TEMA 3. ANÁLISIS DE LA DISTRIBUCIÓN DE LOS DATOS

Contenidos

I. Introducción
II. Distribuciones marginales
    a. Data Table
    b. Feature Statistics
    c. Distributions
    d. Box Plot
III. Distribuciones conjuntas
    a. Scatter Plot
    b. Line Plot
    c. Correlations, Distances, Distance Matrix y Distance Map
    d. Heat Map
IV. Actividad

I. Introducción

Tipos de análisis de la distribución de los datos

- **Univariante:** distribuciones marginales
- **Multivariante**: distribuciones conjuntas (condicionales)
  - Caso particular: análisis bivariante (dos variables)

Representación del análisis

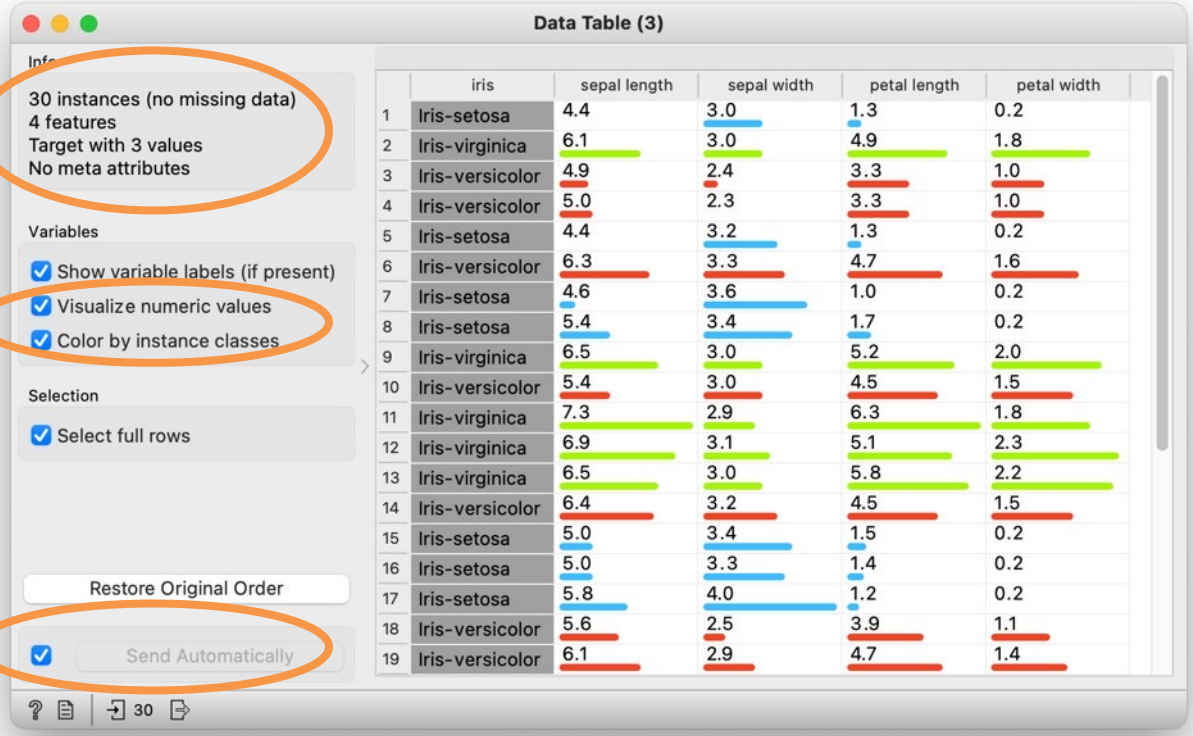
- **Análisis cuantitativo:** tablas numéricas (estadísticas, …)
- **Visualización gráfica**: gráficos de diferentes tipos

II. Distribuciones marginales

Data Table

Data Table

- Objetivo: observar la distribución de los valores de los atributos **sin resumir** (previa ordenación por columna).
- Información del tamaño del dataset, el número y los tipos de atributos.
- Permite visualizar una barra horizontal de rango en atributos numéricos coloreada en función de una clase discreta.
- Se pueden seleccionar **filas** para comunicarlas con otros nodos.



