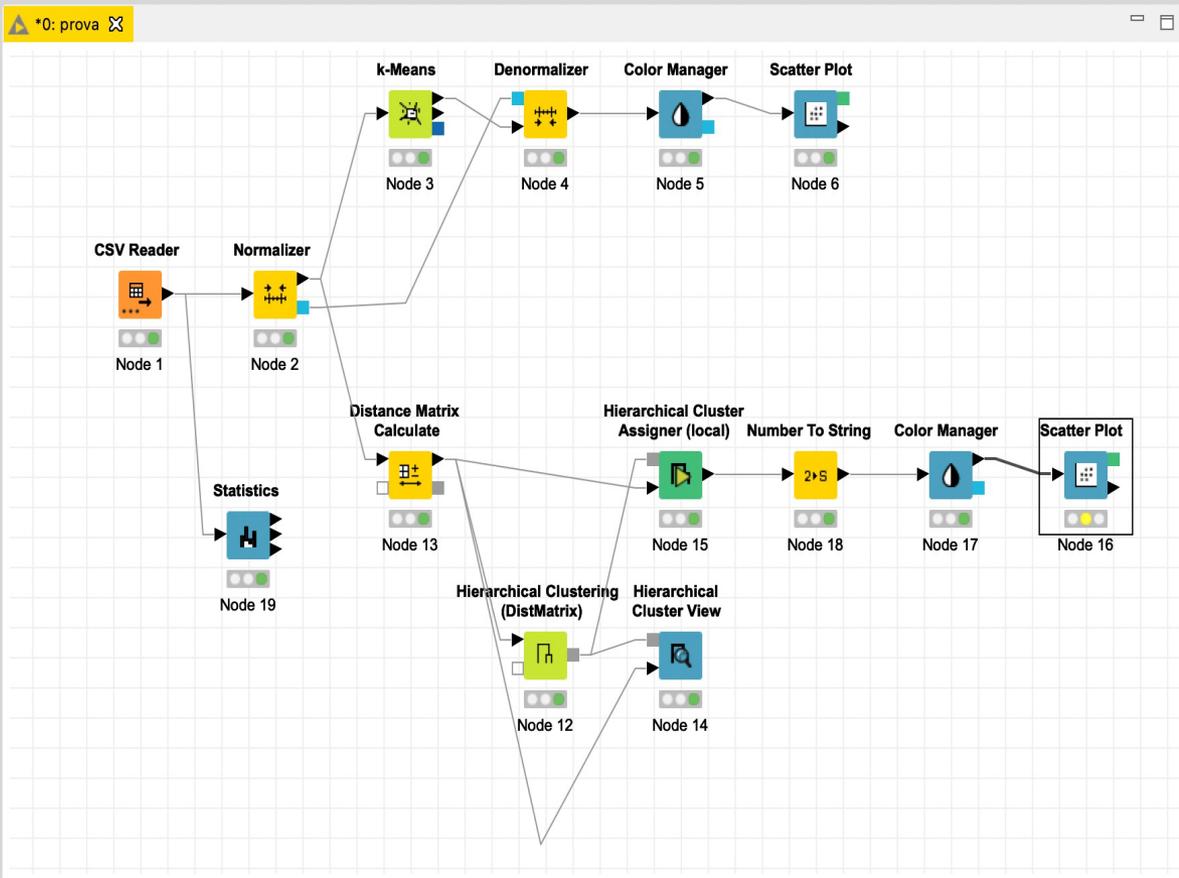
- My-KNIME-Hub (api.hub.knime.com)
- EXAMPLES (knime@api.hub.knime.com)
- LOCAL (Local Workspace)
 - Example Workflows
 - knime_data
 - 01_clusters
 - 02_knn
 - 03_DecisionTree
 - prova
 - clusters.sva

Workflow Coach

[Node recommendations only available with usage data req](#)

Node Repository

- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data
- Scripting
- Tools & Services
- KNIME Labs
- Workflow Control
- Workflow Abstraction
- Reporting



Scatter Plot

A scatter plot using a JavaScript based charting library. The view can be accessed either via the "interactive view" action on the executed node or in KNIME Server web portal page.

