DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Tratamiento Inteligente de Datos

Guión de Prácticas

Práctica 1

Herramientas de Minería de Datos

Introducción a KNIME

Introducción al SPSS

Introducción al Rstudio

Las herramientas de Minería de Datos

Desde que se popularizó el ordenador como herramienta de cálculo científico se desarrollaron herramientas de ayuda al cálculo estadístico, en especial al Análisis Multivariante. En principio estas herramientas aparecieron como librerías de subprogramas para ser llamadas desde un lenguaje de programación, habitualmente desde FORTRAN; pero muy aparecieron herramientas que permitían llamar a programas completos introduciendo los datos en un formato determinado y describiendo de forma muy sintética, cuál era formato de los datos y qué se deseaba que hiciese el programa. Estás herramientas permitían a los analistas despreocuparse de labores de programación y centrarse en el análisis de datos, si bien obviamente ofrecían procedimientos estándar. Obviamente muy pronto se hicieron interactivas

Con el desarrollo de la Estadística Computacional y de la Minería de Datos, las herramientas iniciales fueron evolucionando y surgieron otras nuevas de manera que actualmente existen una gran cantidad de opciones para utilizar. Hay

- Soluciones de software libre (Knime o Weka) y soluciones de software propietario (SPSS, o SAS)
- Soluciones más orientadas a la resolución de problemas estadísticos puros (SPSS o quizás R), y soluciones donde se hace más hincapié en las heurísticas de DM (Weka)
- Soluciones muy cerradas en procedimientos estándar (SPSS), soluciones que admiten módulos programados en Java (Weka, Knime), incluso soluciones que utilizan un lenguaje de programación propio y llamadas a librerías (R).
- Soluciones independientes del sistema de información, y soluciones asociadas a una base de datos (Data-Miner SQL Developer)

Algunas de las herramientas más conocidas:

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, existe un amplio grupo de usuarios que han desarrollado una gran cantidad de extensiones: textos, bioinformática, química etc. 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.

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

http://rapidminer.com/

Otra herramienta similar a KNIME cuya versión inicial, conocida como YALE [*Yet Another Learning Environment*], 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 sigue mantieniendo una versión open-source.

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 Labotatorios 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. (*En la aulas se encuentra instalado SPSS versión 20, no se puede instalar fuera de las aulas debido a las condiciones de las prácticas)*
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 y 2011, según la revista Fortune), con un 2% de turnover rate (empleados que cambian de trabajo al año) frente al 20% de muchos de sus competidores

ORACLE

Data Mining

Se trata de un módulo integrado en SQL-Developer que permite realizar minería de datos directamente sobre las bases de datos implementadas. Recoge los elementos más importantes de los problemas de DM pero no es muy completo. Su mayos ventaja es el procesamiento directo desde bases de datos. Las otras herramientas también permiten acceso a bases de datos no de forma directa y no a todo tipo de BD. La mejor en este aspecto es Knime.

En el aula de prácticas están instalados Knime, SPSS ,Weka y Rstudio. Para entrar en la imagen se utiliza el código *master-tdi*

Introducción a KNIME

http://www.knime.org/

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*].

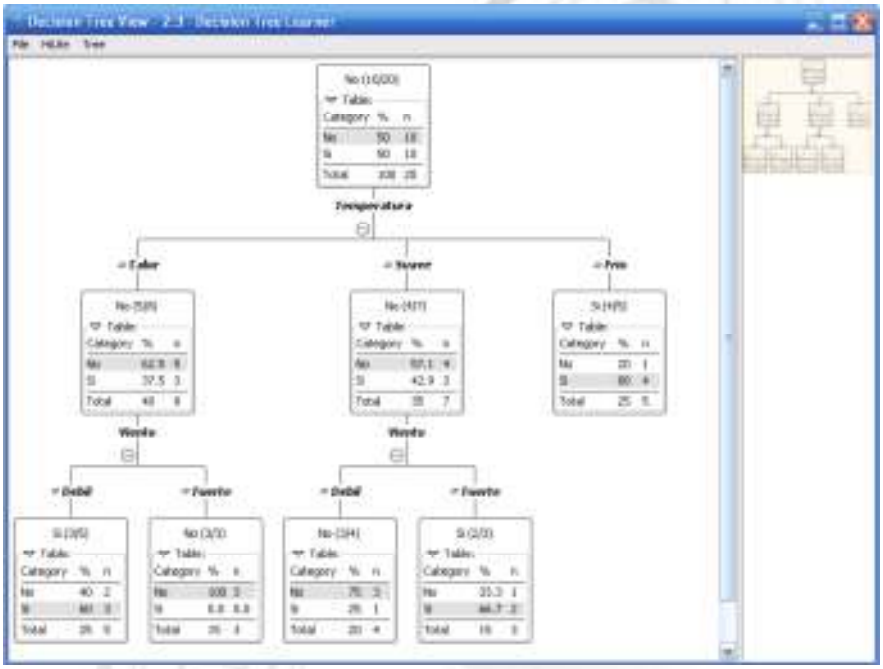
c) **Preprocesamiento** [*Data Manipulation*], para filtrar, discretizar, normalizar, filtrar, seleccionar variables…