II. Distribuciones marginales

Feature Statistics

Feature Statistics

- Objetivo: visualizar mediante histogramas.
- Se pueden colorear según el atributo seleccionado. Si es numérico se usará una paleta continua de colores.
- Estadísticos numéricos descriptivos de cada atributo.
  - Centro: Si son atributos numéricos es la media, si son cualitativos es la moda.
  - Dispersión: Si son atributos numéricos es el coeficiente de variación, si son cualitativos es la entropía.
- Se pueden seleccionar **atributos** para comunicarlos con otros nodos.

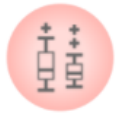
II. Distribuciones marginales

Distributions

- Objetivo: visualizar un histograma para cada atributo.
- Cambiar tamaño de las barras y la distribución.
- Permite ajustar una distribución de probabilidad en función de cada valor de un atributo cualitativo (**Split by**).
- Se pueden seleccionar **filas** para comunicarlas con otros nodos.

II. Distribuciones marginales

Box Plot

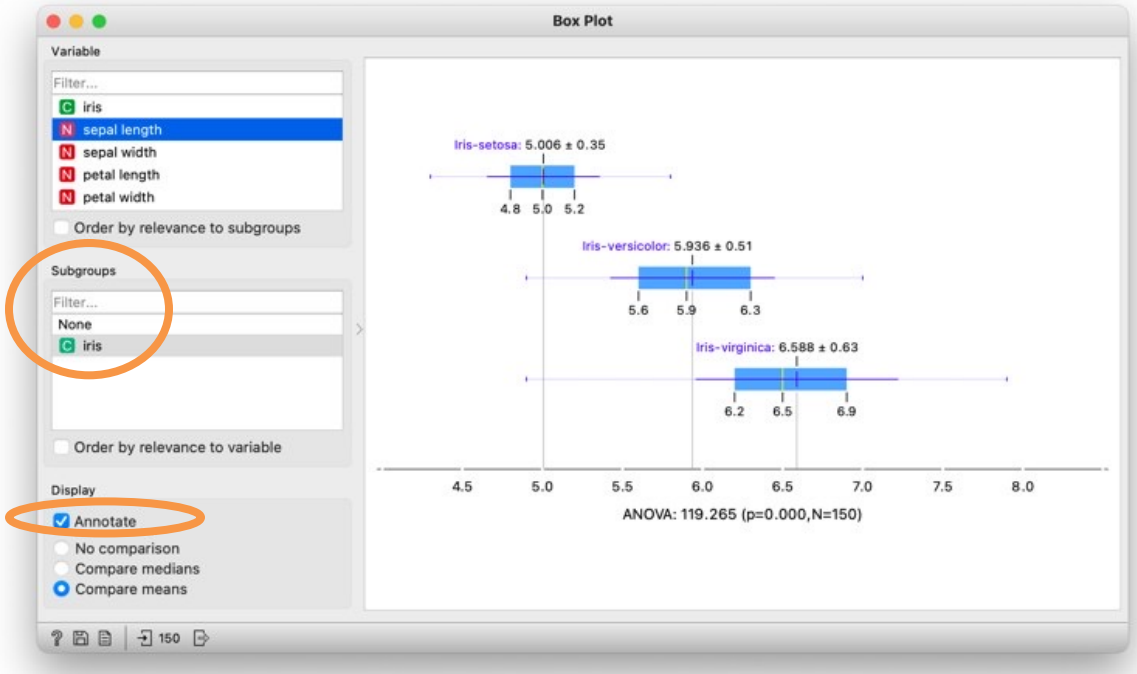


Box Plot

- Objetivo: visualizar los valores mínimo, máximo, media y cuartiles de cada atributo, así como un test de contraste de diferencia de medias con respecto a un atributo (clase) cualitativo.
  - Media: Raya azul oscuro vertical.
  - Mediana: Raya amarilla vertical.
  - Valores entre 1º y 3º cuartil: Zona azul.