The configuration of the node lets you choose the size of a sample to display and to enable certain controls, which are then available in the view. This includes the ability to choose different columns for x and y or the possibility to set a title. Enabling or disabling these controls via the configuration dialog might not seem useful at first glance but has benefits when used in a web portal/wizard execution where the end user has no access to the workflow itself.

Since missing values as well as NaN (not a number) or infinite values cannot be displayed in the view, they will be omitted with a

Outline

Console Node Monitor

Node: Scatter Plot (0:16)

State: CONFIGURED

Port Output: Port 0

Node not executed

Barra dei comandi, bottoni di esecuzione e layout

The screenshot displays the KNIME Analytics Platform interface. At the top, the title bar reads "KNIME Analytics Platform - /Users/roberto/Documents/knime_workspace". Below the title bar is a toolbar with various icons for file operations, execution, and navigation. The main workspace is a grid where a workflow is built using nodes. The workflow starts with a "CSV Reader" (Node 1) and a "Statistics" (Node 19) node. The "CSV Reader" feeds into a "Normalizer" (Node 2). The "Normalizer" feeds into two parallel paths: one leading to a "k-Means" (Node 3) node, and another leading to a "Distance Matrix Calculate" (Node 13) node. The "k-Means" node feeds into a "Denormalizer" (Node 4), which then feeds into a "Color Manager" (Node 5), and finally a "Scatter Plot" (Node 6). The "Distance Matrix Calculate" node feeds into a "Hierarchical Clustering (DistMatrix)" (Node 12) node, which feeds into a "Hierarchical Cluster Assigner (local)" (Node 15). The "Hierarchical Cluster Assigner" feeds into a "Number To String" (Node 18) node, which feeds into a "Color Manager" (Node 17), and finally a "Scatter Plot" (Node 16). The "Statistics" node feeds into the "Distance Matrix Calculate" node. The "Hierarchical Clustering (DistMatrix)" node feeds into the "Hierarchical Cluster View" (Node 14). The "Hierarchical Cluster View" node feeds into the "Number To String" node. The "Number To String" node feeds into the "Color Manager" node. The "Color Manager" node feeds into the "Scatter Plot" node. The "Scatter Plot" node is highlighted with a red box. On the right side of the interface, there is a "Scatter Plot" configuration panel with the following text: "A scatter plot using a JavaScript based charting library. The view can be accessed either via the 'interactive view' action on the executed node or in KNIME Server web portal page. The configuration of the node lets you choose the size of a sample to display and to enable certain controls, which are then available in the view. This includes the ability to choose different columns for x and y or the possibility to set a title. Enabling or disabling these controls via the configuration dialog might not seem useful at first glance but has benefits when used in a web portal/wizard execution where the end user has no access to the workflow itself. Since missing values as well as NaN (not a number) or infinite values cannot be displayed in the view, they will be omitted with a". At the bottom of the interface, there is an "Outline" panel on the left and a "Console" and "Node Monitor" panel on the right. The "Node Monitor" panel shows the selected "Scatter Plot (0:16)" node with the state "CONFIGURED" and "Port Output" set to "Port 0". The "Node Monitor" panel also shows "Node not executed".

Barra laterale del workspace e selezione dei nodi

The image displays the KNIME Analytics Platform interface. The main workspace shows a workflow with the following nodes: CSV Reader (Node 1), Normalizer (Node 2), Statistics (Node 19), k-Means (Node 3), Denormalizer (Node 4), Color Manager (Node 5), Scatter Plot (Node 6), Distance Matrix Calculate (Node 13), Hierarchical Cluster Assigner (local) (Node 15), Number To String (Node 18), Color Manager (Node 17), Scatter Plot (Node 16), Hierarchical Clustering (DistMatrix) (Node 12), and Hierarchical Cluster View (Node 14).

The left sidebar contains the following sections:

- KNIME Explorer**: Shows the local workspace structure, including 'My-KNIME-Hub', 'EXAMPLES', and 'LOCAL (Local Workspace)' with subfolders like 'Example Workflows', 'knime_data', and 'prova'.
- Workflow Coach**: Provides node recommendations, with a note that they are only available with usage data recording.
- Node Repository**: A searchable list of nodes categorized by function, such as IO, Manipulation, Views, Analytics, DB, and Scripting.