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

También se encuentran dos documentos para leer en la documentación:

- Primeros pasos.pdf
- Quickstart.pdf

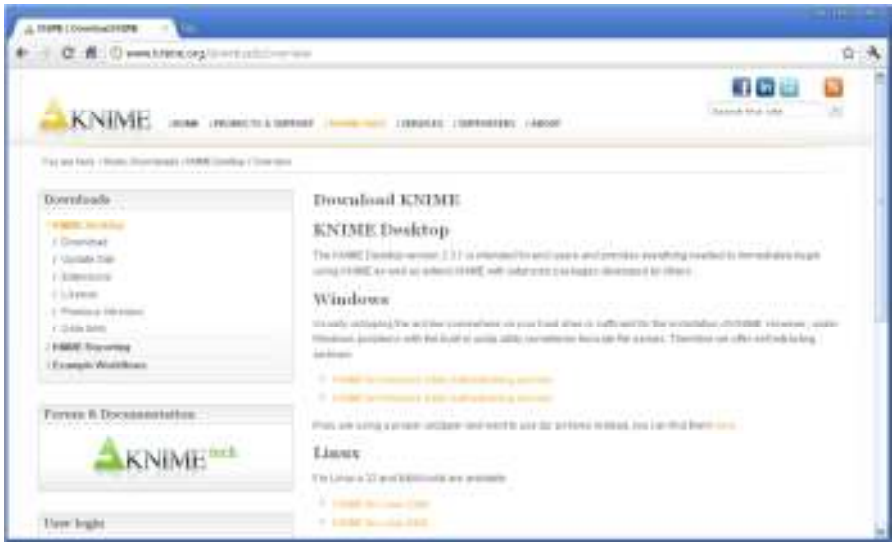
De forma complementaria, también puede consultar la siguiente presentación de KNIME en castellano:

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

Hay que familiarizarse con la herramienta siguiendo los documentos introductorios

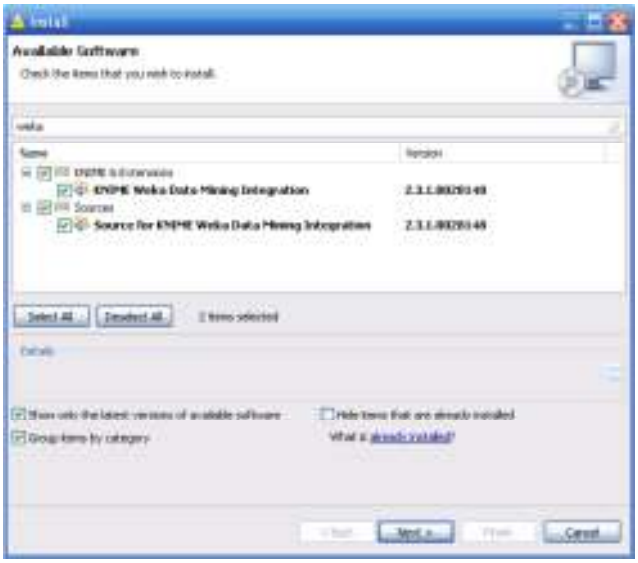
Instalación de KNIME (para trabajar autónomamente}

- 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.
- 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, Conexión con R, Tratamiento de textos y Cálculo de distancias para el cluster jerárquico, 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*. En la pantalla se describe la instlación de la extensión de Weka.



Información adicional en la guía de instalación paso a paso:

