

K-MEANS (ORANGE DATAMINING).

. Volvemos a arrastrar al lienzo un modulo file. Y luego cargamos un archivo .csv en este caso utilizaremos un origen llamado Drugs.

The screenshot shows the 'File' widget in the Orange Data Mining software. The 'File' radio button is selected, and the file 'Drugs.csv' is loaded. The 'Info' section displays the dataset statistics: 600 instances, 7 features (no missing values), no target variable, and 0 meta attributes. The 'Columns' section shows a table with 7 columns: Name, Type, Role, and Values. The 'Values' column is highlighted. The 'Reset' and 'Apply' buttons are visible at the bottom right.

File: Drugs.csv

URL:

Info

600 instance(s)
7 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	Age	N numeric	feature	
2	Sex	C categorical	feature	F, M
3	BP	C categorical	feature	HIGH, LOW, NORMAL
4	Cholesterol	C categorical	feature	HIGH, NORMAL
5	Na	N numeric	feature	
6	K	N numeric	feature	
7	Drug	C categorical	feature	drugA, drugB, drugC, drugX, drugY

Browse documentation datasets

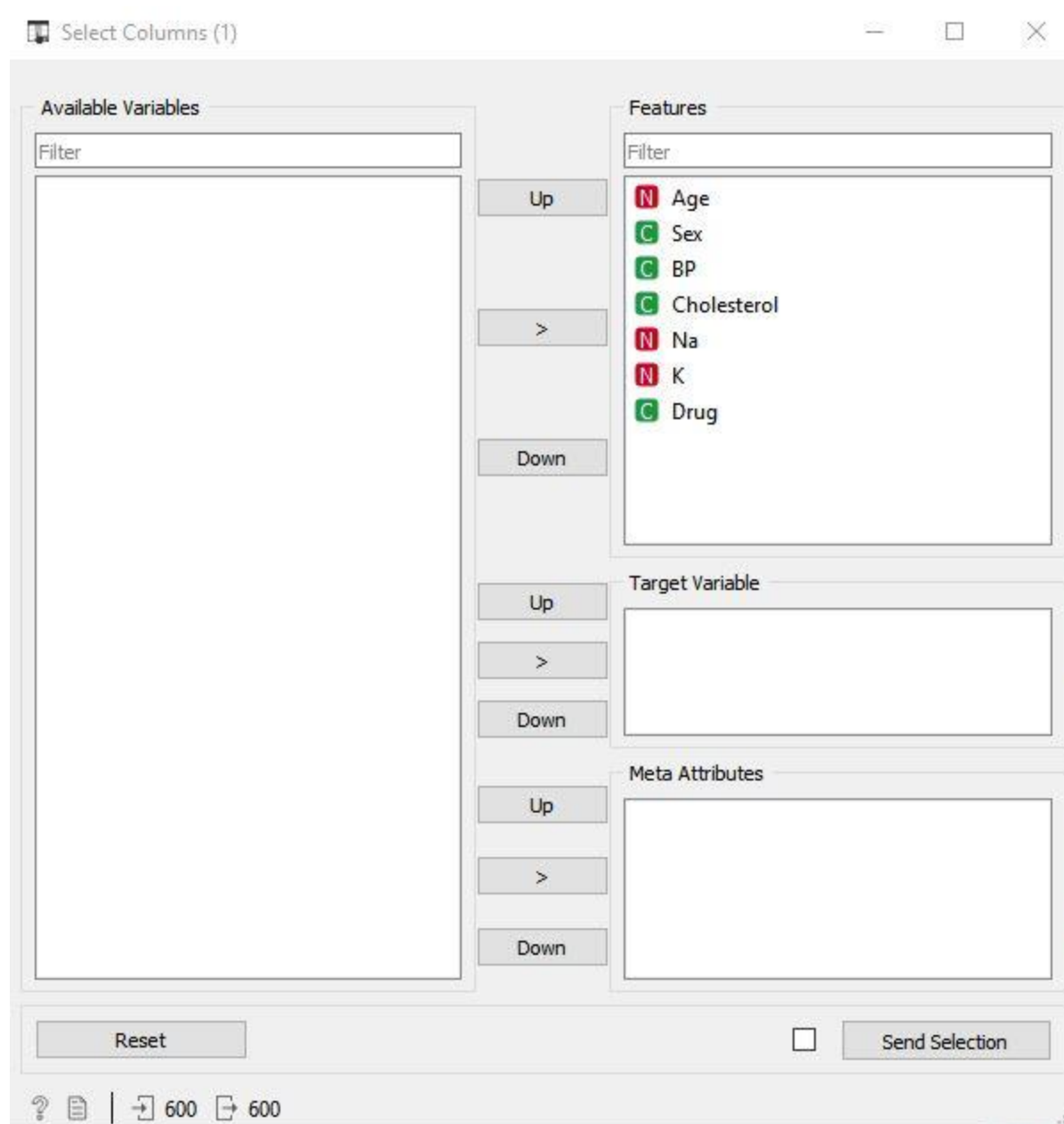
Reset Apply

? | 600

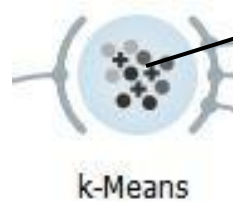
. Tomamos también un data table para poder ver mas detalladamente los datos que contiene el origen.

Drugs							
Info							
600 instances (no missing data)							
7 features							
No target variable.							
No meta attributes							
Variables							
<input checked="" type="checkbox"/> Show variable labels (if present)							
<input type="checkbox"/> Visualize numeric values							
<input checked="" type="checkbox"/> Color by instance classes							
Selection							
<input checked="" type="checkbox"/> Select full rows							
Restore Original Order							
<input checked="" type="checkbox"/> Send Automatically							
Age	Sex	BP	Cholesterol	Na	K	Drug	
