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

Tratamiento Inteligente de Datos

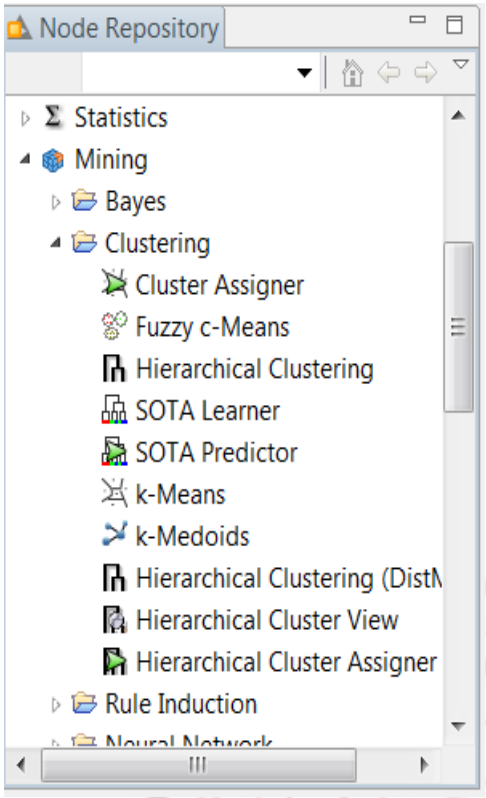
Guión de practicas 3

Métodos de agrupamiento [Clustering]

FICHEROS DE DATOS

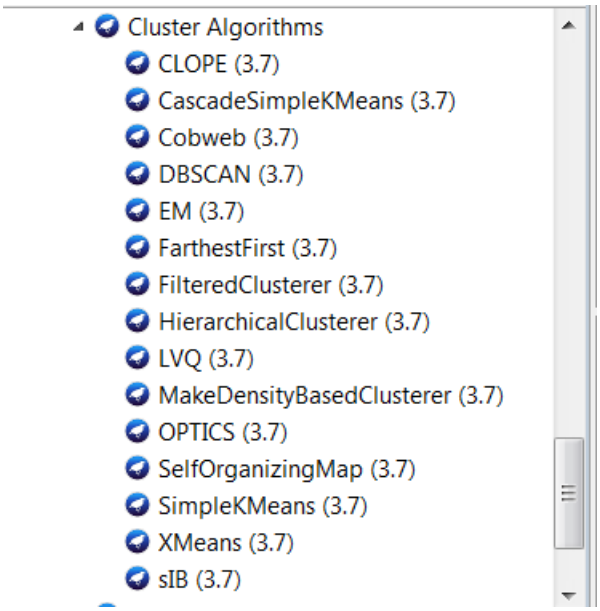
```
bankloan.csv
iris.csv
demo.csv
demo_cs.csv
```

Clustering en Knime



Como puede verse en la figura Knime proporciona los elementos más habituales en clustering, K-medias, k-medoides, y cluster jerarquico incluyendo, como elemento adicional un módulo de tratamiento de funciones de distancia que veremos posteriormente.

El módulo adicional de Weka es más completo



Pero al ser importado no permite casi ninguna opción y selecciona todos los elementos de los ficheros, por ello si se va a usar se recomienda antes preparar los datos con cuidado, seleccionando las variables, normalizando etc.

Como mecanismo de actuación la mejor solución es trabajar con los algoritmos de Knime de forma exploratoria y si la solución no es satisfactoria , o se quiere refinar, utilizar algún método adicional pero teniendo ya más claro que variables vamos a usar y por qué buscamos otro método.

K-Means en KNIME (utilizaremos datos de iris)

Cree un proyecto en KNIME llamado `k-medias-iris` con los siguientes nodos que nos permitan agrupar los datos de `iris dataset`

- Un nodo para leer el fichero buscar el data set

- Un nodo *Data Manipulation* > *Column* > *Normalizer* para normalizar los datos, ya que KNIME implementa k-Means sin normalizar previamente las variables. Con los datos obtenidos desde el applet no hay ningún problema, ya que ambas variables tienen magnitudes similares. Sin embargo, es bueno que nos acostumbremos a normalizar las variables.