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

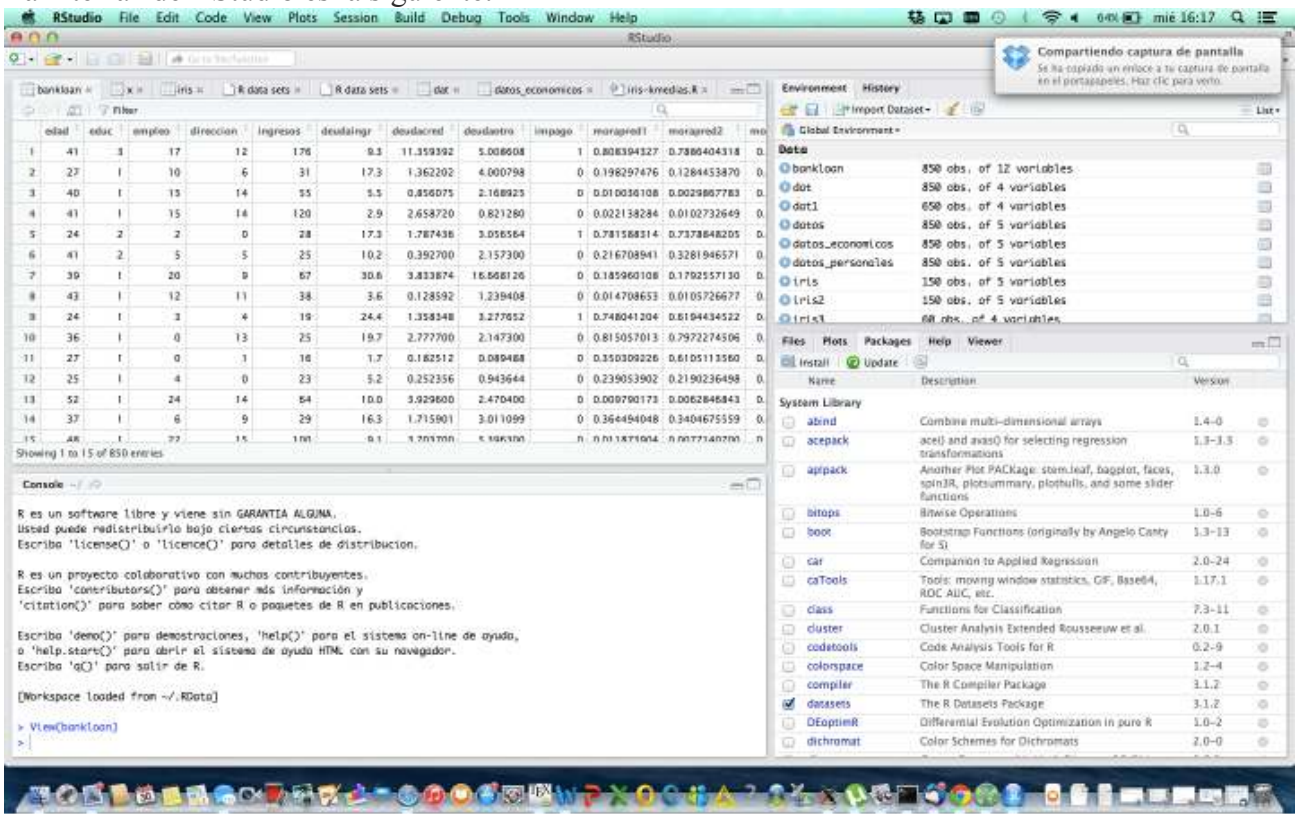
NOTA: KNIME no requiere programa de instalación. El paquete se desinstala, simplemente, borrando la carpeta en la que lo hayamos descomprimido.

Introducción a R y Rstudio

Como ya se comentó en el apartado anterior R es un entorno de software libre orientado, AL principio al cálculo estadístico y gráfico. R utiliza un lenguaje de programación sencillo, en línea de comandos que permite llamadas a funciones, de hecho se basa fundamentalmente en llamadas a funciones. Se puede extender muy fácilmente a través de paquetes de funciones ya que so programación es bastante simple y debido a ello se ha constituido en un referente de software libre para el desarrollo de técnicas estadísticas, de Minería de Datos, y en general de todo tipo de técnicas de tratamiento de datos. En **Short-refcard.pdf** se pueden encontrar un resumen de los comandos de programación de R y en **RDataMining-reference-card.pdf** se pueden encontrar las principales librerías de R que se utilizan en Minería de Datos. Para una información mas actualizada y exahustiva ir a https://cran.rstudio.com/web/views/

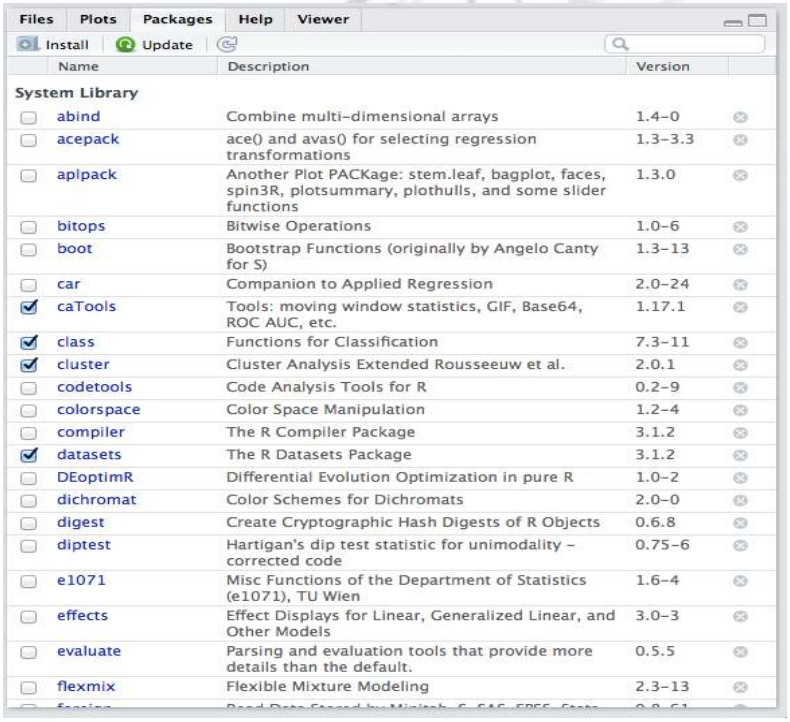
Si bien R se plantea como un lenguaje de comandos existen varias interfaces de usuario que hacen más fácil su utilización , R-commander, Tinn-R y RStudio, esta última será la interfaz que utilizaremos para las prácticas. Un buen resumen introductorio al R y RStudio se puede encontrar en **A (very) short introduction to R.pdf .**

