

## PRÁCTICA 4

### 1. Spotfire

Spotfire es una herramienta de visualización de la información desarrollada por la empresa de desarrollo TIBCO.

Podemos descargarnos la herramienta de <http://spotfire.tibco.com/>

Es necesario pedir una licencia de 30 días de uso registrándose. Si queréis registraros solo 1 y os juntáis en un ordenador.

a) Abre la base de datos breastCancer.csv.

Ésta contiene 569 registros, cada uno describiendo la imagen de microscopio de masa extraída por una aguja fina pinchada en el pecho de una mujer.

Cada registro se describe con 10 métricas que describen la imagen:

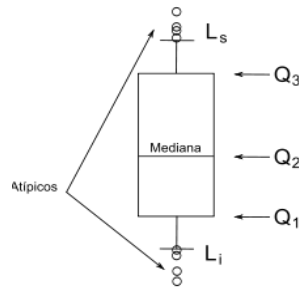
1. Clump Thickness	1 - 10
2. Uniformity of Cell Size	1 - 10
3. Uniformity of Cell Shape	1 - 10
4. Marginal Adhesion	1 - 10
5. Single Epithelial Cell Size	1 - 10
6. Bare Nuclei	1 - 10
7. Bland Chromatin	1 - 10
8. Normal Nucleoli	1 - 10
9. Mitoses	1 - 10
10. Class:	B benign, M malignant

b) Por defecto aparece una gráfica de dispersión, con los ejes X e Y mostrando las 2 primeras variables, y el color indicando la tercera variable de la base de datos.

- Arréglalo para que el color indique el tipo de tumor, ¿ves clusters en el tipo de tumor según los valores de ClumpThickness y UniformSize?
- Está claro que no vemos 569 casos. Con el ratón sobre el espacio blanco junto al gráfico, presione en el símbolo tic de 'correcto', y en 'Appearance' aumenta el jitter.
- Busca más formaciones de clusters cambiando las variables de los ejes X e Y.

c) Ahora añade una matriz de dispersión en 3D.

- a. Vuelve a indicar que el color se relacione con el tipo de tumor, e interacciona con el gráfico utilizando las herramientas de zoom y giro para conseguir una perspectiva desde la que se puedan observar con claridad los ejes y los casos.
  - b. ¿Puedes ver que los casos Benignos están todos en los valores pequeños de las variables ClumpThickness, UniformSize y UniformShape? Busca más conjuntos de 3 variables que consigan patrones identificables.
  - c. Selecciona algún caso o conjunto de casos que te interesen. Fíjate cómo se selecciona también automáticamente en la otra gráfica.
- d) Elimina las gráficas y ahora añade un histograma que indique cuántos registros hay para cada posible valor de ClumpThickness.
  - a. Vuelve a añadir el color. Fíjate que te divide cada posible valor en 2 barras: los casos benignos y los malignos.
  - b. En la parte derecha de la pantalla puedes seleccionar qué registros mostrar según el valor de alguna variable. Elimina los casos que tengan un tamaño de célula (CellSize) pequeño (de 1 a 4). ¿Qué barras del histograma reducen más su tamaño? ¿Podemos pensar que los casos con poco grosor (ClumpThickness) tienen además un tamaño pequeño? Demuestra esto con una Gráfica de Líneas contrastando estas 2 variables (eje Y la media de ClumpThickness de los casos con el valor correspondiente de CellSize del eje X).
  - c) Une las barreras en Appearance → Layout → stacked bars
- e) Como en XmdvTool, Spotfire te permite realizar TreeMaps y Paralelas Coordinadas. Haz un TreeMap con 2 niveles de anidación con las 2 variables que quieras, más una de color.
- f) El tipo de gráfico Heat Map es parecido a un TableLens, solo que no amplifica las filas seleccionadas. Haz un HeatMap seleccionando que muestre todas las variables. Ordena las filas por tipo de tumor haciendo click sobre el nombre de la Columna. ¿Qué valores suele tomar las variables para tumores Benignos y Malignos?
- g) La gráfica Box Plot es la gráfica que vimos en el Tema 2 con nombre de Gráfico de Caja y Bigotes, que nos sirve para ver la distribución de los datos. Recuerda la siguiente estructura:



- Los bigotes (barras verticales) representan los límites superior e inferior de los datos, excepto aquellos casos cuyo valor supere:
  - $1.5 \times \text{IQR}$  : atípicos
  - $3 \times \text{IQR}$ : outliers
  
- a. Crea este gráfico para la variable ClumpThickness en el eje Y, y en el eje X los 2 posibles tumores. ¿En torno a qué valores de la mediana está ClumThickness para cada tipo de tumor? ¿En qué rango de valores se mueve la variable para cada tipo de tumor?
  
- b. Para no olvidar tus conclusiones, o para compartirlas con alguien, puedes escribirlas pulsando el botón 'New Text Area'.
  
- h) Spotfire también permite realizar tests estadísticos muy sencillamente, para ver si existe relación entre variables, presentando distintos tests según el tipo de variables a comparar. En Tools→Data Relationships:
  - a. Comprueba si hay relación estadística entre cada variable numérica y la variable Tumor. (se condiera relación existente para  $p\text{-values} < 0.05$ ).
  - b. Comprueba que hay relación entre la variable CellSize y ClumpThickness como se intuyó en el punto 2.d.b.