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

# Tratamiento Inteligente de Datos

## Guión de Prácticas

## Práctica 2

## Preprocesamiento de datos

## Introducción

La Estadística es fundamental en minería de datos:

- Proporciona técnicas muy utilizadas en minería de datos (como, por ejemplo, el análisis de componentes principales, las técnicas de regresión o el análisis factorial).

- Sirve de filtro previo a la realización de distintos estudios de minería de datos. Por ejemplo, en un estudio que analiza qué variables son importantes para predecir el comportamiento de otra (clasificación), ¿hay variables correladas que se pudiesen suprimir antes de proceder a dicho estudio?

- Se utiliza como parte de las técnicas propias de minería de datos

A la hora de aplicar técnicas estadísticas, hemos de tener en cuenta lo siguiente:

1. Las técnicas estadísticas suelen requerir que el experto diga exactamente lo que quiere comprobar.

2. Cuando se aplican técnicas estadísticas "clásicas", hay que tener cuidado de que se cumplan ciertos "requerimientos" o "hipótesis de partida". En caso contrario, hay que aplicar técnicas "no paramétricas".

En general todas las herramientas de DM incluyen técnicas de análisis estadísticos, si bien algunas de ellas son más completas que otras. En ese sentido SPSS y R son obviamente más completos. Si bien Knime también ofrece un buen grupo de técnicas. En esta práctica utilizaremos las tres herramientas. El objetivo es:

- Familiarizarnos con los tipos de datos
- Usar técnicas de exploración de datos, tanto estadísticas como de visualización.
- Familiarizarnos con las herramientas de transformación de datos.
- Trabajar con componentes principales y con cambios de escala
- Tratar con valores perdidos.

## Tipos de datos.

*En SPSS.*

1. Abrir SPSS y cargar el fichero Iris.csv
2. En el visor de variables: ver los tipos de datos que ofrece SPSS.
3. Se pueden reconocer todos datos vistos en teoría

En SPSS, los nombres de las variables no pueden tener más de 8 letras, pero se les puede poner una etiqueta más larga que luego saldrá en los gráficos (columna *Etiqueta*).

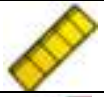
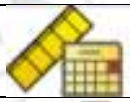
A la hora de declarar variables, es muy importante escoger adecuadamente la combinación Tipo de dato – Medida

**Medidas** (establecen qué mide la variable):

- Nominal: Una variable que toma valores no ordenados (p.ej. color de pelo).
- Ordinal: Una variable que toma valores ordenados (p.ej. nivel de satisfacción, medido de 0 a 5).
- Escala: Una variable que toma valores numéricos, para los que tiene sentido la operación de resta (p.ej. edad).

**Tipos** (establece cómo codificamos lo que la variable mide):

- Numérico (con una precisión determinada).
- Cadena (típica cadena de caracteres)
- Otros: Dólar, fecha, etc.

| **Nivel de Medida** | **Tipo de datos** | | | |
|---|---|---|---|---|
| | Numérico | Cadena | Fecha | Tiempo |
| Escala |  | n/a |  |  |
| Ordinal |  |  |  |  |
| Nominal |  |  |  |  |

*En Knime*

Para trabajar con el mismo dataset en SPSS que en Knime debemos preparar un fichero .csv que tenga idénticas características. Para ello:

1. Se guarda el fichero como .csv con el SPSS
2. Se abre un nodo de lectura csv en Knime y se configura:
   a. Delimitado por ;
   b. Con el título de columnas
3. Se ejecuta y puede verse (ver tabla) que las variables numéricas con decimales las ha pasado a string (nadie es perfecto, ni siquiera Knime)
4. Tendremos que convertir estas variable utilizando
   a. Un nodo de cambio de string a número (configurar con coma decimal)
   b. Un nodo de escritura en un nuevo fichero csv que ya estará correcto.

5. Para comprobar que está bien abrir otro nodo de lectura, configurar ahora separado por comas y ver que está bien.

A través de estos pasos pueden verse los tipos de variables que usa Knime. En términos muy simples utiliza diversos tipos de variables numéricas y variables string que son las únicas que pueden ser usadas como variables nominales para categorizar.