The right sidebar shows the configuration for the selected 'Scatter Plot' node. It includes a description: 'A scatter plot using a JavaScript based charting library. The view can be accessed either via the "interactive view" action on the executed node or in KNIME Server web portal page.' It also mentions that the configuration allows for choosing the size of a sample to display and enabling certain controls. The configuration panel shows 'Port Output' set to 'Port 0' and a 'Load data' button. The status is 'CONFIGURED' and the node has not been executed.

Workspace principale

The screenshot displays the KNIME Analytics Platform workspace. The main area shows a workflow with the following nodes and connections:

- Node 1:** CSV Reader
- Node 2:** Normalizer (receives input from Node 1)
- Node 3:** k-Means (receives input from Node 2)
- Node 4:** Denormalizer (receives input from Node 2)
- Node 5:** Color Manager (receives input from Node 3)
- Node 6:** Scatter Plot (receives input from Node 4)
- Node 13:** Statistics (receives input from Node 1)
- Node 19:** Statistics (receives input from Node 1)
- Node 12:** Hierarchical Clustering (DistMatrix) (receives input from Node 13)
- Node 15:** Hierarchical Cluster Assigner (local) (receives input from Node 12)
- Node 18:** Number To String (receives input from Node 15)
- Node 17:** Color Manager (receives input from Node 18)
- Node 16:** Scatter Plot (receives input from Node 17)

The left sidebar contains the **Node Repository** with categories like IO, Manipulation, Views, Analytics, DB, and Scripting. The bottom panel shows the **Node Monitor** for the selected **Scatter Plot (0:16)** node, which is in a **CONFIGURED** state. The console output shows "Node not executed".

Scatter Plot

A scatter plot using a JavaScript based charting library. The view can be accessed either via the "interactive view" action on the executed node or in KNIME Server web portal page.

The configuration of the node lets you choose the size of a sample to display and to enable certain controls, which are then available in the view. This includes the ability to choose different columns for x and y or the possibility to set a title. Enabling or disabling these controls via the configuration dialog might not seem useful at first glance but has benefits when used in a web portal/wizard execution where the end user has no access to the workflow itself.

Since missing values as well as NaN (not a number) or infinite values cannot be displayed in the view, they will be omitted with a

Descrizione dei nodi e del workspace

The image displays the KNIME Analytics Platform interface. The main workspace shows a workflow with the following nodes:

- CSV Reader (Node 1)
- Normalizer (Node 2)
- Statistics (Node 19)
- Distance Matrix Calculate (Node 13)
- k-Means (Node 3)
- Denormalizer (Node 4)
- Color Manager (Node 5)
- Scatter Plot (Node 6)
- Hierarchical Clustering (DistMatrix) (Node 12)
- Hierarchical Cluster Assigner (local) (Node 15)
- Number To String (Node 18)
- Color Manager (Node 17)
- Scatter Plot (Node 16)

The right-hand side of the interface shows the configuration panel for the selected "Scatter Plot" node (Node 16). The panel includes a title "Scatter Plot" and a description: "A scatter plot using a JavaScript based charting library. The view can be accessed either via the 'interactive view' action on the executed node or in KNIME Server web portal page." Below the description, it states: "The configuration of the node lets you choose the size of a sample to display and to enable certain controls, which are then available in the view. This includes the ability to choose different columns for x and y or the possibility to set a title. Enabling or disabling these controls via the configuration dialog might not seem useful at first glance but has benefits when used in a web portal/wizard execution where the end user has no access to the workflow itself." At the bottom, it notes: "Since missing values as well as NaN (not a number) or infinite values cannot be displayed in the view, they will be omitted with a".