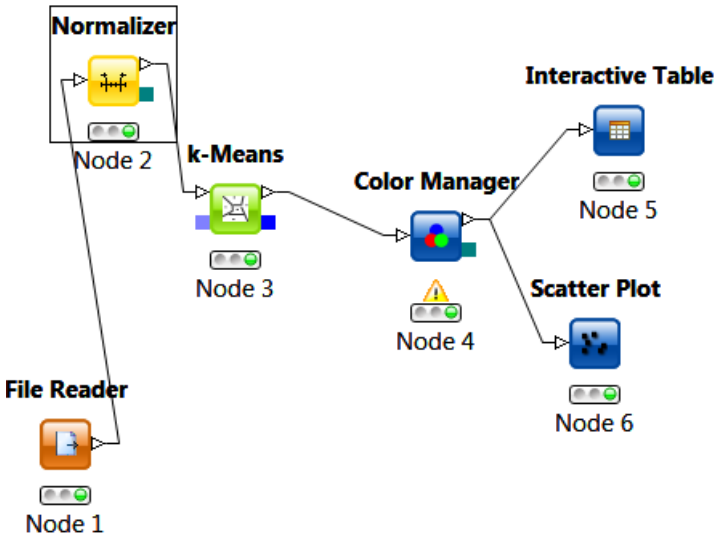
  NOTA: Recordemos que, en SPSS, se hacía de la forma siguiente: *Analizar* > *Estadísticos Descriptivos* > *Descriptivos* > *Guardar valores tipificados como variables*.

- Un nodo *Mining* > *Clustering* > *k-Means* para realizar el clustering. Este nodo añadirá una columna *Cluster*, indicando el agrupamiento asignada a cada tupla de nuestro conjunto de datos. En su configuración, debemos indicar los atributos que se usarán para establecer los clusters. En nuestro ejemplo, usaremos ambos.

- Un nodo *DataViews* > *Property* > *Color Manager* para colorear los datos correspondientes a la columna *Cluster* del nodo anterior.

- Un nodo *DataViews* > *Interactive Table* para ver los resultados.

- Dos nodos *DataViews* > *Scatter Plot* para ver la nube de puntos original y la coloreada con los clusters obtenidos.



Para ver los centroides obtenidos, seleccione con la derecha el nodo *k-Means* y muestre los resultados (*View: Cluster View*).

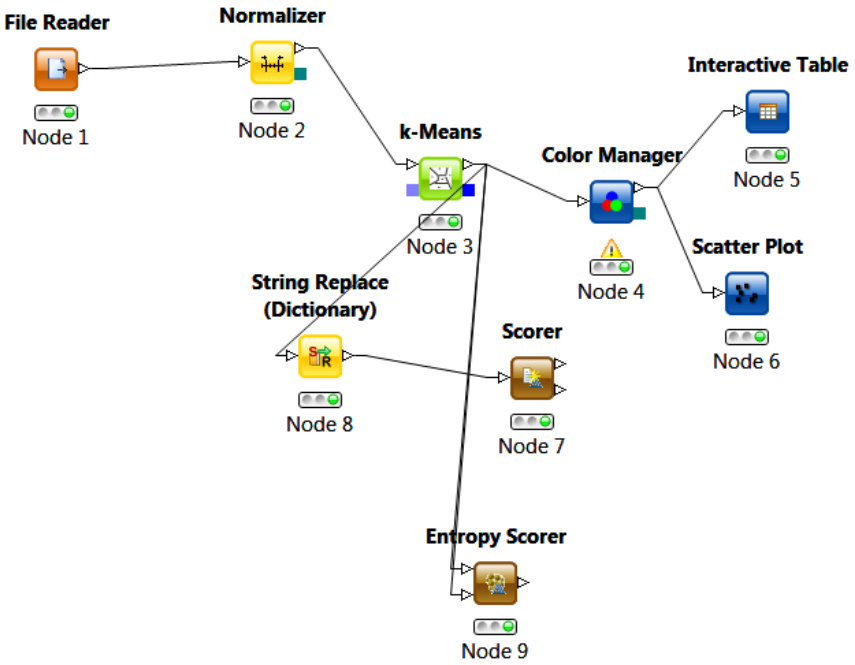
Para ver la calidad de la solución obtenida abrimos *Mining>Scoring>Entropy Scorer* y lo conectamos con k-means, este nodo permite comparar la solución obtenida con una variable de referencia que en este caso será Species. La salida es:

Row ID	Size	Entropy	Norm...	Quality
cluster_2	50	0	0	?
cluster_0	39	0.391	0.247	?
cluster_1	61	0.777	0.49	?
Overall	150	0.418	0.264	0.736