*En R (Rstudio)*

R opera con tipos básicos de datos que son, números reales, enteros, valores lógicos y cadenas de caracteres. Sobre estos elementos básicos se pueden montar: vectores, matrices, data.frames , listas, y factores que son un tipo especial de vectores de datos categorizados.
Para ampliar estos conceptos, ver notaciones etc. conviene repasar los primeros capítulos de "An Introduction to R"

## Ejercicios que son el objetivo de esta sección de prácticas

El conjunto de ejercicios a realizar en las distintas herramientas se corresponde con los conceptos y técnicas explicados en el tema 2. Por ello, para cada dataset se deben obtener:

1. Medidas estadísticas. Media, mediana, varianza, cuantiles etc. Covarianzas y/o correlaciones
2. Gráficos: histogramas y diagramas de frecuencias con distintos intervalos, diagramas de sectores, diagramas de cajas. Gráficos de puntos. Otros gráficos multidimensionales
3. Cambios de escala, normalizaciones.
4. Análisis en componentes principales

En clase se desarrollarán los ejercicios correspondientes al Iris. Como trabajo personal de los alumnos se deben replicar los ejercicios con los datos de bankloan

## Exploración de datos usando estadística descriptiva y visualización. Ejercicios (1,2)

### *En SPSS*

Existen dos formas de hacerlo:

- Entrando en análisis- estadística descriptiva.
- Utilizando el generador de gráficos
-

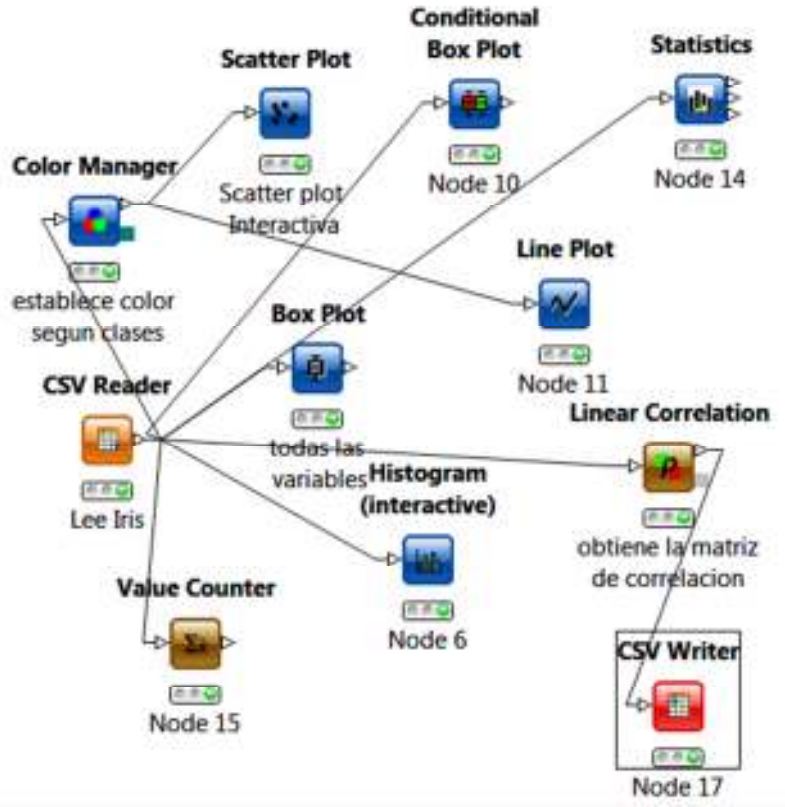
***Se recomienda entrar en ambos módulos y probar la potencialidad de la herramienta con los datos de iris.***

### *En Knime*

Se utilizan dos grupos de nodos:

- El grupo de estadística que permite calcular los datos más habituales y trabajar con la matriz de correlación
- El grupo de visión de datos, que incluye herramientas generales propiedades como la de manejo del color (permite colorear casos según los valores de un atributo categórico y que eso se transmita a los gráficos) y herramientas para la generación de gráficos de muy diversos tipos.

La siguiente figura muestra un ejemplo de flujo de datos:




.

