



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

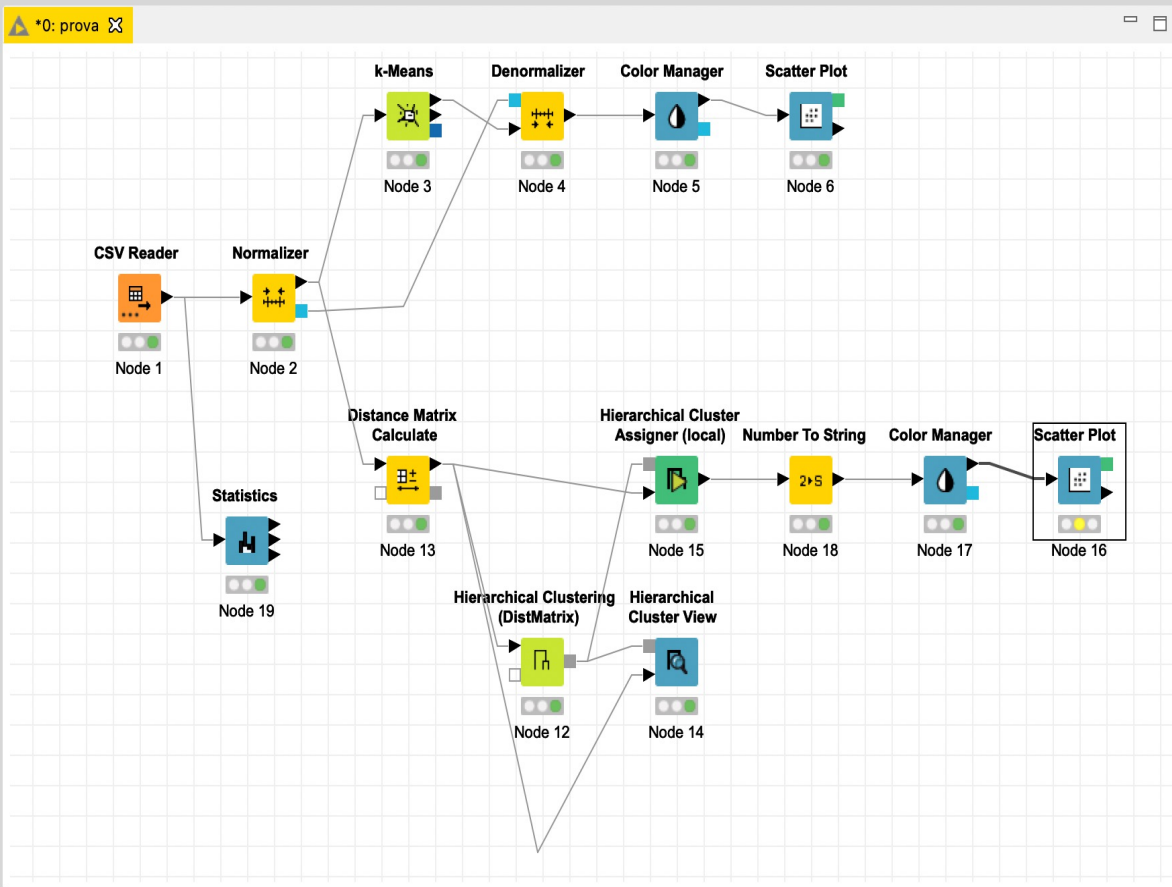
- My-KNIME-Hub (api.hub.knime.com)
- EXAMPLES (knime@api.hub.knime.com)
- LOCAL (Local Workspace)
 - Example Workflows
 - knime_data
 - 01_clusters
 - 02_knn
 - 03_DecisionTree
 - prova
 - clusters.sva