- "Annotate": Muestra los valores finales.

- Se pueden seleccionar **filas** para comunicarlas con otros nodos.



III. Distribuciones conjuntas

Scatter Plot

- Objetivo: visualizar las instancias del dataset proyectadas en 2 dimensiones (2 atributos). Útil para análisis *bivariante*.

- Permite encontrar las proyecciones que mejor discriminan la clase con “Find Informative Projections”.

- Es posible visualizar hasta 6 atributos a la vez (color, forma, tamaño y etiqueta).

- Se pueden seleccionar **filas** para comunicarlas con otros nodos.

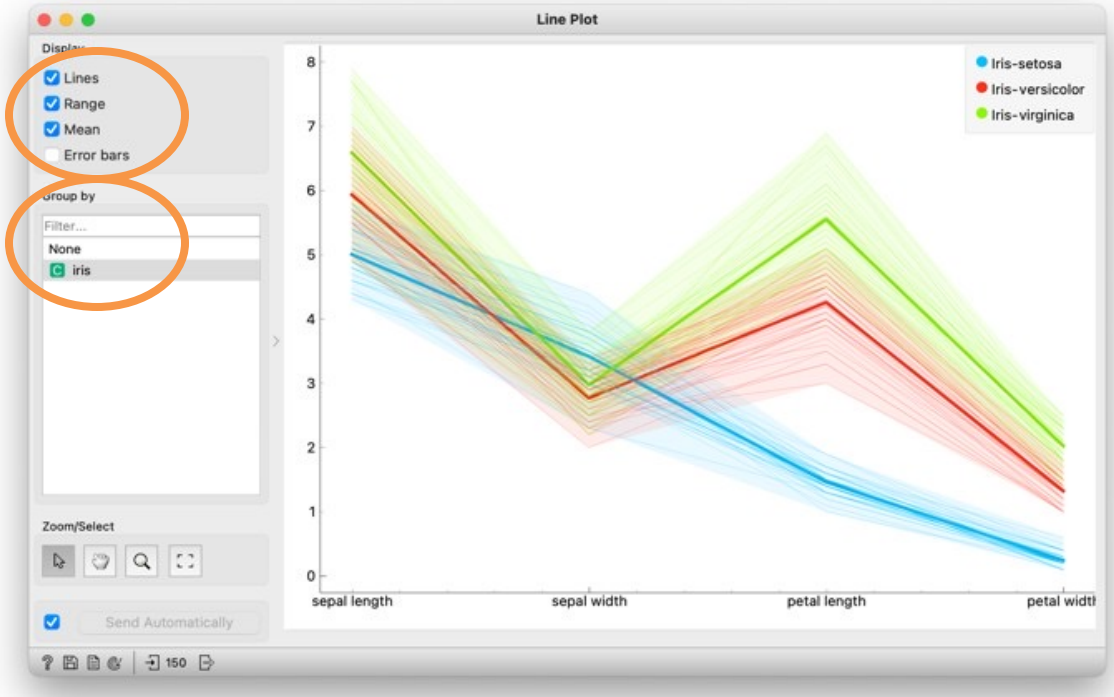
Scatter Plot

III. Distribuciones conjuntas

Line Plot

- Objetivo: visualizar los valores de cada atributo conectados mediante líneas por pertenecer a la misma instancia.
- Permite representar una línea promedio (“Mean”), líneas de percentiles 10% - 90% (“Range”) y barras de desviación típica (“Error”).
- “Group by” para unir instancias según un atributo cualitativo.
- Se pueden seleccionar **filas** para comunicarlas con otros nodos.