***Se deben rehacer los flujos de datos, aplicándolos a distintas variable de Bankloabn. Hacer uso del manejo de color en gráficos. Obtener la matriz de correlación de manera que pueda ser usada como archivo csv y visualizada en Excel***

*En Rstudio*

Como ya se ha indicado en su momento R trabaja con ficheros script. Los ficheros que permiten obtener todos estos resultados se encuentran en el material de la asignatura. Concretamente se pueden trabajar con los scripts: ***estadística-descriptiva-unidimensionl*** y ***estadística-descriptiva-bidimensional***

***Se deben probar los scripts viendo los resultados y estudiando las funciones utilizadas, se deben rehacer los scripts aplicándolos a algunas variables de Bankloan***

## Tranformación de los datos. (4)

*En SPSS*

En SPSS se pueden transformar los datos utilizando el módulo de transformaciones, permite agrupar datos categóricos, permite discretizar datos numéricos por distintos criterios y reemplazar valores perdidos de variable numéricas. Se puede normalizar mediante transformaciones.

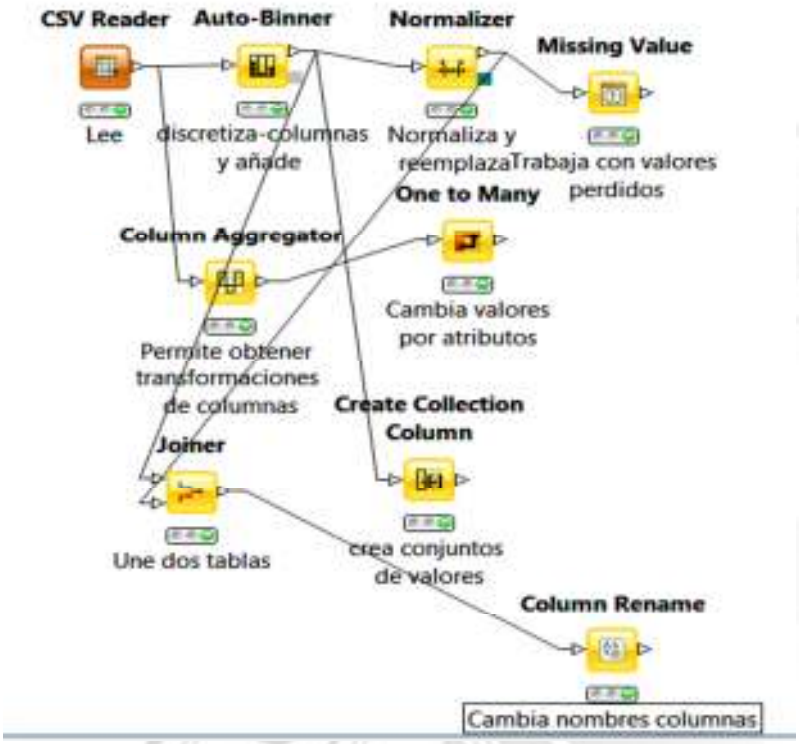
Hay una herramienta más completa de análisis de valores perdidos en módulo de analizar.

***Se recomienda entrar en ambos módulos y probar la potencialidad de la herramienta con los datos de iris.***

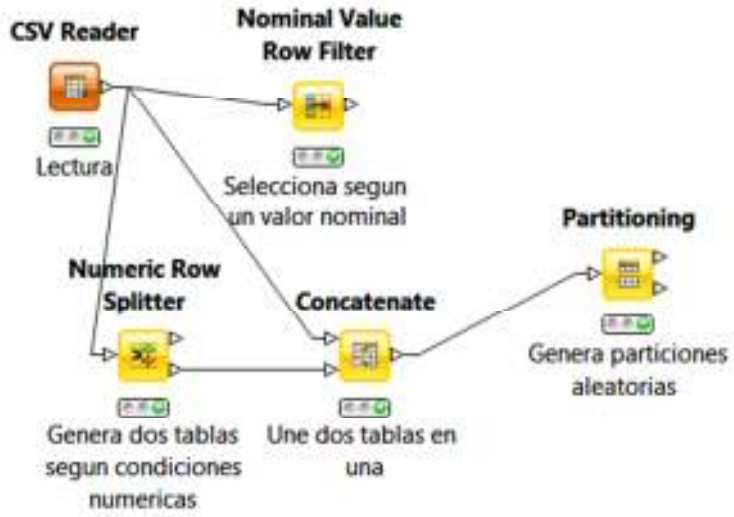
*En Knime*

Knime posee un módulo de manipulación de datos que recoge un amplio número de posibilidades de transformación del data set:

- Por columnas: discretización, transformaciones y generación de columnas de muy diversos tipos, agregación, normalización según distintos criterios etc… La siguiente figura muestra un flujo de ejemplo




- Por filas, permite agrupar, partir filtrar etc.. La siguiente figura muestra un flujo de ejemplo:




***Explorar las posibilidades de Knime para generar las transformaciones que se han visto en teoría. Generar nuevos conjuntos de datos csv con las variables transformados. Se aplicarán a bankloan***

*En R/Rstudio*

Al ser realmente un lenguaje de programación R es mucho más flexible en el tratamiento de datos que las otras herramientas, siempre es posible, mediante un script adecuado:

- Seleccionar columnas de un data.frame, transformarlas (normalizar etc.) y generar un nuevo dataset, combinando transformaciones y columnas originales, se puede ver en el script ***generacion de variables normalizadas.R***
- Filtrar y extraer distintas filas un data set. En este sentido se recomienda probar las posibilidades de R que se encuentran en R-reference-card y explorar las posibilidades de los paquetes ***dplyr*** y ***tidyr***. A título de ejemplo se propone un ejemplo de filtrado con la función "filter"que está en el paquete dplyr que está en el sript ***transformación por filas***. Otro interesante ejemplo de filtrado se encuentra en el script ***otro ejemplo con más variables.R***

## Reducción de variables

*En SPSS*

En general se permite siempre seleccionar las variables que van a participar en el proceso pero no hay un procedimiento automático de selección

*En Knime*

Algunas de las transformaciones de columnas de Knime podrían utilizarse en un procedimiento de reducción de variables pero depende muy exclusivamente del modelo.

En general todos los nodos de minería permiten seleccionar las variables que van a intervenir en el proceso, pero no hay un procedimiento automático de selección

*En R/Rstudio*

En general se permite seleccionar siempre las variables que van a participar en el proceso pero no hay un procedimiento automático de selección

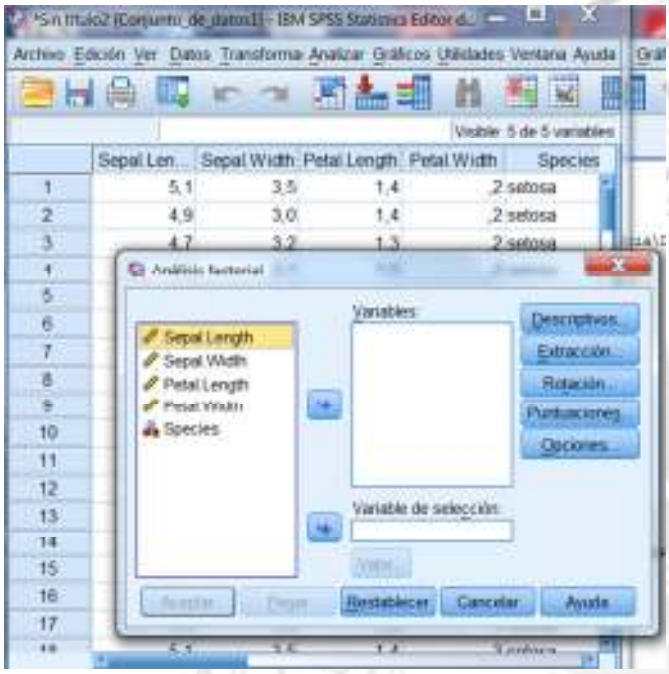
## Analisis en componentes principales o análisis factorial

*En SPSS*

SPSS contempla el análisis factorial en analizar-reducción de variables. Permite almacenar los factores como columnas adicionales del mismo fichero. Probablemente es la herramienta que mejor desarrolla el análisis factorial de todas las consideradas.
Permite elegir variables y diversas opciones acerca de:

- Método de calculo
- Criterio de selección de factores
- Forma de salida, gráficos, puntuaciones añadidas a los datos etc.

En el siguiente ejemplo:
1.- Seleccionamos variables y forma de calcular




2.- La salida, tomando 2 factores, método de componentes principales y rotación varimax es:

## A. factorial

[Conjunto_de_datos1]

**Comunalidades**

| | Inicial | Extracción |
|---|---|---|
| Sepal.Length | 1,000 | ,923 |
| Sepal.Width | 1,000 | ,991 |
| Petal.Length | 1,000 | ,994 |
| Petal.Width | 1,000 | ,936 |

Método de extracción: Análisis de Componentes principales.