La interfaz de RStudio es la siguiente:



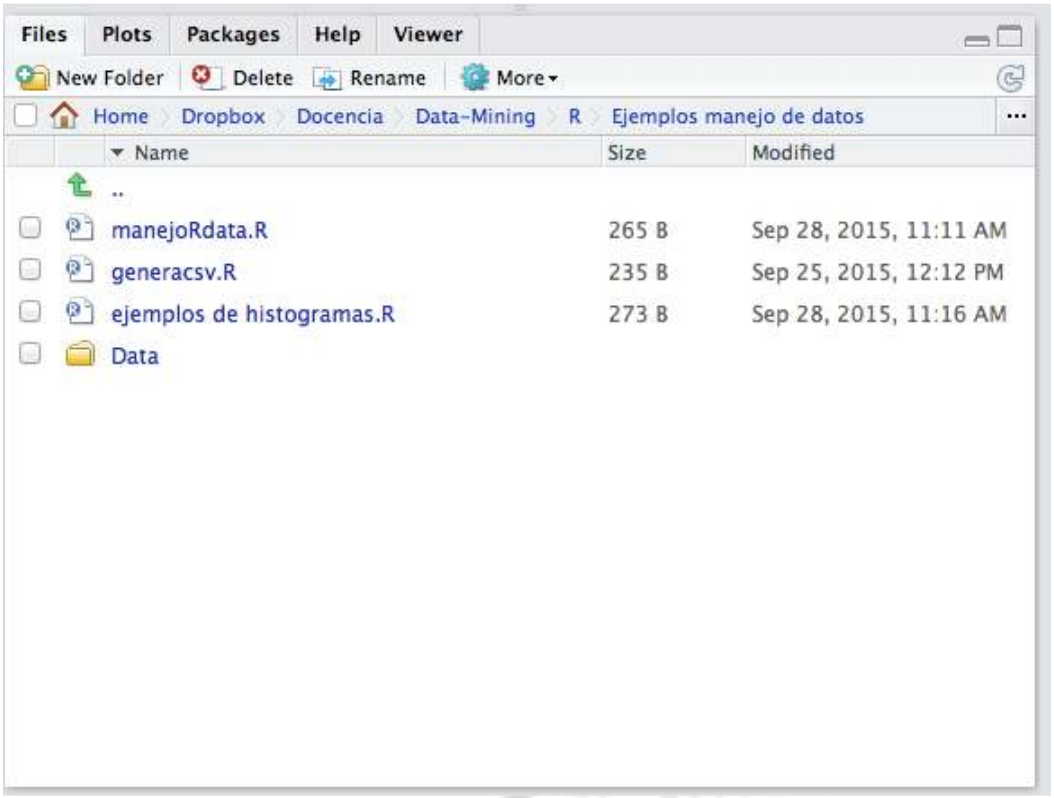
Como puede verse existen fundamentalmente cuatro ventanas.

1. La ventana inferior izquierda es la ventana de comandos de R
2. La inferior derecha permite distintas acciones dependiendo de las pestañas que se usan:
   - Se pueden cargar, y/o instalar librerías (packages)