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) connected to a "Normalizer" (Node 2). From Node 2, the flow branches into two paths: one leading to "k-Means" (Node 3) and "Denormalizer" (Node 4), and another leading to "Distance Matrix Calculate" (Node 13). Node 13 connects to "Hierarchical Clustering (DistMatrix)" (Node 12), which then connects to "Hierarchical Cluster View" (Node 14). Another path from Node 2 goes to "Statistics" (Node 19). The "k-Means" path continues to "Color Manager" (Node 5) and "Scatter Plot" (Node 6). The "Hierarchical Clustering" path continues to "Hierarchical Cluster Assigner (local)" (Node 15), "Number To String" (Node 18), "Color Manager" (Node 17), and finally "Scatter Plot" (Node 16). On the left side, there are three panels: "KNIME Explorer" showing a file tree with "LOCAL (Local Workspace)" containing "Example Workflows", "knime_data", "01_clusters", "02_knn", "03_DecisionTree", "prova", and "clusters.sva"; "Workflow Coach" with a link to "Node recommendations only available with usage data rec"; and "Node Repository" with a search bar and a list of categories like IO, Manipulation, Views, Analytics, DB, etc. On the right side, a "Scatter Plot" node is selected, and a panel on the right shows its configuration and description. The description text reads: "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, there are panels for "Outline", "Console", and "Node Monitor". The "Node Monitor" panel shows "Node: Scatter Plot (0:16)", "State: CONFIGURED", "Port Output: Port 0", and "Load data" button. Below it, it says "Node not executed".

Barra laterale del workspace e selezione dei nodi

The image shows the KNIME Analytics Platform interface. The main workspace contains 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 is divided into three sections:

- KNIME Explorer:** Shows the local workspace structure, including folders like 'Example Workflows', 'knime_data', and files like '01_clusters', '02_knn', '03_DecisionTree', 'prova', and 'clusters.sva'.
- Workflow Coach:** Provides node recommendations, with a note: 'Node recommendations only available with usage data rec'.
- Node Repository:** A searchable list of nodes categorized by function, such as IO, Manipulation, Views, Analytics, DB, Other Data Types, Structured Data, Scripting, Tools & Services, KNIME Labs, Workflow Control, Workflow Abstraction, and Reporting.

The right sidebar displays the configuration for the selected 'Scatter Plot' node. It includes a title 'Scatter Plot', 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.', and a detailed configuration section: '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, the 'Console' and 'Node Monitor' panels are visible. The Node Monitor shows the selected node is 'Scatter Plot (0:16)' with a state of 'CONFIGURED'. The console output indicates 'Node not 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 right-hand panel shows the configuration for the selected **Scatter Plot** node (Node 16). The configuration includes:

- Node:** Scatter Plot (0:16)
- State:** CONFIGURED
- Port Output:** Port 0
- Load data:** [button]
- Node not executed:** [text area]

The Node Repository on the left lists various nodes under categories like IO, Manipulation, Views, Analytics, DB, Other Data Types, Structured Data, Scripting, Tools & Services, KNIME Labs, Workflow Control, Workflow Abstraction, and Reporting.

Descrizione dei nodi e del workspace

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

- CSV Reader (Node 1)** feeds into **Normalizer (Node 2)**.
- Normalizer (Node 2)** feeds into **k-Means (Node 3)** and **Distance Matrix Calculate (Node 13)**.
- k-Means (Node 3)** feeds into **Denormalizer (Node 4)**.
- Denormalizer (Node 4)** feeds into **Color Manager (Node 5)**.
- Color Manager (Node 5)** feeds into **Scatter Plot (Node 6)**.
- Distance Matrix Calculate (Node 13)** feeds into **Hierarchical Clustering (DistMatrix) (Node 12)** and **Hierarchical Cluster Assigner (local) (Node 15)**.
- Hierarchical Clustering (DistMatrix) (Node 12)** feeds into **Hierarchical Cluster View (Node 14)**.
- Hierarchical Cluster Assigner (local) (Node 15)** feeds into **Number To String (Node 18)**.
- Number To String (Node 18)** feeds into **Color Manager (Node 17)**.
- Color Manager (Node 17)** feeds into **Scatter Plot (Node 16)**.
- Statistics (Node 19)** is connected to the workflow but has no outgoing connections.