The bottom of the interface shows the "Node Monitor" panel for the selected node, displaying the following information:

- Node: Scatter Plot (0:16)
- State: CONFIGURED
- Port Output: Port 0
- Load data button
- Node not executed

Outline e console

The image displays the KNIME Analytics Platform interface. The main workspace shows a workflow with the following nodes: CSV Reader (Node 1), Normalizer (Node 2), k-Means (Node 3), Denormalizer (Node 4), Color Manager (Node 5), Scatter Plot (Node 6), Statistics (Node 19), Distance Matrix Calculate (Node 13), Hierarchical Cluster Assigner (local) (Node 15), Number To String (Node 18), Color Manager (Node 17), Scatter Plot (Node 16), Hierarchical Clustering (DistMatrix) (Node 12), and Hierarchical Cluster View (Node 14). The interface includes a KNIME Explorer on the left, a Workflow Coach, and a Node Repository. The Outline view at the bottom left shows a tree structure of the workflow nodes. The Console view at the bottom right shows the configuration for the selected Scatter Plot (0:16) node, which is in a CONFIGURED state. The console output is empty, indicating the node has not been executed.

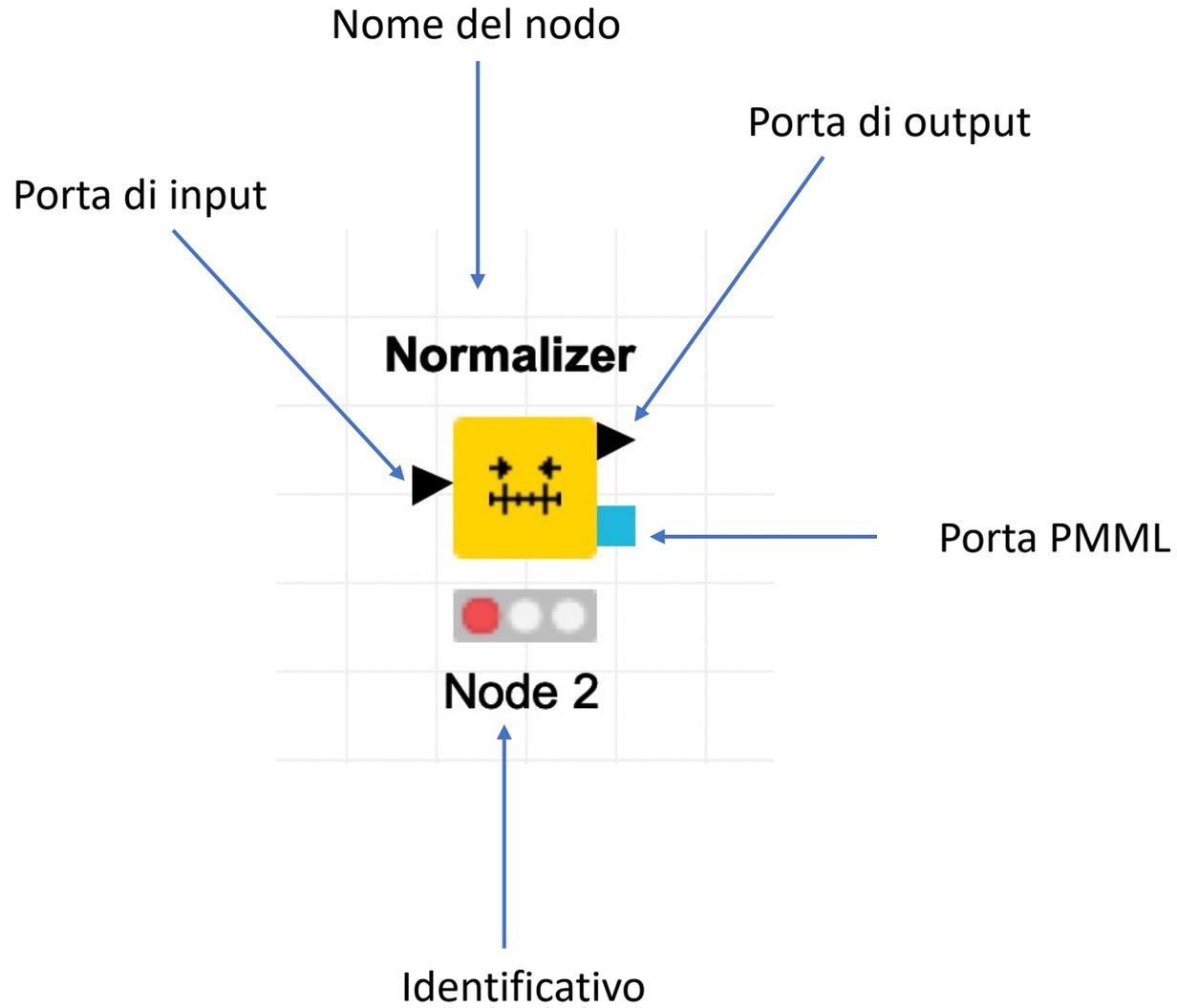
Scatter Plot

A scatter plot using a JavaScript based charting library. The view can be accessed either via the "interactive view" action on the executed node or in KNIME Server web portal page.

The configuration of the node lets you choose the size of a sample to display and to enable certain controls, which are then available in the view. This includes the ability to choose different columns for x and y or the possibility to set a title. Enabling or disabling these controls via the configuration dialog might not seem useful at first glance but has benefits when used in a web portal/wizard execution where the end user has no access to the workflow itself.

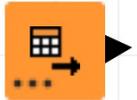
Since missing values as well as NaN (not a number) or infinite values cannot be displayed in the view, they will be omitted with a

