



**DECSAI**

**Departamento de Ciencias de la Computación e I.A.**

Universidad de Granada

## **Sistemas Inteligentes de Gestión**

### **Guión de Prácticas de Minería de Datos**

#### **Práctica 1**

#### **Herramientas de Minería de Datos**



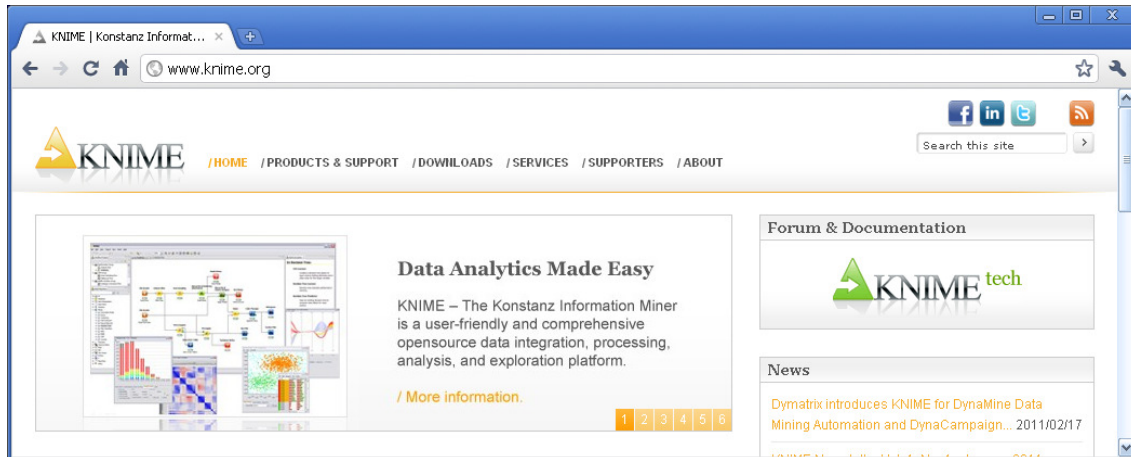
#### **Introducción a KNIME**

*© Juan Carlos Cubero & Fernando Berzal*

# Introducción a KNIME

KoNstanz Information MinEr

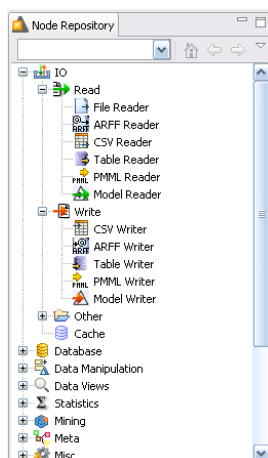
<http://www.knime.org/>



KNIME es un entorno totalmente gratuito para el desarrollo y ejecución de técnicas de minería de datos. KNIME fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold. En la actualidad, la empresa KNIME.com GmbH, radicada en Zúrich, Suiza, continúa su desarrollo, además de prestar servicios de formación y consultoría.

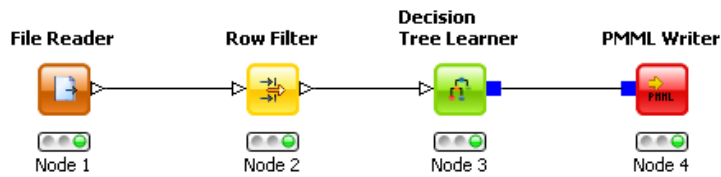
KNIME está desarrollado sobre la plataforma Eclipse y programado, esencialmente, en Java. Como otros entornos de este tipo, algunos de los cuales aparecen referenciados al final de este documento, su uso se basa en el diseño de un flujo de ejecución que plasme las distintas etapas de un proyecto de minería de datos.

Para ello, KNIME proporciona distintos nodos agrupados en fichas, como por ejemplo:



- a) **Entrada de datos** [*IO > Read*].
- b) **Salida de datos** [*IO > Write*].
- c) **Preprocesamiento** [*Data Manipulation*], para filtrar, discretizar, normalizar, filtrar, seleccionar variables...
- d) **Minería de datos** [*Mining*], para construir modelos (reglas de asociación, clustering, clasificación, MDS, PCA...).
- e) **Salida de resultados** [*Data Views*] para mostrar resultados en pantalla (ya sea de forma textual o gráfica).