Line Plot



III. Distribuciones conjuntas

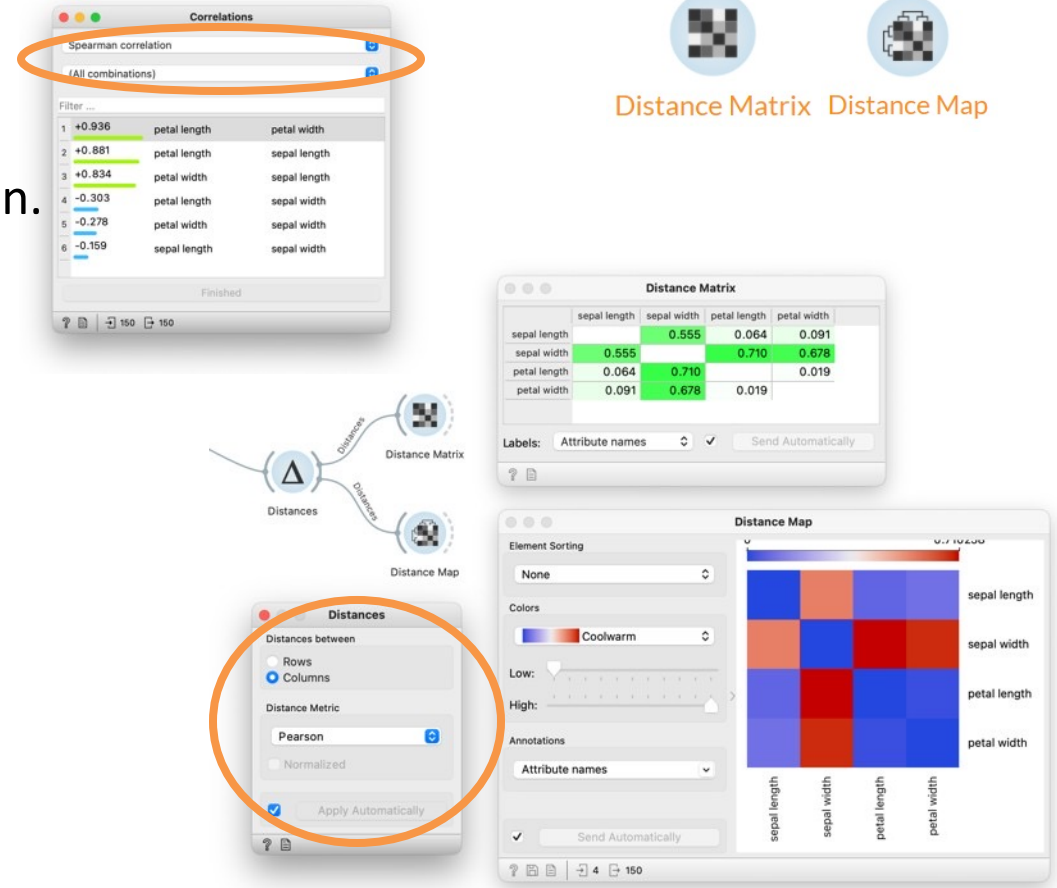
Correlations, Distances, Distance Matrix y Distance Map

- Objetivo: calcular las correlaciones entre pares de atributos.

- El nodo **Correlations** permite tanto Pearson como Spearman. Devuelve tanto la tabla de correlaciones como el par de atributos seleccionado.

- El nodo **Distances** computa distancias entre filas o entre columnas (normalizando) y permite más tipos de correlaciones (funciones de distancia).

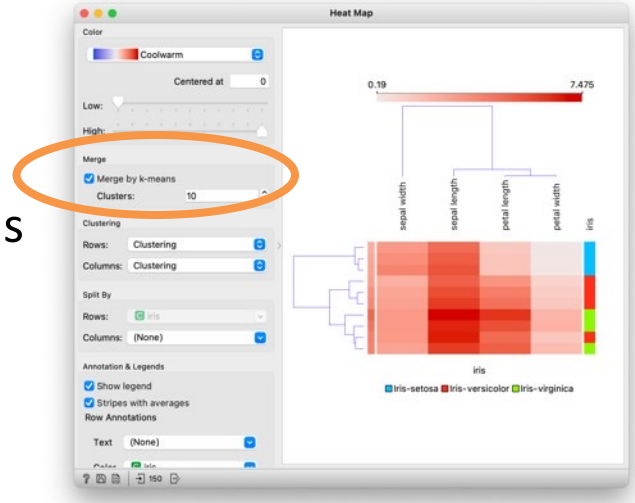
- Los nodos **Distance Matrix** y **Distance Map** permiten visualizar las matrices de distancias.