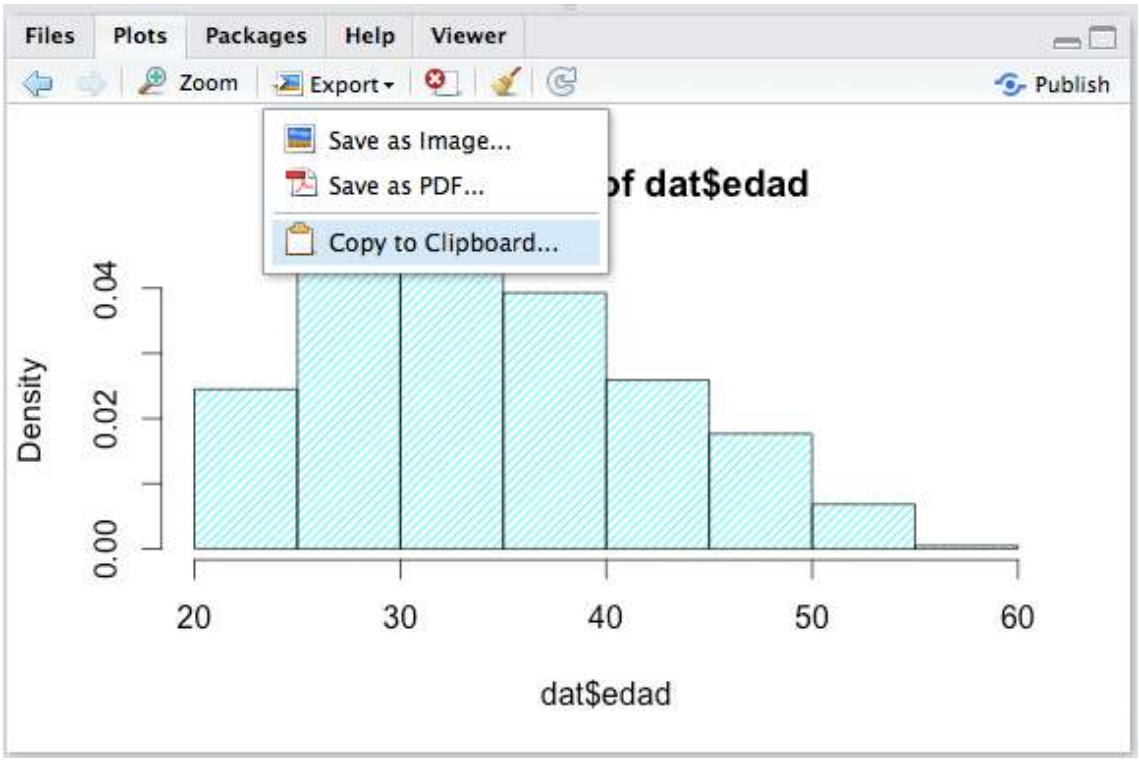


Name	Description	Version
System Library		
abind	Combine multi-dimensional arrays	1.4-0
acepack	ace() and avas() for selecting regression transformations	1.3-3.3
aplpack	Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions	1.3.0
bitops	Bitwise Operations	1.0-6
boot	Bootstrap Functions (originally by Angelo Canty for S)	1.3-13
car	Companion to Applied Regression	2.0-24
caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
class	Functions for Classification	7.3-11
cluster	Cluster Analysis Extended Rousseeuw et al.	2.0.1
codetools	Code Analysis Tools for R	0.2-9
colorspace	Color Space Manipulation	1.2-4
compiler	The R Compiler Package	3.1.2
datasets	The R Datasets Package	3.1.2
DEoptimR	Differential Evolution Optimization in pure R	1.0-2
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create Cryptographic Hash Digests of R Objects	0.6.8
diptest	Hartigan's dip test statistic for unimodality – corrected code	0.75-6
e1071	Misc Functions of the Department of Statistics (e1071), TU Wien	1.6-4
effects	Effect Displays for Linear, Generalized Linear, and Other Models	3.0-3
evaluate	Parsing and evaluation tools that provide more details than the default.	0.5.5
flexmix	Flexible Mixture Modeling	2.3-13

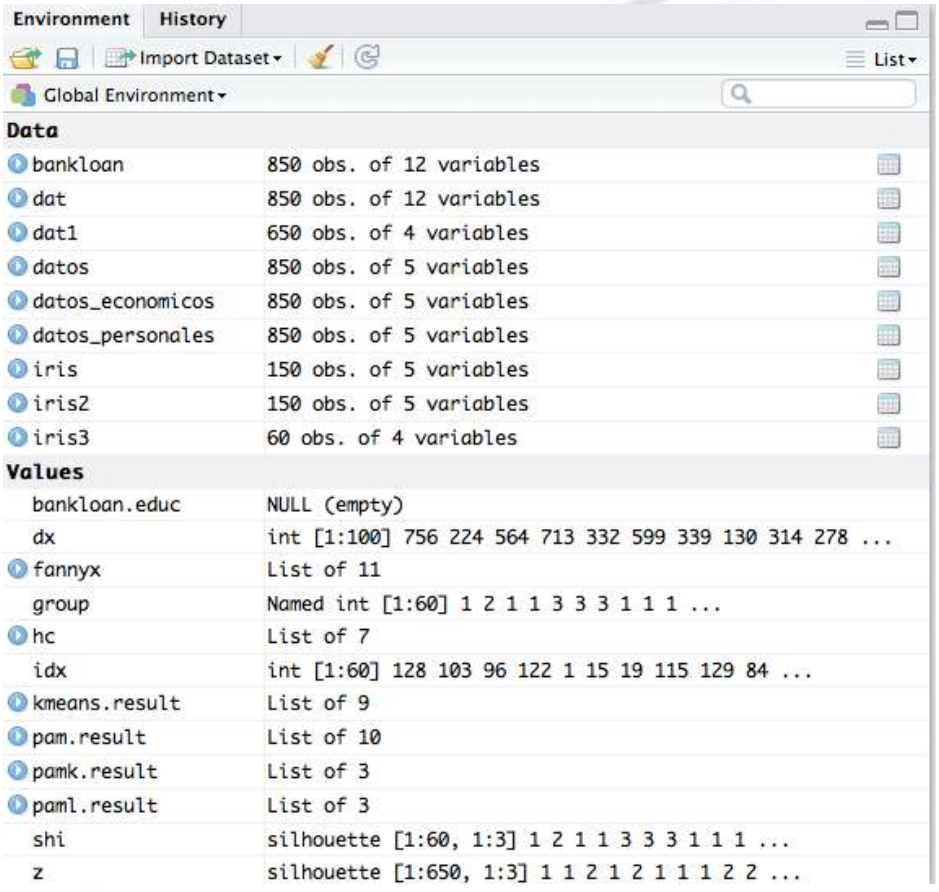
- Seleccionar el directorio de trabajo y dentro de él ficheros de datos o bien ficheros de comandos de R (scripts)