Para crear un flujo de ejecución, las salidas de unos nodos se utilizan como entradas de otros. Por ejemplo, un flujo básico podría ser de la forma:

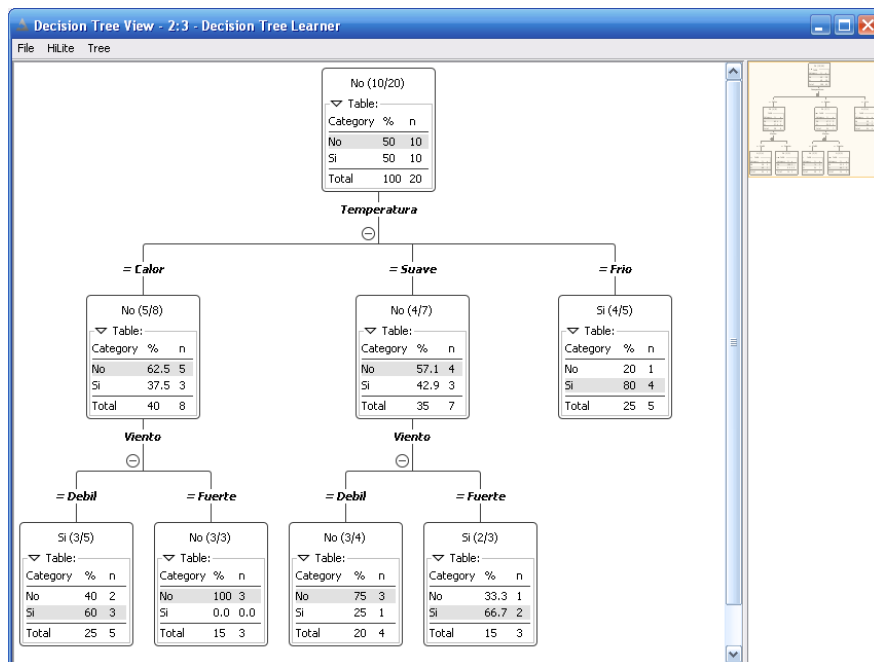


Nodo de lectura de datos

→ Nodo de preprocesamiento

→ Nodo de modelado (por ejemplo, modelo de clasificación)

→ Nodo de salida de resultados.



Lea ahora la introducción a KNIME que se encuentra en la siguiente URL:

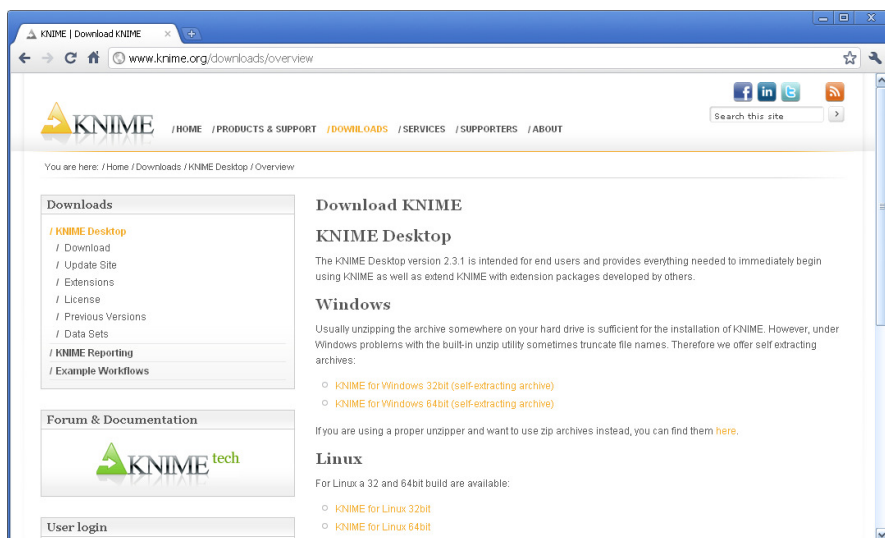
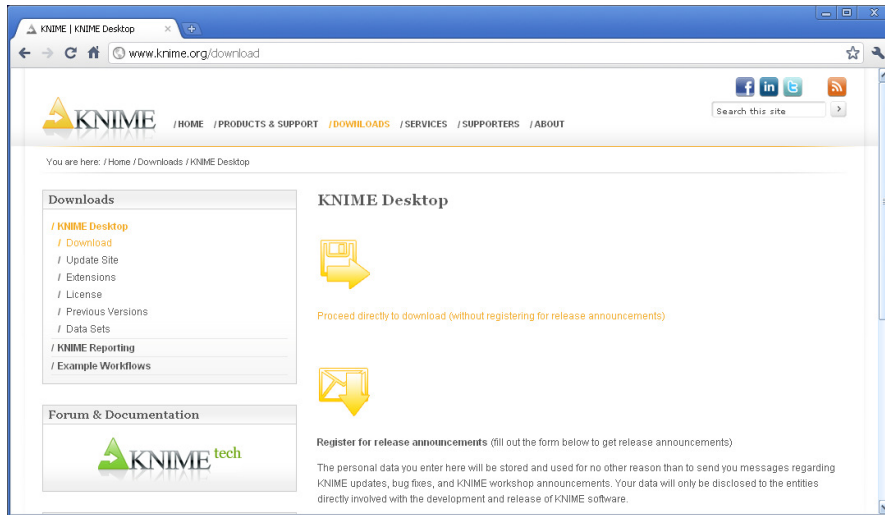
<http://tech.knime.org/getting-started>