1	25 F	HIGH	HIGH	0.675996	0.074834	drugA	
2	17 F	HIGH	HIGH	0.539756	0.030081	drugY	
3	23 M	LOW	NORMAL	0.556453	0.03618	drugY	
4	24 M	NORMAL	NORMAL	0.845236	0.055498	drugY	
5	74 F	LOW	HIGH	0.849624	0.076902	drugC	
6	40 F	NORMAL	HIGH	0.67683	0.049634	drugX	
7	32 F	HIGH	HIGH	0.581664	0.024803	drugY	
8	70 M	LOW	HIGH	0.716359	0.036936	drugY	
9	64 M	HIGH	NORMAL	0.640789	0.078302	drugB	
10	45 M	HIGH	HIGH	0.664105	0.047819	drugA	
11	33 F	LOW	NORMAL	0.821805	0.027674	drugY	
12	74 F	LOW	NORMAL	0.772225	0.04794	drugY	
13	73 M	HIGH	NORMAL	0.792131	0.062171	drugB	
14	38 F	LOW	HIGH	0.794318	0.051825	drugY	
15	72 F	HIGH	NORMAL	0.533558	0.021289	drugY	
16	27 F	HIGH	NORMAL	0.555064	0.04665	drugA	
17	62 M	HIGH	NORMAL	0.51015	0.071463	drugB	
18	72 M	HIGH	NORMAL	0.819483	0.073802	drugB	
19	19 M	HIGH	NORMAL	0.552701	0.032598	drugY	
20	28 M	HIGH	HIGH	0.584039	0.068375	drugA	
21	54 M	NORMAL	NORMAL	0.605629	0.076527	drugX	
22	37 F	NORMAL	HIGH	0.743515	0.021146	drugY	
23	33 F	NORMAL	NORMAL	0.815439	0.064816	drugX	
24	73 F	LOW	HIGH	0.522929	0.02639	drugY	
25	41 F	NORMAL	HIGH	0.796671	0.075183	drugX	
26	41 M	LOW	HIGH	0.845565	0.054674	drugY	
27	46 F	NORMAL	NORMAL	0.598635	0.032928	drugY	
28	59 F	NORMAL	HIGH	0.658724	0.022489	drugY	
29	71 F	LOW	HIGH	0.56486	0.065305	drugC	
30	49 M	HIGH	HIGH	0.56226	0.0309	drugY	
31	17 M	HIGH	NORMAL	0.788461	0.068818	drugA	
32	22 M	LOW	NORMAL	0.803766	0.020025	drugY	
33	69 M	HIGH	HIGH	0.744017	0.030104	drugY	
34	48 F	LOW	NORMAL	0.701044	0.024175	drugY	

. . Cuando tengamos un select column ya al haberlo unido desde el origen de datos, en este caso solo tomaremos todos los campos de la tabla y no será necesario tener una variable objetivo, ya que en este caso que se usará K-means leerá todo lo que esta en el origen y en el módulo de select column.