The interface includes a left sidebar with **KNIME Explorer** (showing a local workspace with folders like `knime_data` and files like `01_clusters`), **Workflow Coach**, and **Node Repository**. The bottom panel shows the **Outline**, **Console**, and **Node Monitor** for the selected **Scatter Plot (0:16)** node, which is in a **CONFIGURED** state and 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

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), 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 right sidebar shows the description for the selected 'Scatter Plot' node (Node 16):

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.

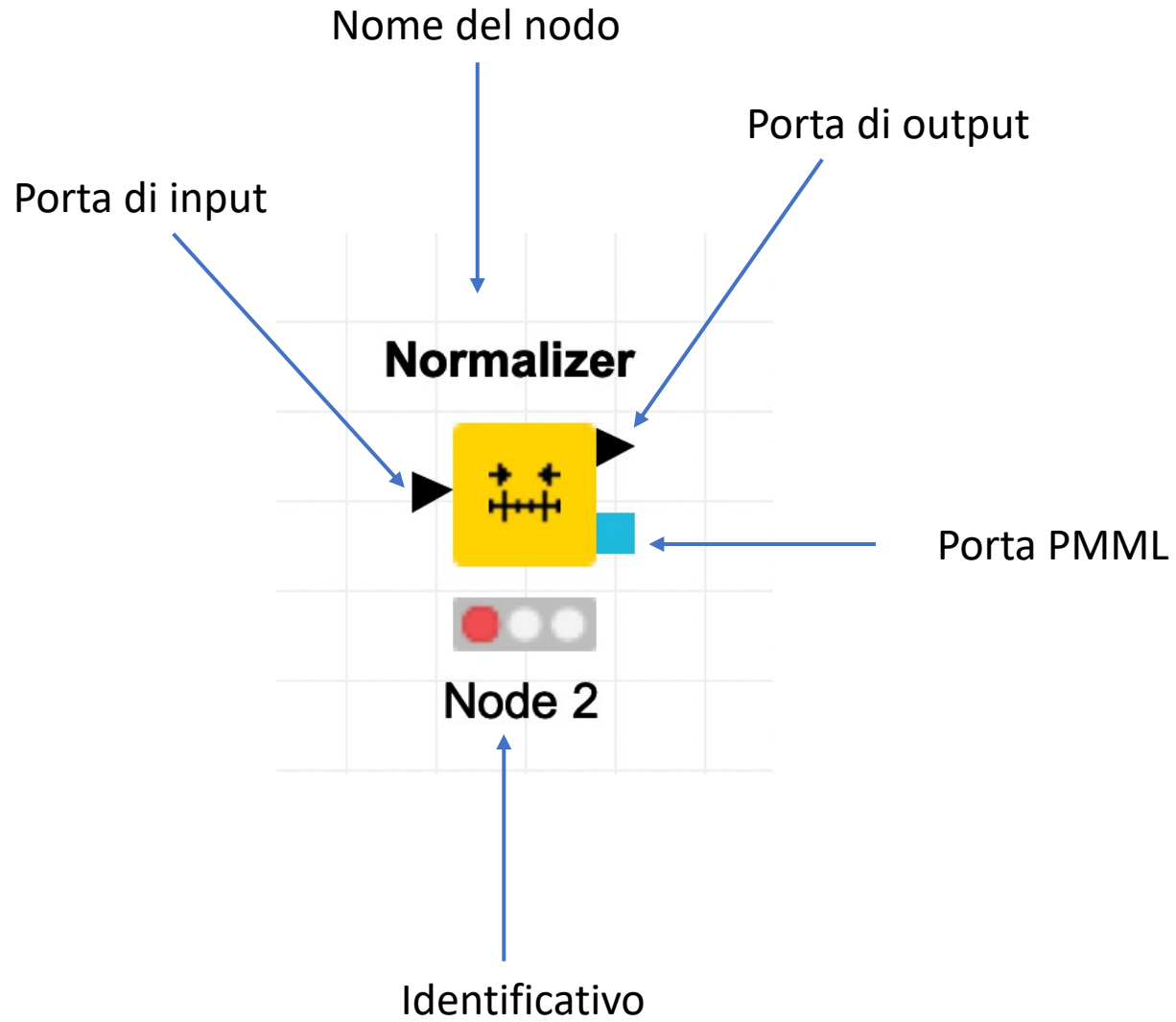
The configuration of the node lets you choose the size of a sample to display and to enable certain controls, which are then available in the view. This includes the ability to choose different columns for x and y or the possibility to set a title. Enabling or disabling these controls via the configuration 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