**Varianza total explicada**

| Componente | Autovalores iniciales | | | Sumas de las saturaciones al cuadrado de la extracción | |
|---|---|---|---|---|---|
| | Total | % de la varianza | % acumulado | Total | % de la varianza |
| 1 | 2,918 | 72,962 | 72,962 | 2,918 | 72,962 |
| 2 | ,914 | 22,851 | 95,813 | ,914 | 22,851 |
| 3 | ,147 | 3,669 | 99,482 | | |
| 4 | ,021 | ,518 | 100,000 | | |

**Varianza total explicada**

| Componente | Sumas de las saturaciones al ... | Suma de las saturaciones al cuadrado de la rotación | | |
|---|---|---|---|---|
| | % acumulado | Total | % de la varianza | % acumulado |
| 1 | 72,962 | 2,700 | 67,489 | 67,489 |
| 2 | 95,813 | 1,133 | 28,325 | 95,813 |
| 3 | | | | |
| 4 | | | | |

Método de extracción: Análisis de Componentes principales.

**Matriz de componentes[a]**

| | Componente | |
|---|---|---|
| | 1 | 2 |
| Sepal.Length | ,890 | ,361 |
| Sepal.Width | -,460 | ,883 |
| Petal.Length | ,992 | ,023 |
| Petal.Width | ,965 | ,064 |

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

**Matriz de componentes rotados[a]**

| | Componente | |
|---|---|---|
| | 1 | 2 |
| Sepal.Length | ,959 | ,046 |
| Sepal.Width | -,143 | ,985 |
| Petal.Length | ,944 | -,306 |
| Petal.Width | ,932 | -,259 |

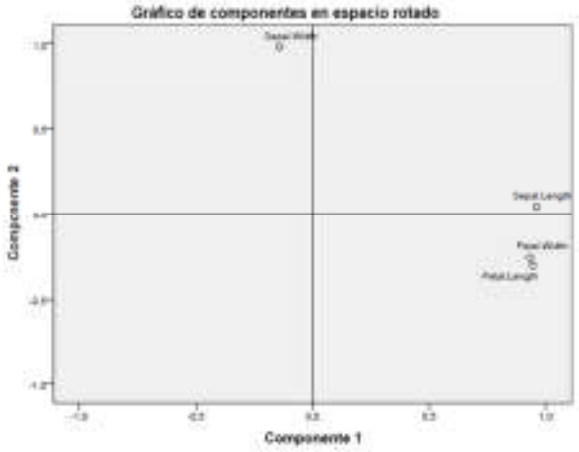
Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

**Matriz de transformación de las componentes**

| Componente | 1 | 2 |
|---|---|---|
| 1 | ,944 | -,331 |
| 2 | ,331 | ,944 |

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varimax con Kaiser.

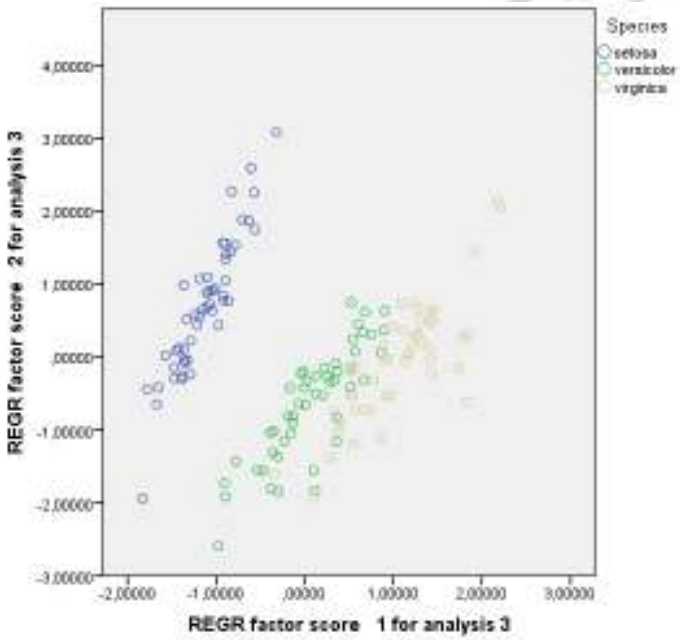



3.- Tambien se puede elegir la salida de las puntuaciones (scores) de los datos con respecto a los factores aparecen como columnas nuevas en la tabla de datos