Struttura di un nodo



Stato del nodo

CSV Reader



Node 1

Da configurare

CSV Reader



Node 1

Pronto all'esecuzione

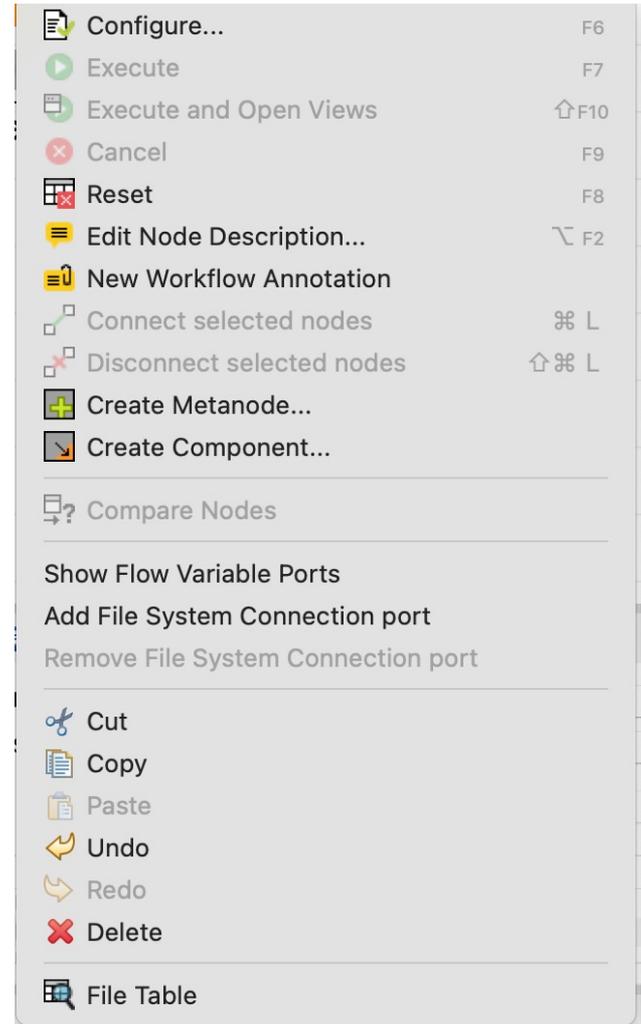
CSV Reader



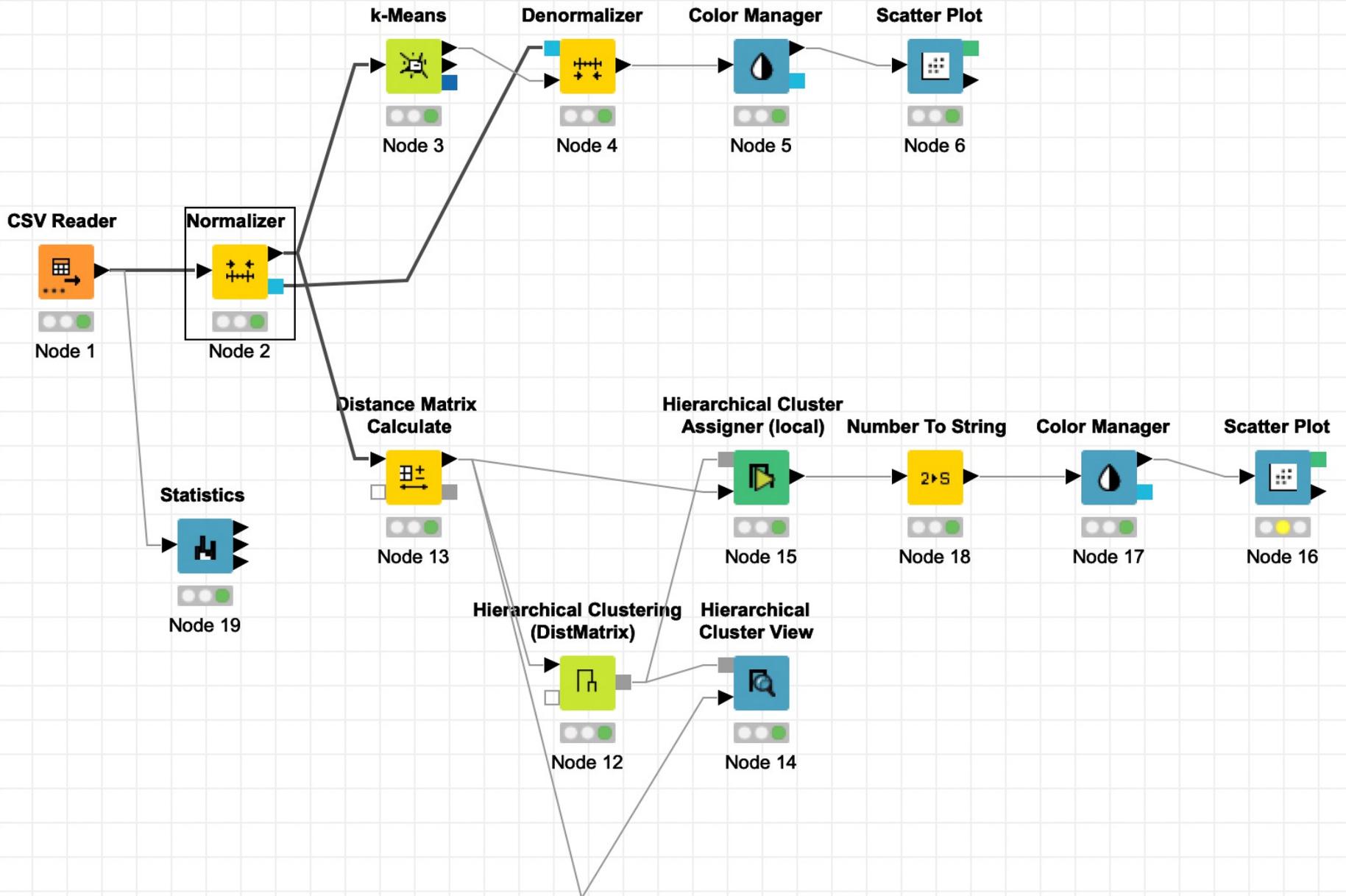