The bottom console shows the following information for the selected node:

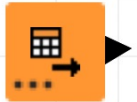
Node: Scatter Plot (0:16)
State: CONFIGURED
Port Output: Port 0 [dropdown] [Load data]
Node not executed

Struttura di un nodo



Stato del nodo

CSV Reader



Node 1

Da configurare

CSV Reader



Node 1

Pronto all'esecuzione

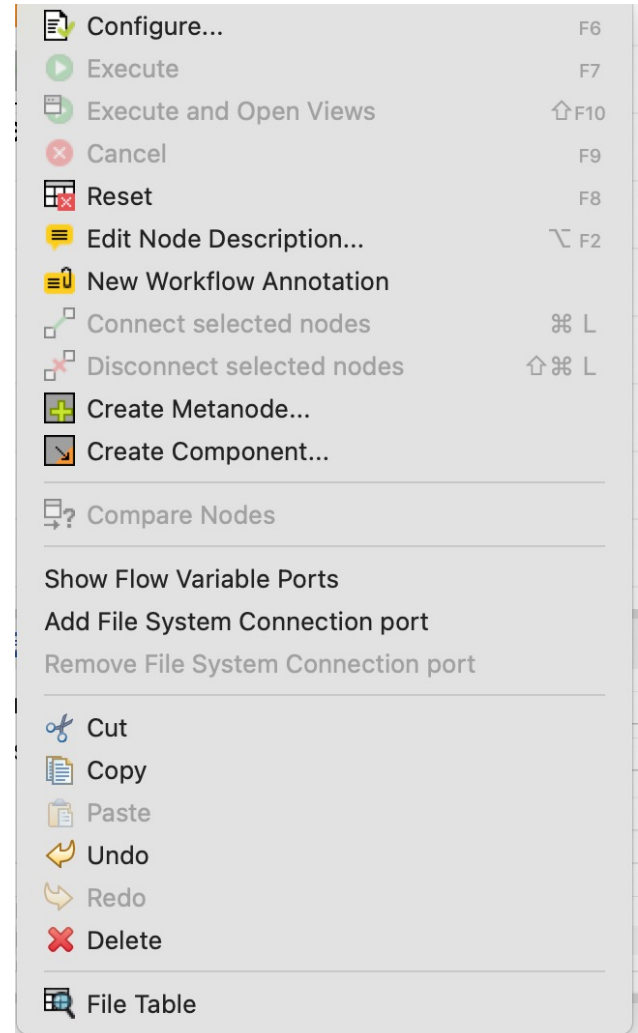
CSV Reader



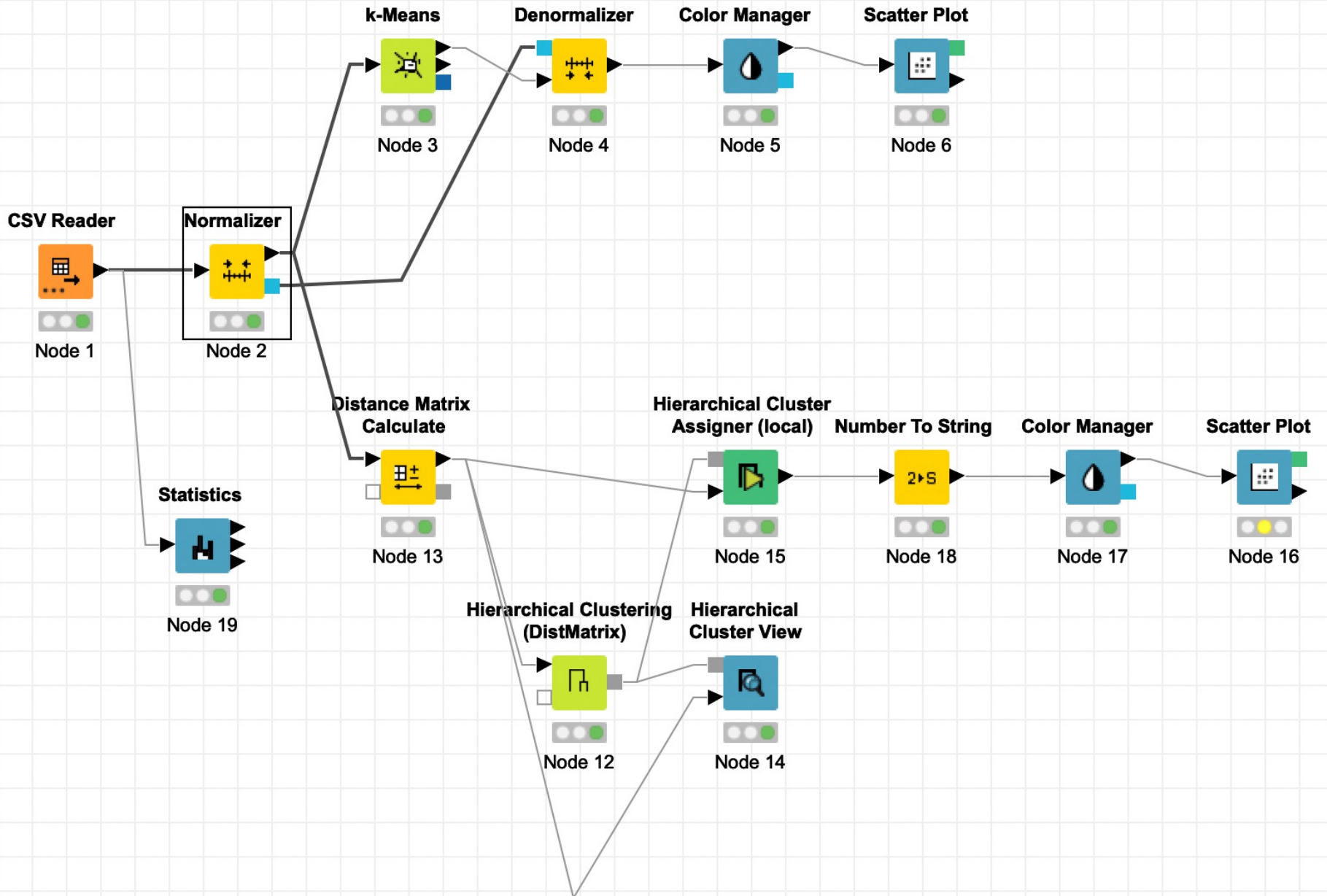
Node 1

Eseguito

Menù contestuale:
tasto destro del mouse



Un workflow di esempio



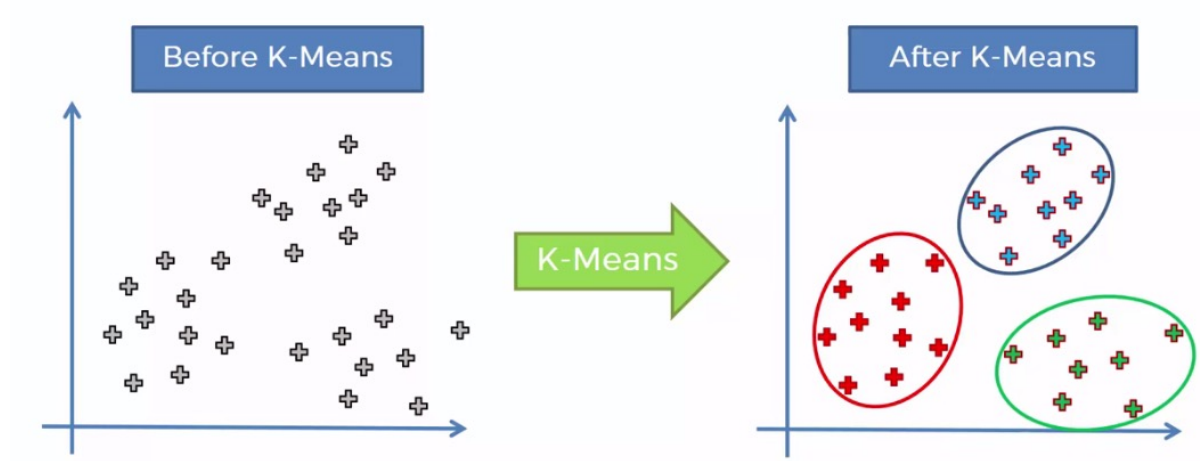
Normalizer

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

Normalizer

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