Tambien se puede ver la calidad del agrupamiento con medias de precisión. Esto se hace comparando el agrupamiento y la especie como si fuesen clasificaciones distintas. Para ellos se utiliza *Mining>Scoring> Scorer* . Pero hay que generar previamente una columna donde se identifica cada cluster con su clase más probable. En este caso, vista la tabla con colores sería:

virginica	cluster_0
versicolor	cluster_1
setosa	cluster_2

El nodo que se usa es *string-replace(dictionary)*, habiendo creado previamente desde Excel una tabla .csv donde están los datos de identificación. El flujo completo sería:



Y la salida que obtenemos es para Scorer es:

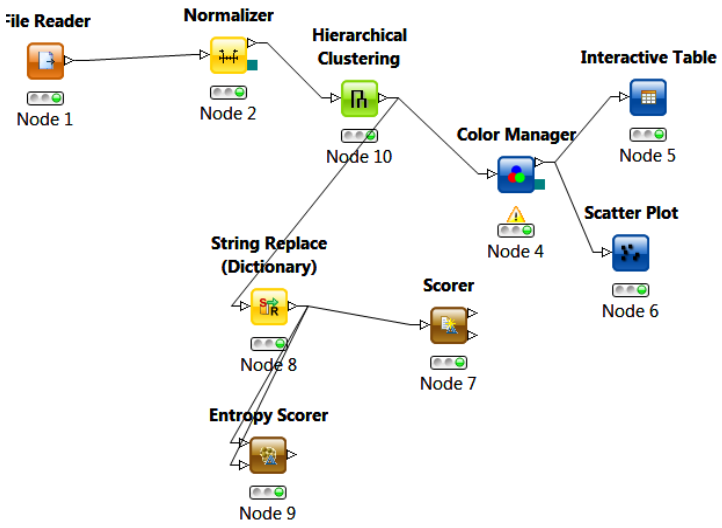
Row ID	TrueP...	FalseP...	TrueN...	False...	Recall	Precisi...	Sensit...	Specifity	F-mea...	Accur...	Cohen...
setosa	50	0	100	0	1	1	1	1	1	?	?
versicolor	47	14	86	3	0.94	0.77	0.94	0.86	0.847	?	?
virginica	36	3	97	14	0.72	0.923	0.72	0.97	0.809	?	?
Overall	?	?	?	?	?	?	?	?	?	0.887	0.83

Ejercicios:

- **Intentar mejorar la calidad del agrupamiento seleccionando adecuadamente las variables.**
- **Probar otros métodos de agrupamiento con numero de grupos conocidos como k-medias difusas, dbscan o k-medioides comparando resultados.**
- **Rehacer el flujo para un conjunto de variables de Bankloan**