- Es en esta ventana donde aparecen los dibujos que se realizan y donde se permiten guadarlos como pdf o imágenes



3. La ventana superior derecha muestra los datos con los que se trabaja, los datos tipo "dataset" (data.frames) en R y los datos tipo values, que en R son objetos de muy distintos tipo, vectores , matrices, variables, También permite salvar, obtener e importar datos y ver el historiar dela ventana de comandos en R



4. La ventana superior izquierda permite edita ficheros de comandos (script) , generarlos, corregirlos y ejecutarlos.

también permite visualizar los data.frames.

Instalación de R y Rstudio

Para instalar R ir a http://www.r-project.org/ ir a CRAN buscar Spain y descargar e instalar.
Una vez instalado R se instala Rstudio y ya se puede trabajar, existen un conjunto de paquetes preinstalados y de datasets que se pueden usar, entre ellos el Iris que es uno de los que usaremos en las prácticas. Para activar un paquete basta marcarlo y si no se tiene preinstalado basta usar la pestaña install.

Introducción a SPSS

En estas prácticas utilizaremos también SPSS debido a su uso generalizado en la realización de estudios de tipo estadístico.

- SPSS es un paquete software para realizar análisis estadísticos.
- SPSS utiliza menús descriptivos y cuadros de diálogo simples para realizar las funciones solicitadas por el usuario.
- SPSS ofrece la posibilidad de ejecutar una serie de comandos especificados en los denominados ficheros de sintaxis.
- SPSS posee una estructura tipo modular.
- El módulo base forma el núcleo del sistema e incluye, tanto comandos de lectura y transformación de datos y ficheros, como procedimientos estadísticos básicos.
- En estas prácticas, utilizaremos como ejemplo la versión 20.0.

Ejecución de SPSS

Al ejecutar el programa desde el menú de inicio, se muestra una ventana desde la que se nos ofrecen diversas opciones para abrir ficheros de datos, introducir nuevos datos o ejecutar un tutorial.

La ayuda del SPSS es excelente y permite trabajar con él de manera autónoma. Como ejemplo estudiar los tipos de variables que se usan en SPSS siguiendo el tutorial. Se pueden encontrar en el apartado de datos.

Abrir además algún fichero de ejemplo (bankloan.sav) para ver qué variables tenemos.

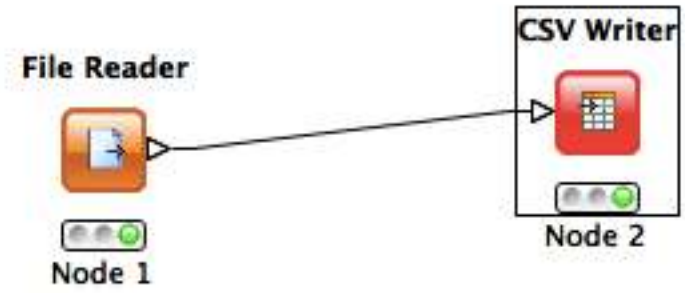
Ejemplos de trabajo

Vamos a familiarizarnos con las herramientas y para ello vamos a leer y transformar los datasets que se van a usar en las prácticas. En estas prácticas vamos a usar principalmente dos datasets:

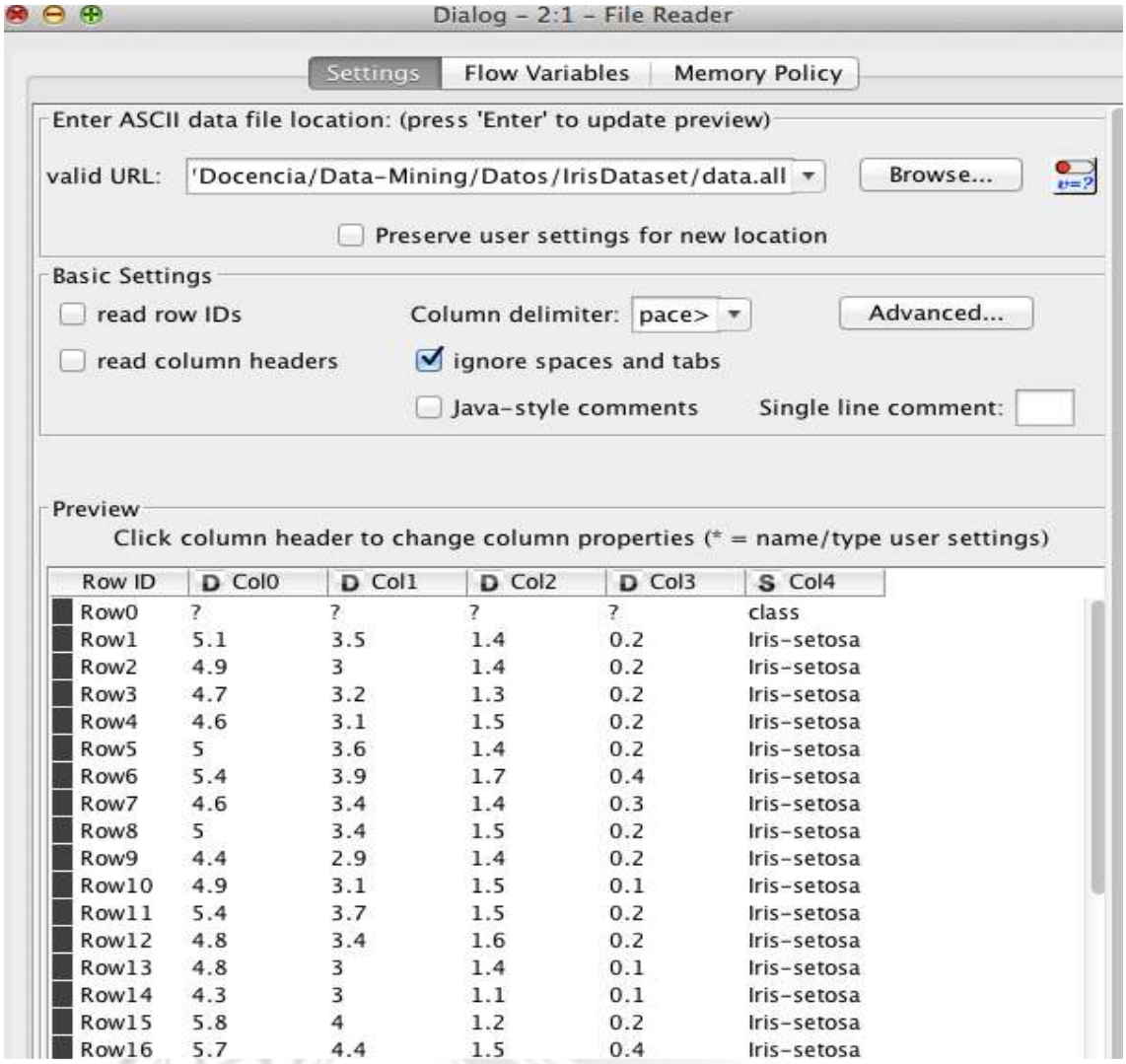
- Iris, es un dataset muy clásico que recoge 150 casos sobre los petalos y sepalos de varias especies de flores
- Bankloan, es un dataset de 850 casos que recoge datos sobre los clientes un banco que piden préstamos, incluyendo el hecho de que hayan pagado o no. Se incluyen datos personales, y datos de sueldo. También datos estimados de probabilidad de impago. En total 12 variables

Ejemplo de lectura de datos de dataset Iris y escritura del mismo como .csv

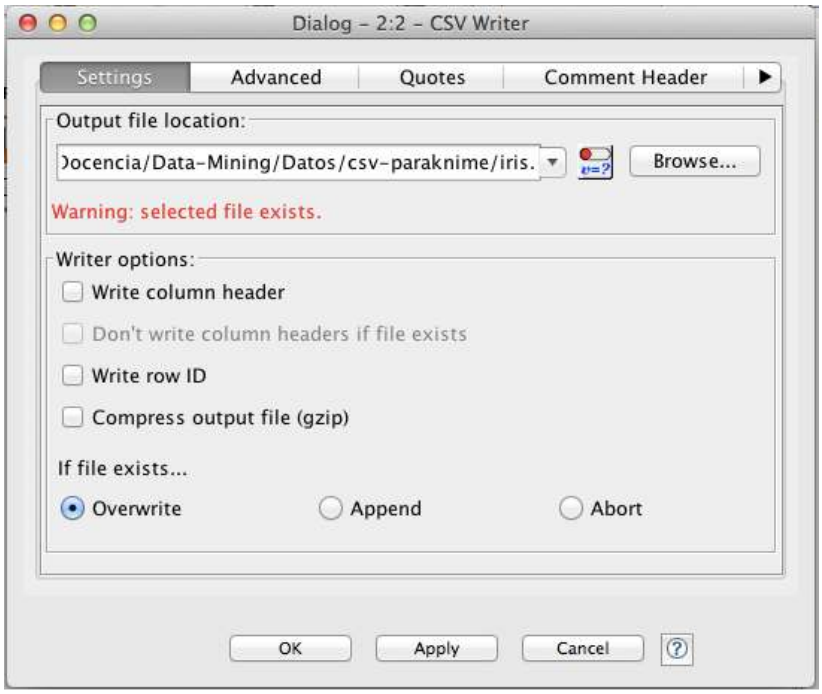
- Se puede leer un fichero de datos típico de Knime (.data) y excribirlo como .csv (formato de texto intercambiable para distintas herramientas SPSS, R Excel etc)
  Este es el flujo es