Node 1

Eseguito

Menù contestuale:
tasto destro del mouse



Un workflow di esempio



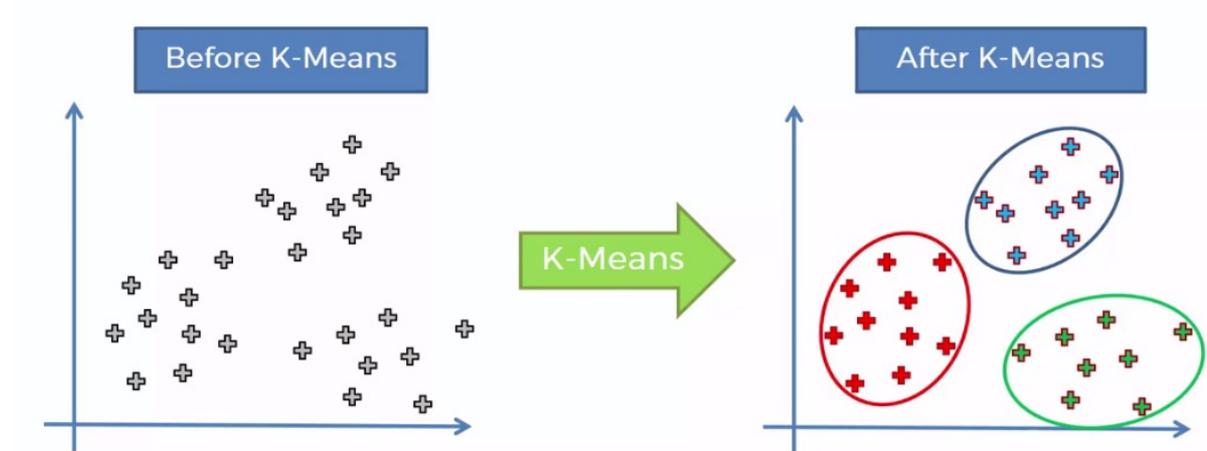
Normalizer

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Normalizer