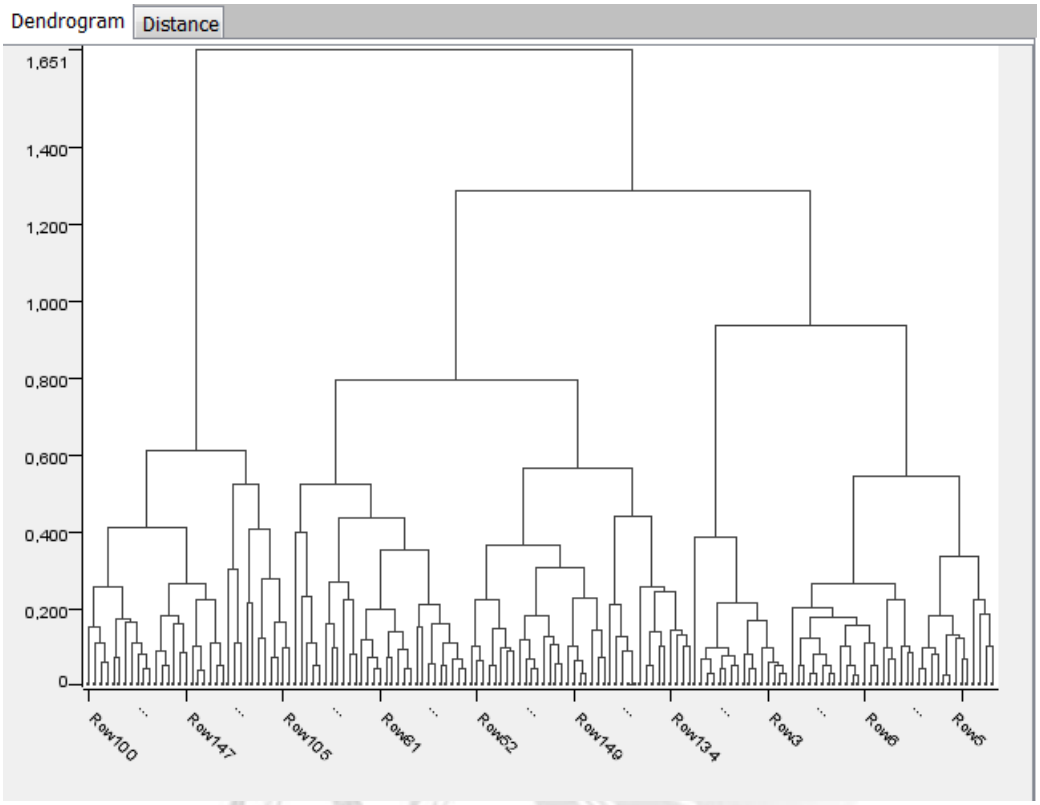
Clustering jerárquico en KNIME

El flujo anterior se puede rehacer sustituyendo el nodo k-means por hieralchical clustering, en la configuración se puede ver que permite elegir distancias y método de agrupamiento.

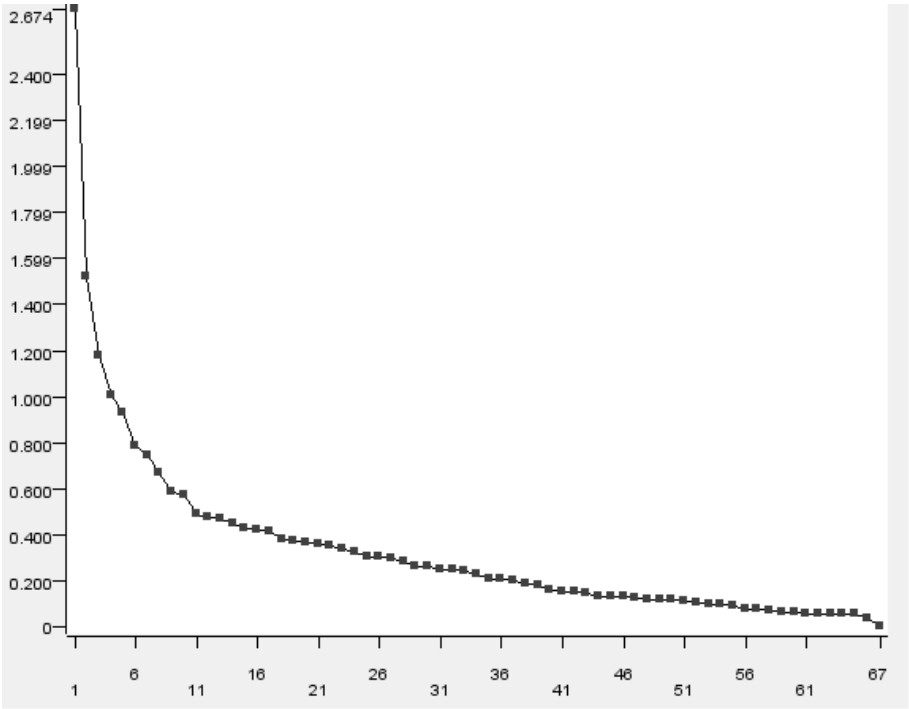


Ejecutando el nodo, obtenemos dos gráficos:

1.- En una se muestra el dendrograma y permite deducir cómo se agrupan los puntos



2.- En la ventana en la que se muestra el dendrograma, hay una pestaña *Distance* que muestra los valores de SSE (la suma de las distancias al cuadrado de cada punto al centroide de su cluster). Esto puede resultar bastante útil para determinar el valor más adecuado de k. Cuanto mayor sea k, menor será el valor de SSE, pero habrá un punto a partir del cual la ganancia no será significativa. Este será el punto que determinará el valor más adecuado para k.



Esta gráfica nos permite ajustar mejor el número de grupos en un entorno completamente no supervisado y nos da una medida no supervisada de la bondad del agrupamiento.