| | Sepal Len... | Sepal Width | Petal Length | Petal Width | Species | FAC1_1 | FAC1_2 |
|---|---|---|---|---|---|---|---|
| 1 | 5,1 | 3,5 | 1,4 | ,2 | setosa | -1,32123 | -1,08159 |
| 2 | 4,9 | 3,0 | 1,4 | ,2 | setosa | -1,21404 | -1,37808 |
| 3 | 4,7 | 3,2 | 1,3 | ,2 | setosa | -1,37930 | -1,41959 |
| 4 | 4,6 | 3,1 | 1,5 | ,2 | setosa | -1,34147 | -1,47191 |
| 5 | 5,0 | 3,6 | 1,4 | ,2 | setosa | -1,39424 | -1,09002 |
| 6 | 5,4 | 3,9 | 1,7 | ,4 | setosa | -1,21093 | -,62979 |
| 7 | 4,6 | 3,4 | 1,4 | ,3 | setosa | -1,42585 | -1,32931 |
| 8 | 5,0 | 3,4 | 1,5 | ,2 | setosa | -1,30265 | -1,15256 |
| 9 | 4,4 | 2,9 | 1,4 | ,2 | setosa | -1,36203 | -1,66977 |
| 10 | 4,9 | 3,1 | 1,5 | ,1 | setosa | -1,27434 | -1,36432 |
| 11 | 5,4 | 3,7 | 1,5 | ,2 | setosa | -1,26383 | -,83321 |
| 12 | 4,8 | 3,4 | 1,6 | ,2 | setosa | -1,35707 | -1,23496 |
| 13 | 4,8 | 3,0 | 1,4 | ,1 | setosa | -1,29425 | -1,47258 |
| 14 | 4,3 | 3,0 | 1,1 | ,1 | setosa | -1,53616 | -1,78111 |
| 15 | 5,8 | 4,0 | 1,2 | ,2 | setosa | -1,28275 | -,56979 |
| 16 | 5,7 | 4,4 | 1,5 | ,4 | setosa | -1,31978 | -,32008 |
| 17 | 5,4 | 3,9 | 1,3 | ,4 | setosa | -1,28791 | -,70436 |
| 18 | 5,1 | 3,5 | 1,4 | ,3 | setosa | -1,32285 | -1,02782 |