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

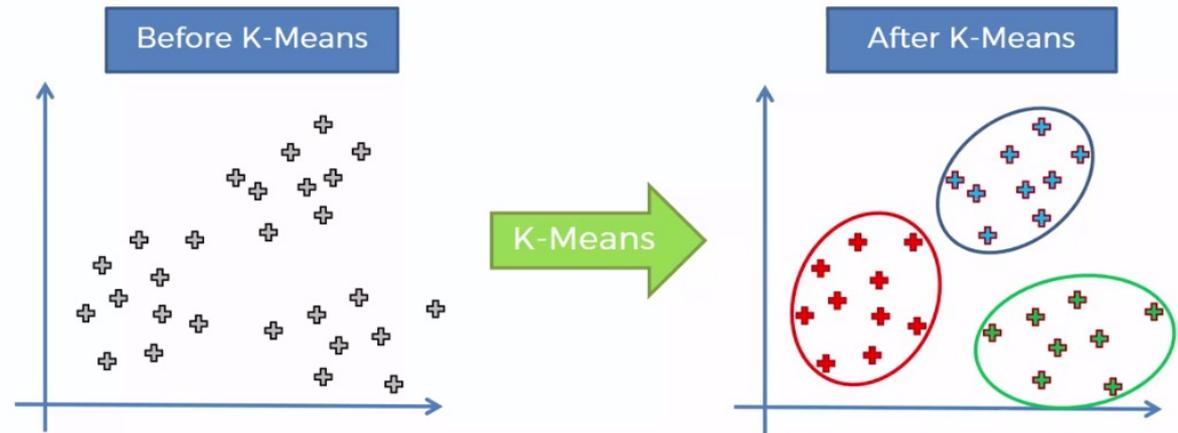
K-means



Normalizer

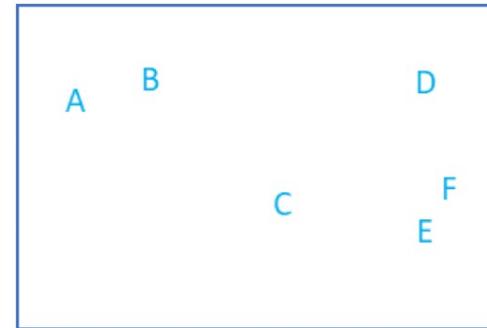
$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