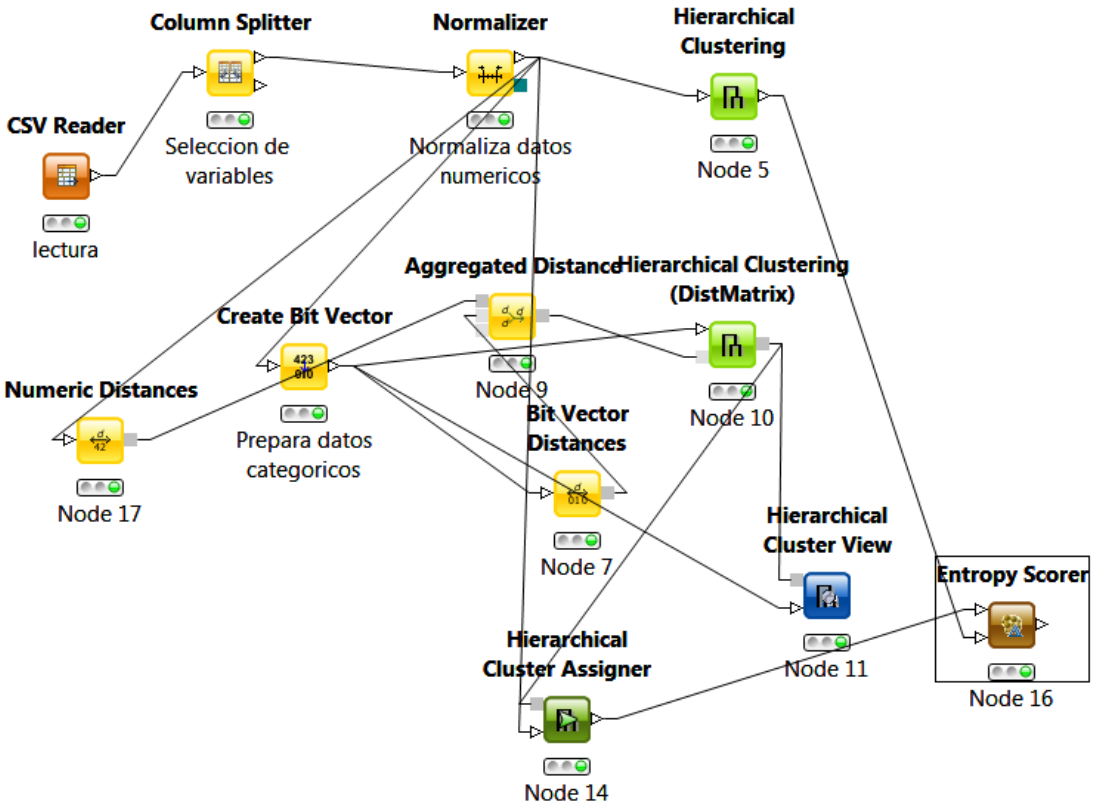
Tratamiento de las distancias en clusering jerarquico en Knime

Como se indicó en teoría existen tipos de variables para las cuales las distancias tipo euclídea no son las más adecuadas, (variables binarias, categóricas etc.), en Knime hay una extensión que permite tratar de forma independiente conjuntos de variables para trabajar con distancias adecuadas a ellas y luego combinar a las distancias obtenidas para obtener una única función de distancia.

Los módulos a utilizar son *data manipulation > distance calculation>distance function y mining>clustering>hieralchical clustering(dist. Matrix)* y luego los de visualización y asignación asociados. Hay que tener en cuenta que en este caso la preparación es mucho mayor ya que hay que entrar con las funciones de distancias de forma adecuada.
Aquí puede verse un ejemplo donde:

- Se leen datos
- Se filtran las columnas que interesan
- Se divide el fichero en dos
    - Uno de datos numéricos que se normalizan y para los que se define una distancia numérica
    - Otro de datos binarios para los que se crea un vector de 0 y 1 y se define una distancia tipo bit.
- Se agregan las distancias y realiza el proceso de clustering jerárquico

Este ejemplo se aplica al fichero demo_cs.csv, tomado de demo.sav mediante transformación que incluye varias variables binarias.



Ejercicio:

Rehacer el flujo con bankloan utilizando dos funciones de distancia distintas, una para datos personales y otra para datos de situación económica

Clustering en SPSS

K-Means: iris

Ahora, lea los datos del fichero desde SPSS. Para ello, seleccione *Archivo > Leer datos de texto* e indique que en el fichero de datos utilizará ; como delimitadores para separar los valores de las distintas variables:

> IMPORTANTE: SPSS no normaliza las variables antes de ejecutar k-Means, por lo que habría que crear las variables normalizadas (*Analizar > Estadísticos Descriptivos > Descriptivos > Guardar valores tipificados como variables*) y usar k-Means sobre las estas nuevas variables, no sobre las originales.