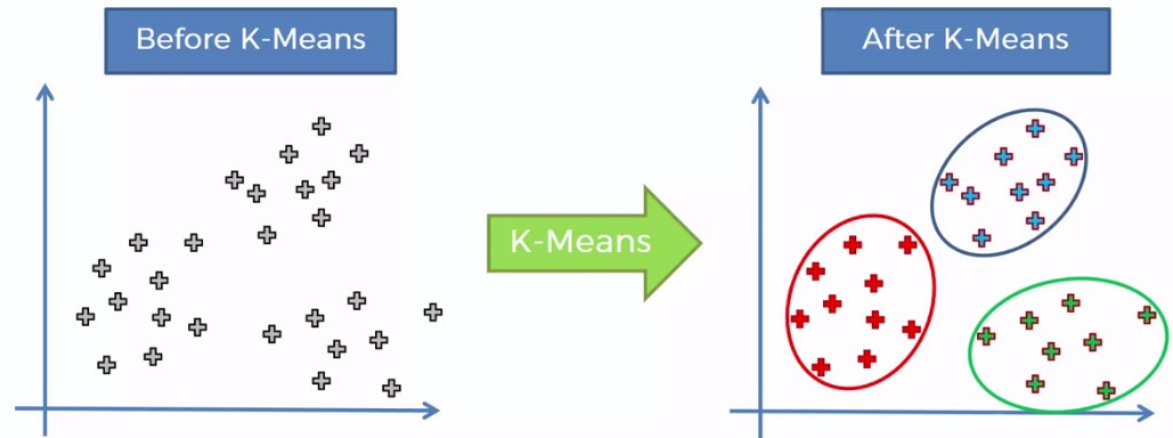
K-means



Normalizer

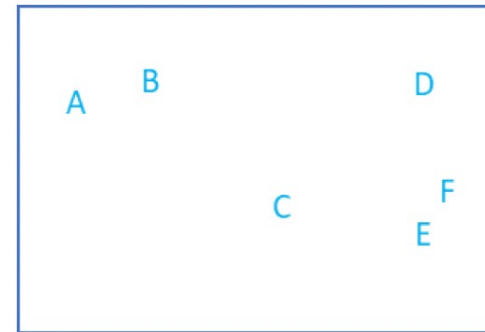
$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

K-means

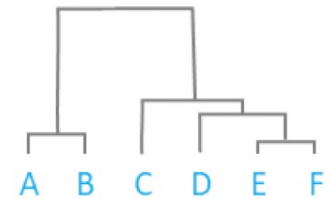


Cluster Gerarchico

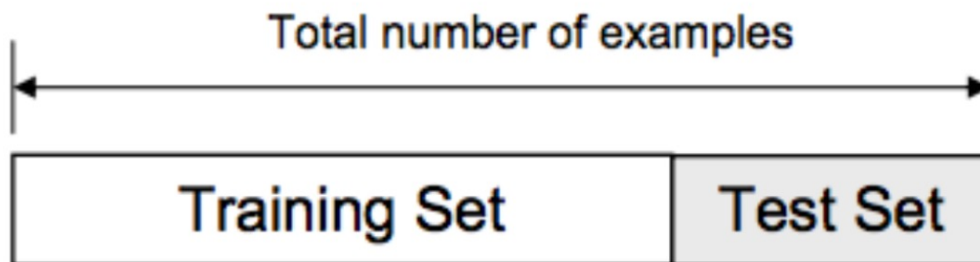
	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0



Dendrogram



Train e test



Knn