K-means

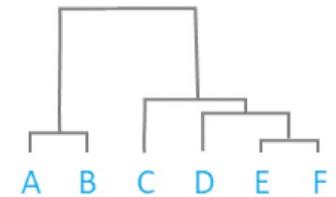


Cluster Gerarchico

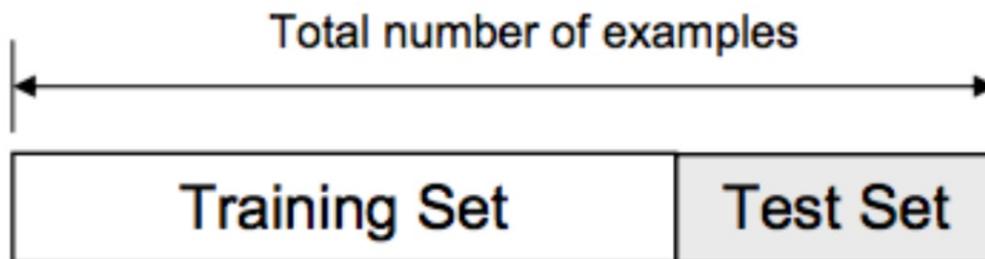
	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0



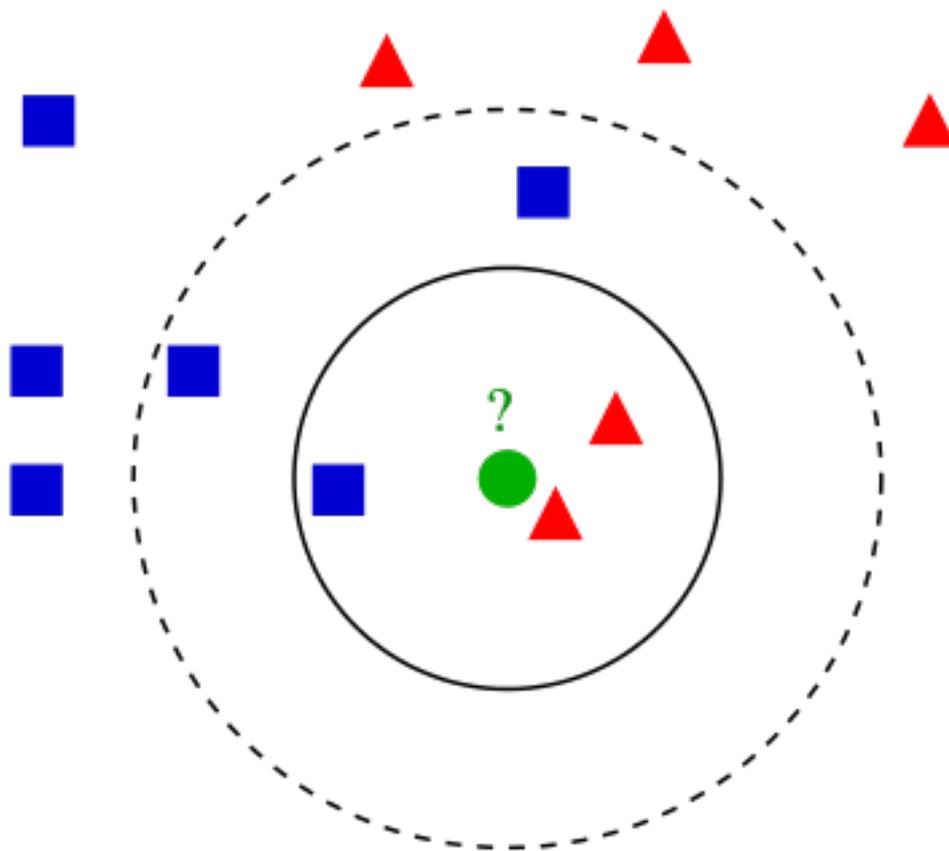
Dendrogram



Train e test

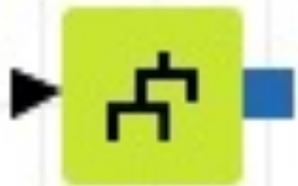


Knn



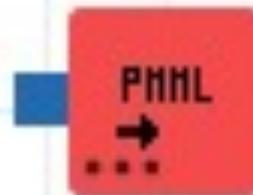
Alcuni classificatori ci restituiscono il modello allenato come un dato PMML (notare la porta in output) Possiamo poi salvarlo per riutilizzarlo più avanti.

**Decision
Tree Learner**



Node 15

PMML Writer



Node 16

PMML Reader



Node 17