A continuación, seleccione *Analizar > Clasificar > Conglomerado de K-medias.* Asegúrese de incluir las dos variables de nuestro conjunto de datos y escoja 3 como número de "conglomerados" (el valor del parámetro k). Lance el "análisis de conglomerados" y, una vez ejecutado el algoritmo de las K Medias, aparecerán los centroides de los agrupamientos. Cuando trabaje con variables normalizados, debe tener

en cuenta que los centroides también estarán normalizados. Para averiguar cuáles serían los valores originales (no normalizados) correspondientes a los centroides, debe deshacer la tipificación. Es decir, debe multiplicar el valor normalizado por la desviación típica y sumar la media aritmética (para cada una de las variables) .

Clustering jerárquico Iris

Seguiremos trabajando con el fichero que tenemos abierto

Para lanzar un algoritmo de clustering jerárquico en SPSS, seleccione
Analizar > *Clasificar* > *Conglomerados Jerárquicos*.

Como puede verse se pueden obtener el gráfico del dendrograma o un gráfico "de tempanos " que nos dá para cada punto cuando se une a un cluster. Tambien se pueden elegir distintas opciones tales como, el método, y la función de distancia, pero solo la euclídea si se desan mezclar variables numericas con variables binarias. En el caso de que se usan solo variables binarias es posible trabajar con ellas tomando distancias propias de las mismas.

Clustering en R/RStudio

En R el clustering o agrupamiento es uno de los problemas más estudiados. Si revisamos el resumen ***RDataMining-reference-card*** puede observarse que se recogen todos los enfoques del agrupamiento así como numerosas variantes distintas técnicas. En comparación con otros problemas no supervisados como reglas de asociación, o supervisados como clasificación, es con mucho el enfoque más elaborado.
También en ***RDataMining-reference-card*** se referencian las librerías de R dedicadas al agrupamiento (más de 20). Por nuestra parte nos centraremos en las más utilizadas que son las que proporcionan los métodos más conocidos ests librerías son:

- ***stats*** que recoge técnicas de cluster jerarquico (calculo, análisis del dendrograma, etc..), y de las k-medias, junto con funciones para el calculo de distancias
- ***cluster*** que recoge los enfoques de L. Kaufman, P, J. Rousseeuw incluyendo métodos como DIANA, AGNES, K-medias difusas, CLARA etc. Tambien permite obtener y representar el coeficiente de silueta,
- ***fpc*** donde se encuentra el dbscan junto con algunas interesantes métodos para el cálculo de la bondad.

Los distintos ejemplos de uso que se adjuntan del directorio ***Scripts agrupamiento.***

1. Los siguientes scripts:
    a. ***iris-jerarquico-nomalizado.R***
    b. ***iris-kmedias-normalizadas.R***
    c. ***iris-medoides-normalizado.R***
    d. ***iris-medoides-normalizado-valor óptimo de grupos.R***
    e. ***iris-dbscan.R***
    f. ***iris-fuzzykmeans-normalizado.R***

    Están orientados a trabajar con distancia euclídea y los datos de Iris. El esquema general de estos scritps es:

- normalizar previamente los datos,
- aplicar la técnica correspondiente
- Analizar la bondad del resultado mediante gráficas, el coeficiente de silueta y algunas medidas estadísticas adicionales

2. Se proponen también sripts que utilizan otras medidas de distancia. Para ello se utiliza un data set alternativo, el fichero ***demo_cs-R.csv*** que mezcla variables numéricas y binarias. Los scripts que se incluyen son:
   a. ***distancias-jerarquico.R***
   b. ***distancias-k-medias.R***
   c. ***distancias-k-medoides.R***

   El esquema general de estos scripts es:
   - Separar los datos por tipos de variables
   - Calcular las distancias para cada grupo de variables (numéricas y binarias)
   - Agregar distancias y aplicar la técnica correspondiente a la matriz de distancias agregada.
   - Evaluar la bondad de los resultados

Ejercicio:
Rehacer los scripts con bankloan:

1. **Trabajando con una única función de distancia**
2. **Utilizando dos funciones de distancia distintas, una para datos personales y otra para datos de situación económica**
3. **Probar a realizar una selección de variables, agrupando de una parte por datos personales y por otra por datos de situación económica. ¿Mejoran los resultados?¿Se pueden interpretar los grupos?**

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)
- L. Kaufman, P, J. Rousseeuw Finding Groups in Data An Introduction to Cluster Analysis (2005)