Y las configuraciones de los nodos de lectura y escritura



Row ID	D Col0	D Col1	D Col2	D Col3	S Col4
Row0	?	?	?	?	class
Row1	5.1	3.5	1.4	0.2	Iris-setosa
Row2	4.9	3	1.4	0.2	Iris-setosa
Row3	4.7	3.2	1.3	0.2	Iris-setosa
Row4	4.6	3.1	1.5	0.2	Iris-setosa
Row5	5	3.6	1.4	0.2	Iris-setosa
Row6	5.4	3.9	1.7	0.4	Iris-setosa
Row7	4.6	3.4	1.4	0.3	Iris-setosa
Row8	5	3.4	1.5	0.2	Iris-setosa
Row9	4.4	2.9	1.4	0.2	Iris-setosa
Row10	4.9	3.1	1.5	0.1	Iris-setosa
Row11	5.4	3.7	1.5	0.2	Iris-setosa
Row12	4.8	3.4	1.6	0.2	Iris-setosa
Row13	4.8	3	1.4	0.1	Iris-setosa
Row14	4.3	3	1.1	0.1	Iris-setosa
Row15	5.8	4	1.2	0.2	Iris-setosa
Row16	5.7	4.4	1.5	0.4	Iris-setosa



Hay que tener en cuenta que en la interfaz de escritura se pueden seleccionar los caracteres de separación y los decimales, para leerlo desde SPSS hay que seleccionar con ; y con coma decimal.

Hay que hacer notar que el fichero Iris.csv no incluye los nombre de las columnas y este datos, como veremos es necesario.

Para incluirlo podemos hacer varias cosas:

- Modificar el archivo mediante Excel
- Modificar el archivo leyendo de SPSS y salvando de nuevo
  1. Esto se hace leyendo el fichero como .csv y siguiendo las opciones de transformación.
     a. Incluir cabeceras
     b. Aceptar separación por ; y texto como “”
     c. Cambiando columna a columna el nombre de las variables
  2. Escribiendo en fichero como .csv.
- Capturando del dataset en Rstudio y salvando este como.csv

A la vista de estos ejemplos concluimos que: para trabajar con las distintas herramientas tomando .csv como formato de unifcado hay que tener en cuenta que:
-Knime y R son intercambiables (utilizan , como separador y . decimal
-SPSS utiliza ; como separador y , decimal.

- Para generar un fichero .csv de este tipo se puede utilizar Knime
- Para leer un fichero generado por SPSS de este tipo:
  - Desde de R se importa automáticamente (es la mejor opción porque luego se puede salvar con separador , y . decimal
  - Desde Knime especificando ; como separador; pero en este caso lee todas las variables como string y hay que convertirlas con un nodo adecuado

Ejemplo de manejo en distintos formatos para trabajo personal.

Los datos de bankloan se encuentran en formato SPSS (.sav) el objetivo es que sean accesibles desde Knime y desde Rstudio para ello tendremos:

a. Pasarlos a bankloan.csv desde SPSS
b. Leerlos con el flujo adecuado en Knime, comprobar mirando la tabla que están correctos. Ver los tipos de datos etc. Transformar algunos datos cambiando las columnas y salvando el fichero de salida como .csv
c. Importarlos desde Rstudio comprobando que están correctos. Obtener algunas columnas, visualizar mediante gráficas etc.

La realización de estas prácticas tiene como objetivo familiarizarse con Knime y con Rstudio para trabajar con los datos una vez leídos. En los pasos b y c se recomienda utilizar los manuales de introducción que se incluyen en el material para tratar con dataset, ver los tipos de variables etc.

Información adicional sobre herramientas software de minería de datos en KDnuggets:
http://www.kdnuggets.com/software/index.html

Fuentes de datos para encontrar datasets:

Además de los proporcionados por SPSS y Knime, se pueden consultar los existentes en el paquete Dataset de R y los siguientes sitios Web:

http://www.kdnuggets.com/datasets/index.html
http://archive.ics.uci.edu/ml/
http://www.webmining.cl/2011/01/15-datasets-gratis-para-data-mining/
http://www.rdatamining.com/resources/data
http://www.inf.ed.ac.uk/teaching/courses/dme/2011/datasets.html
http://www.r-bloggers.com/datasets-to-practice-your-data-mining/
http://www.infochimps.com/datasets
http://www.datawrangling.com/some-datasets-available-on-the-web
http://pages.cs.wisc.edu/~beechung/ref/datasets.html
http://www.cs.waikato.ac.nz/ml/weka/datasets.html
http://www.abbottanalytics.com/data-mining-resources-sets.php
http://www.dmg.org/pmml_examples/
https://www.kaggle.com/competitions
http://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html

Bibliografía

- C.Perez Técnicas de Análisis Inteligente de datos. Aplicaciones con SPSS (Pearson 2004)
- F.Berzal, J.C. Cubero Guiones de prácticas de S.I. de gestión (DECSAI)
- G.Bakor Knime Essentials Packt Publishing 2013
- IBM SPSS Documentos de ayuda (2014)
- J.P. Verma Data Analysis in Management with SPSS Software Springer (2013)
- R. Silipo KNIME Beginner's Luck Knime Press (2014)