De forma complementaria, también puede consultar la siguiente presentación de KNIME en castellano:

<http://www.exa.unicen.edu.ar/catedras/dmining/clases/PresentacionKNime.pdf>

## Instalación de KNIME

- Descargue la versión de KNIME adecuada para su sistema operativo (Windows o Linux, 32 ó 64 bits): <http://knime.org/download>

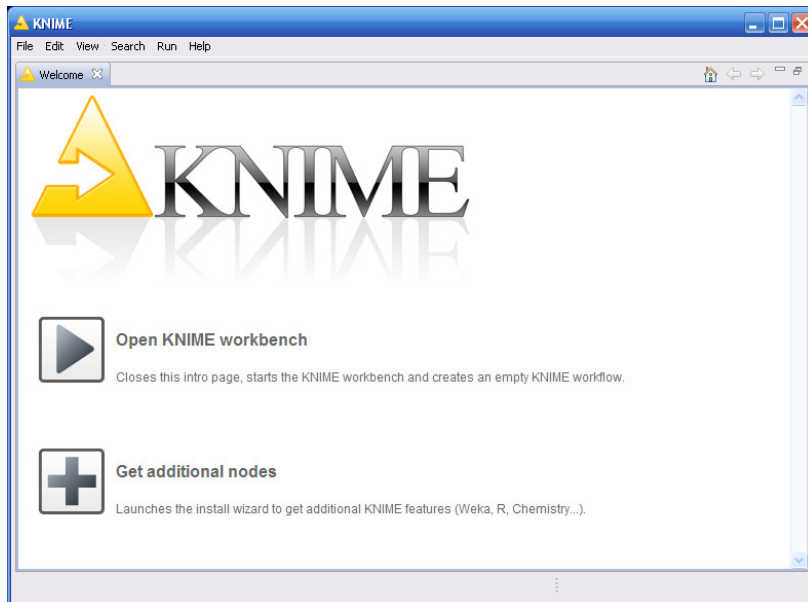


- Si utiliza Windows, ejecute directamente el archivo ZIP autodescomprimible que nos hayamos descargado del sitio oficial de KNIME (p.ej. `knime_2.3.1.win32.win32.x86.exe`) o descomprima manualmente el archivo en la carpeta en la que desee instalar el programa.

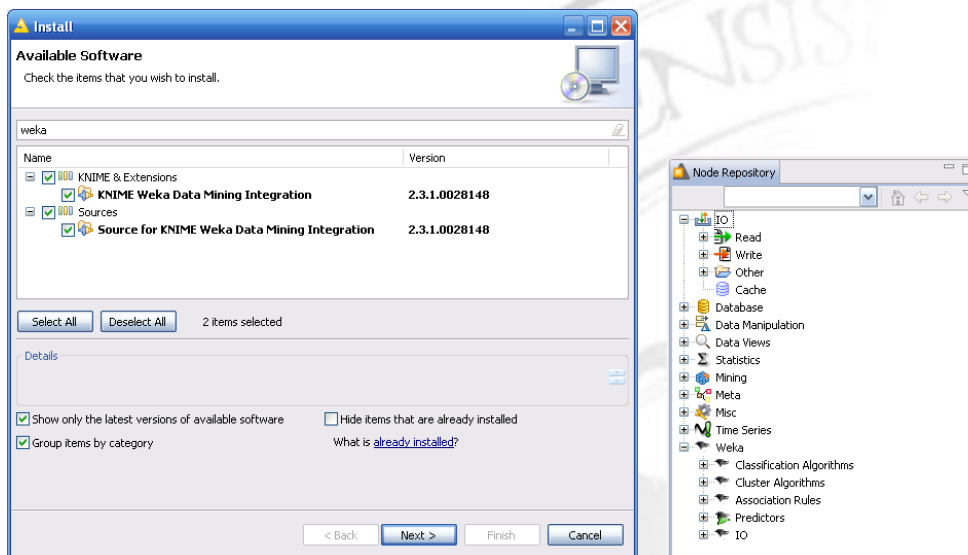
NOTA:

KNIME no requiere programa de instalación. El paquete se desinstala, simplemente, borrando la carpeta en la que lo hayamos descomprimido.