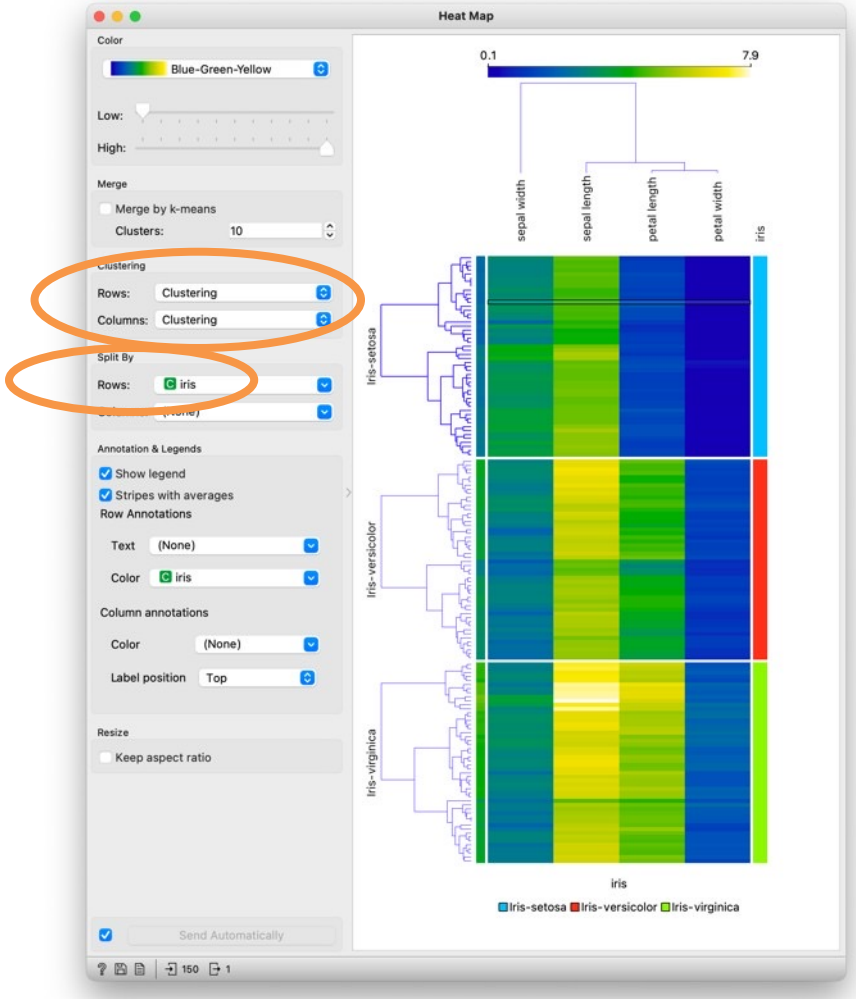


III. Distribuciones conjuntas

Heat Map

- Objetivo: visualizar todas las instancias del dataset con todos sus atributos, coloreadas por una paleta de colores común y agrupadas según uno o dos atributos cualitativos (o la clase).
- También puede hacer las agrupaciones mediante clustering.
- Permite observar dendrogramas (clustering jerárquico), tanto por filas como por columnas.





IV. Actividad de análisis de distribuciones

1. Cargue el dataset **Hearth Disease** mediante el nodo **Datasets**
2. Utilice los nodos estudiados para analizar las distribuciones marginales.
   a) Decida que tipos de visualización son más informativos.
   b) ¿Qué conclusiones puede extraer a partir de la visualización?
3. Utilice los nodos estudiados para analizar las distribuciones conjuntas.
   a) Escoja los tipos de visualización más apropiados para este dataset.
   b) ¿Qué conclusiones puede extraer?