y con un uso adecuado del editor de gráficos

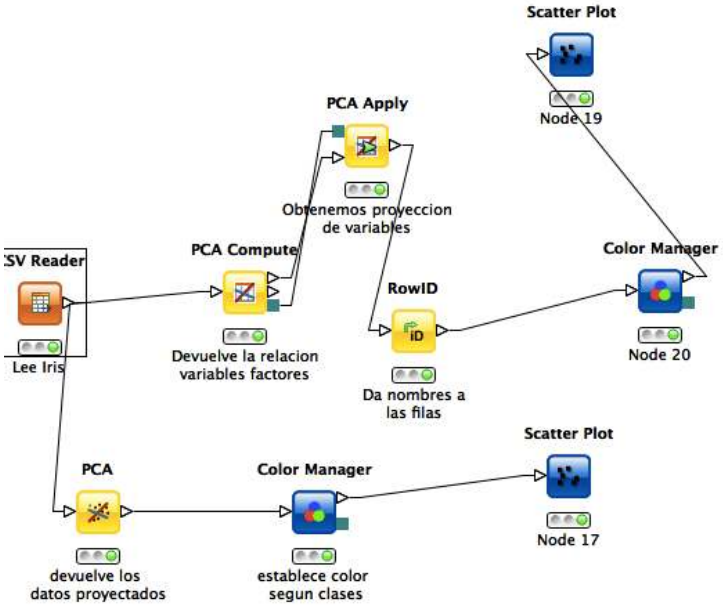

Obtener la representación de los datos según factor y clase



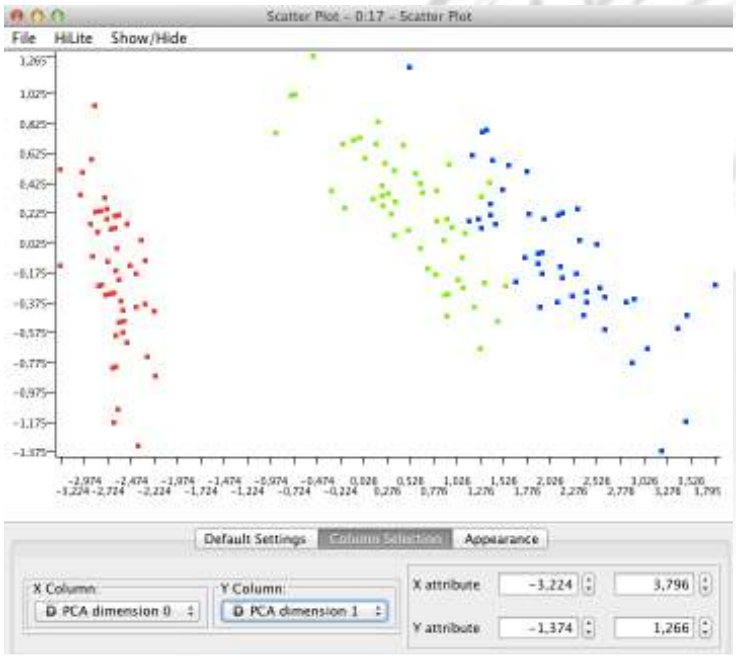

*En Knime*

Se realiza el análisis en componentes principales mediante el subgrupo de nodos PCA que está en el grupo mining

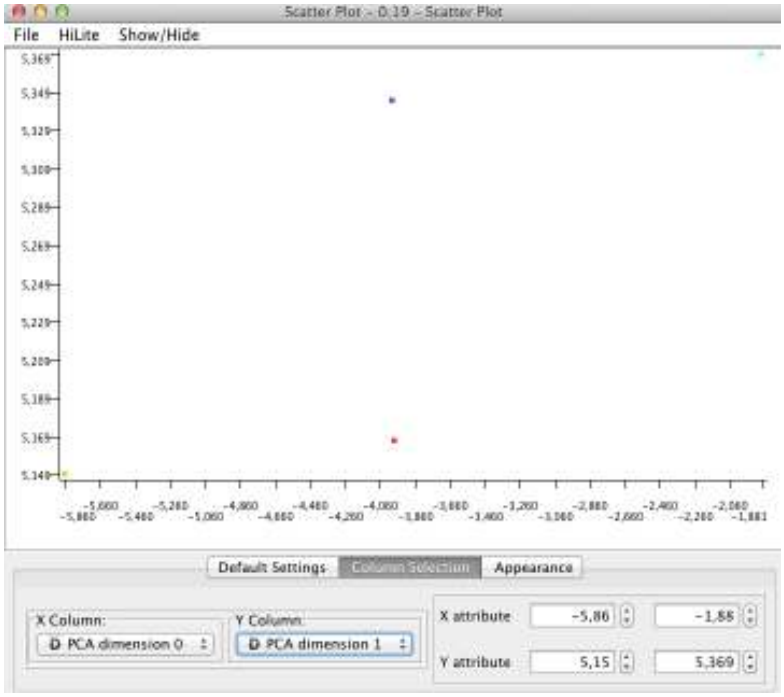
El siguiente flujo muestra la forma de hacerlo para los datos de iris, obteniendo además una representación de los datos con respecto  a dos factores y de los variables con respecto a dos factores




Los gráficos que se obtienen son:
1.- Representación de datos



2.- Representación de variables




*En R/Rstudio*

En R el análisis factorial se realiza mediante la función ***factanal*** que se encuentra dentro del paquete ***stat*** Ejemplos de su utilización sencilla pueden encontrarse en los scripts ***análisis factorial.R*** y ***otro ejemplo con más variables.R***

***Realizar un análisis en componentes principales de los datos numéricos (todos menos nivel educativo e impago) del dataset reproduciendo los resultados expuestos en teoría.***

***Hacerlo con varias herramientas. Cual parece más completo? Cuantos factores se tomarían? . Se pueden identificar los factores?***

## 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)
- Y. Zhao R. and Data Mining. Examples and Case Studies (2013)