- Para ejecutar KNIME, busque el fichero `knime.exe` en la carpeta en la que haya descomprimido el paquete de instalación y ejecútelo:



- Antes de empezar a utilizar KNIME, nos aseguraremos de instalar los componentes de Weka, utilizando la opción “*Get additional nodes*” de la ventana de inicio de KNIME o accediendo a ellos a través del menú *File > Install KNIME Extensions*.



Información adicional en la guía de instalación paso a paso:

<http://tech.knime.org/installation-0>

## Otras herramientas de minería de datos

### *Weka*

Waikato Environment for Knowledge Analysis  
<http://www.cs.waikato.ac.nz/ml/weka/>

Similar a KNIME en cierto modo, Weka incluye distintas interfaces de usuario:

- *Knowledge Flow*, para crear flujos de ejecución similares a los de KNIME, aunque algo menos amigable en su versión actual.
- *Explorer*, para lanzar de forma separada la ejecución de distintas operaciones.
- *Experimenter*, para usuarios avanzados (ejecución sistemática de baterías de experimentos sobre conjuntos de datos).

Comparación frente a KNIME:

- ✚ Weka incorpora un mayor número de componentes.
- ✚ La interfaz de KNIME es más amigable.
- ✚ KNime permite usar los nodos de Weka.
- ✚ KNIME también permite otras extensiones, como las ofrecidas por R

### *RapidMiner*

(anteriormente conocido como YALE, *Yet Another Learning Environment*)  
<http://rapidminer.com/>

Otra herramienta similar a KNIME cuya versión inicial, conocida como YALE, fue desarrollada por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001. Actualmente se encarga de su desarrollo Rapid-I, con sede en Nürnberg, Alemania, y se mantiene una versión open-source.

### *R*

<http://www.r-project.org/>

R es un entorno estadístico tremendamente potente y completo, si bien no ofrece una interfaz de usuario amigable. Las llamadas a R se realizan en línea de comando y sus paquetes, por desgracia, no siempre se utilizan de la misma forma (al provenir de desarrolladores diferentes, lo que dificulta la realización de muchas tareas). Fue desarrollado inicialmente por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland, Nueva Zelanda, en 1993.

R también es un lenguaje de programación, implementación del lenguaje S ideado por John Chambers, Rick Becker y Allan Wilks de los Laboratorios Bell, del cual existe una versión comercial llamada S-Plus, de TIBCO Software Inc. (Palo Alto, CA, USA).

### *SPSS Modeler*

(anteriormente conocido como SPSS Clementine)

<http://www.spss.com/software/modeler/>

Derivado de software para Unix desarrollado para uso interno por una compañía británica llamada ISL (Integral Solutions Limited), la cual fue adquirida en 1999 por SPSS Inc..

SPSS Inc. es una empresa dedicada al desarrollo, distribución y venta del programa SPSS desde 1975. SPSS, que proviene de *Statistical Package for the Social Sciences*, fue creado en 1968 por Norman H. Nie, C. Hadlai (Tex) Hull y Dale H. Bent en la Universidad de Stanford (California). Entre 1969 y 1975, su desarrollo estuvo a cargo de la Universidad de Chicago (Illinois).

SPSS Modeler, bajo la denominación SPSS Clementine, fue la primera herramienta de minería de datos en utilizar una interfaz gráfica de usuario (2000). Posteriormente, fue rebautizado como PASW Modeler (2009), donde PASW hacía referencia a *Predictive Analytics Software*.

Recientemente, SPSS fue adquirida por IBM.

Una licencia de este tipo de herramientas comerciales cuesta miles de dólares...

### *SAS Enterprise Miner™*

(*Statistical Analysis System*)

<http://www.sas.com/>

SAS Enterprise Miner es la herramienta de minería de datos comercializada por SAS Institute, principal competidora de SPSS, y es sólo uno de los muchos componentes del sistema integrado SAS.

El paquete original SAS constaba de numerosos módulos y se ejecutaba inicialmente sobre mainframes de IBM.

SAS Institute tiene su sede central en Cary (Carolina del Norte, EE.UU.) y fue fundada en 1976 por Anthony Barr, James Goodnight, John Sall y Jane Helwig. A día de hoy, sigue siendo una compañía privada que frecuentemente se incluye en la lista de mejores empresas en las que trabajar (#1 en 2010 según la revista Fortune).



Información adicional sobre herramientas software de minería de datos en KDnuggets:

<http://www.kdnuggets.com/software/index.html>