. Luego unimos desde select column hasta el módulo de K-means. Para poder agregar este modulo que esta en oculto, solo arrastramos la línea para conectar, y nos desplegara un menú con todos los modulos y buscamos el que dice K-means. Luego damos click y se nos mostrara un formulario donde podremos configurar el módulo de K-means. En este caso damos en la opción from y de los datos ponemos de 10 a 11, y las demás opciones solo las dejamos por defecto.



k-Means

Number of Clusters

☐ Fixed: 6

☒ From 10 to 11

Preprocessing

☒ Normalize columns

Initialization

Initialize with KMeans++

Re-runs: 10

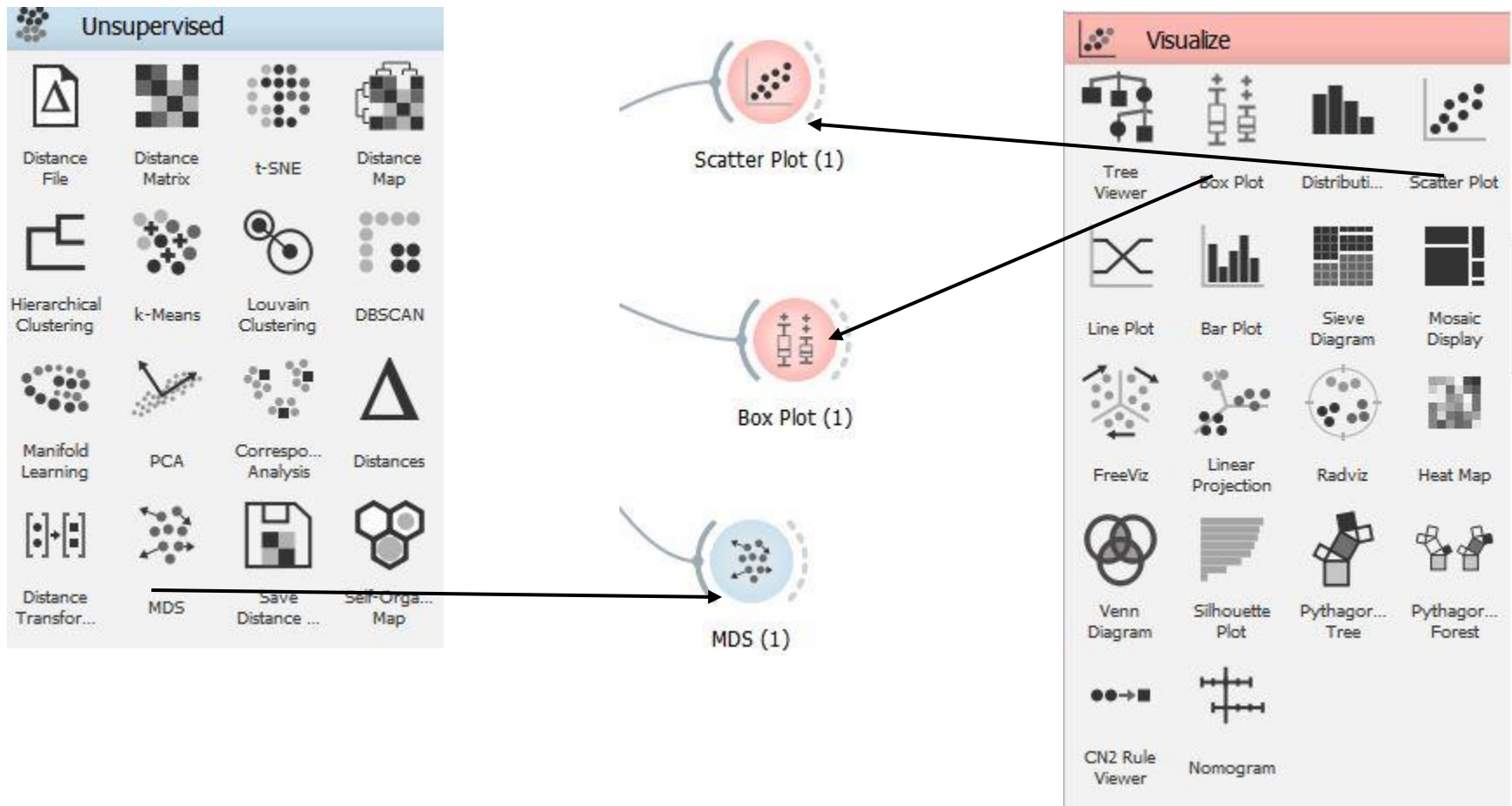
Maximum iterations: 300

Apply

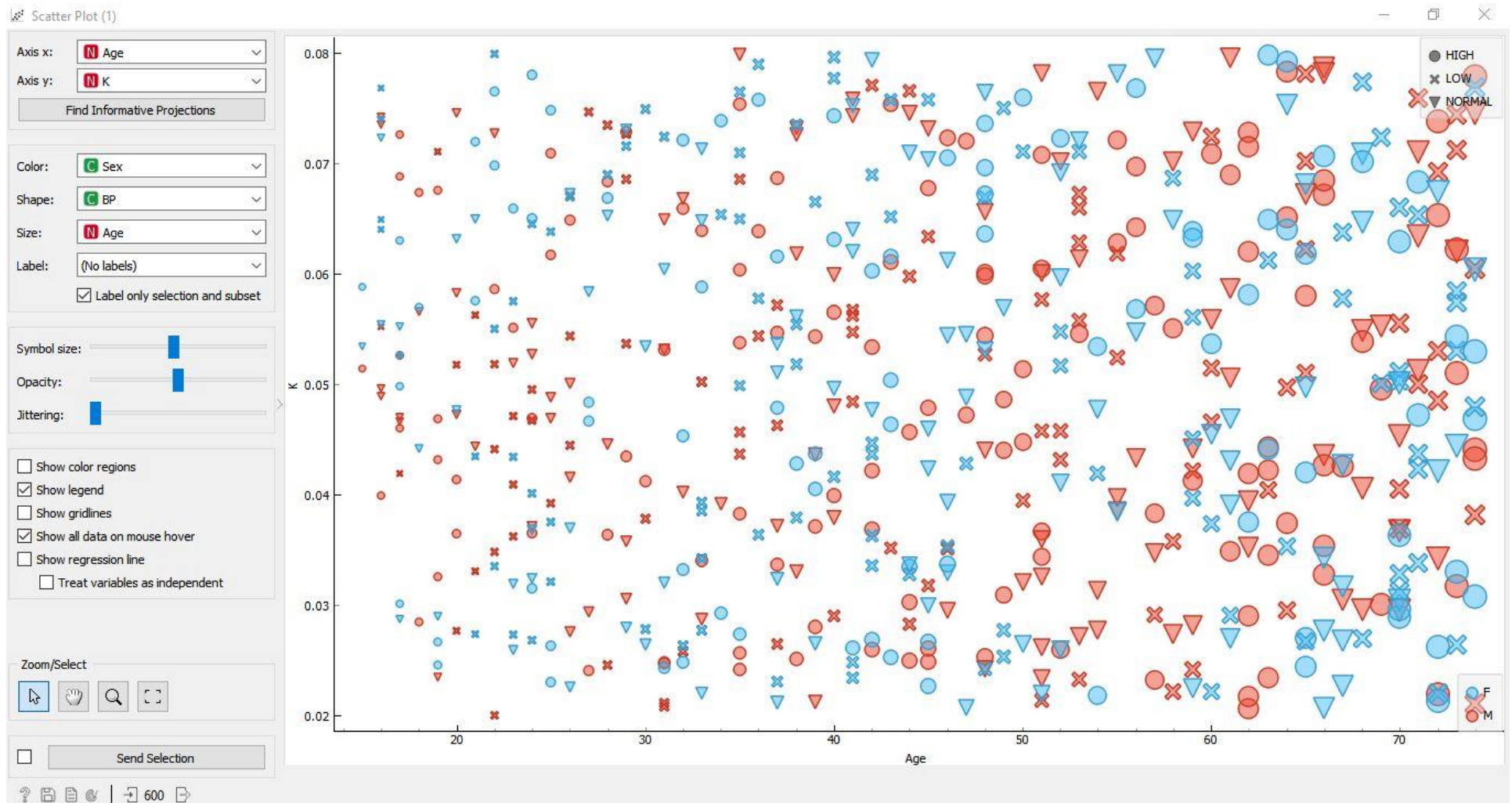
Silhouette Scores

10	0.160
11	0.175

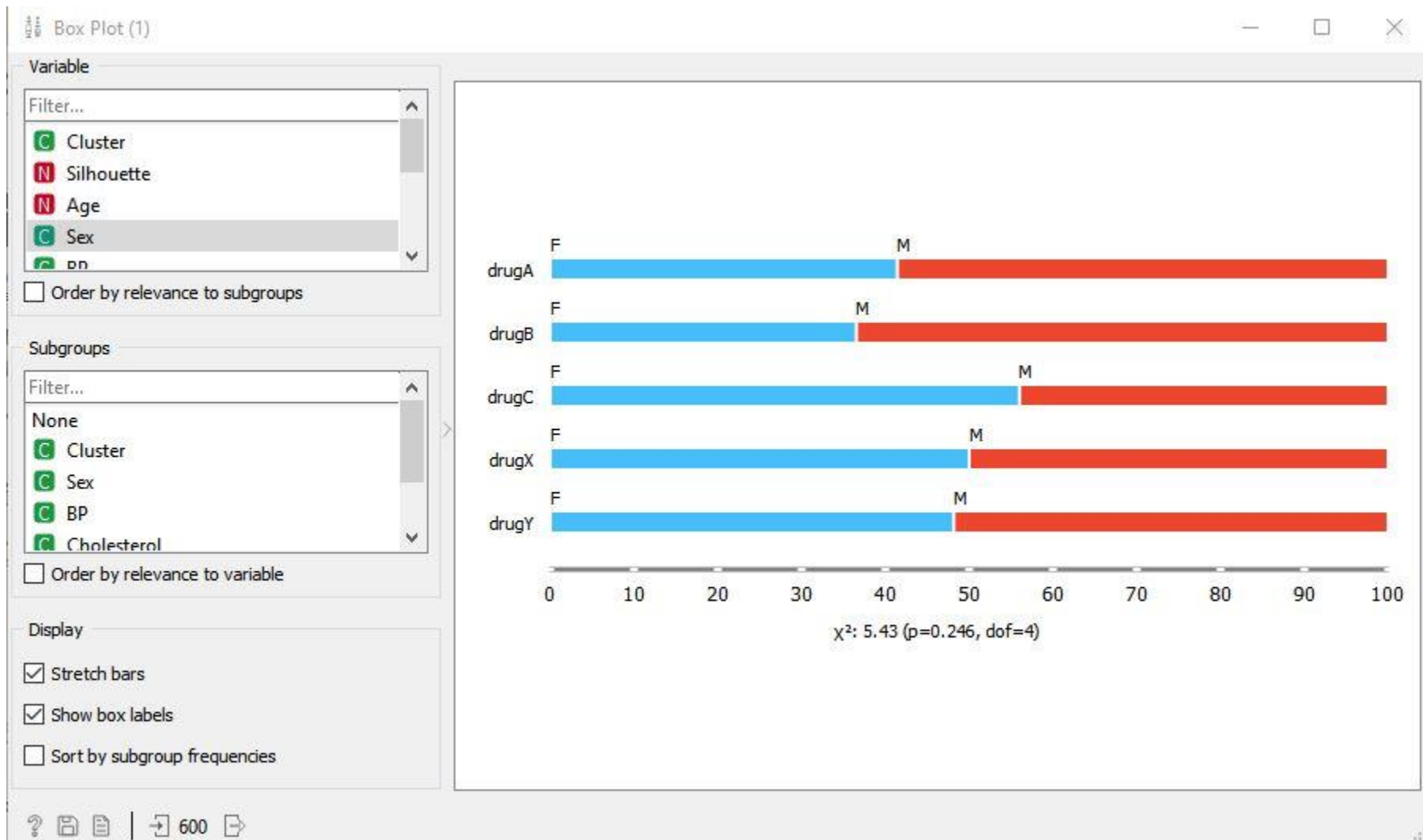
. Ahora vamos al panel de Unsupervised y Visualize. Elegimos los módulos de visualización de Scatter Plot, Box Plot y MDS. ¿Por qué de estos módulos de visualización? Ya que al conectarlos a K-means ya que este algoritmo trabaja con este tipo de visualización de tipo cluster. Esto nos permite ver más gráficamente en un plano cartesiano la data. Y es donde el algoritmo va organizando bien los datos y los filtros que vamos asignándole. Así que estos 3 tipos de visualización los conectamos directamente al módulo de K-means.



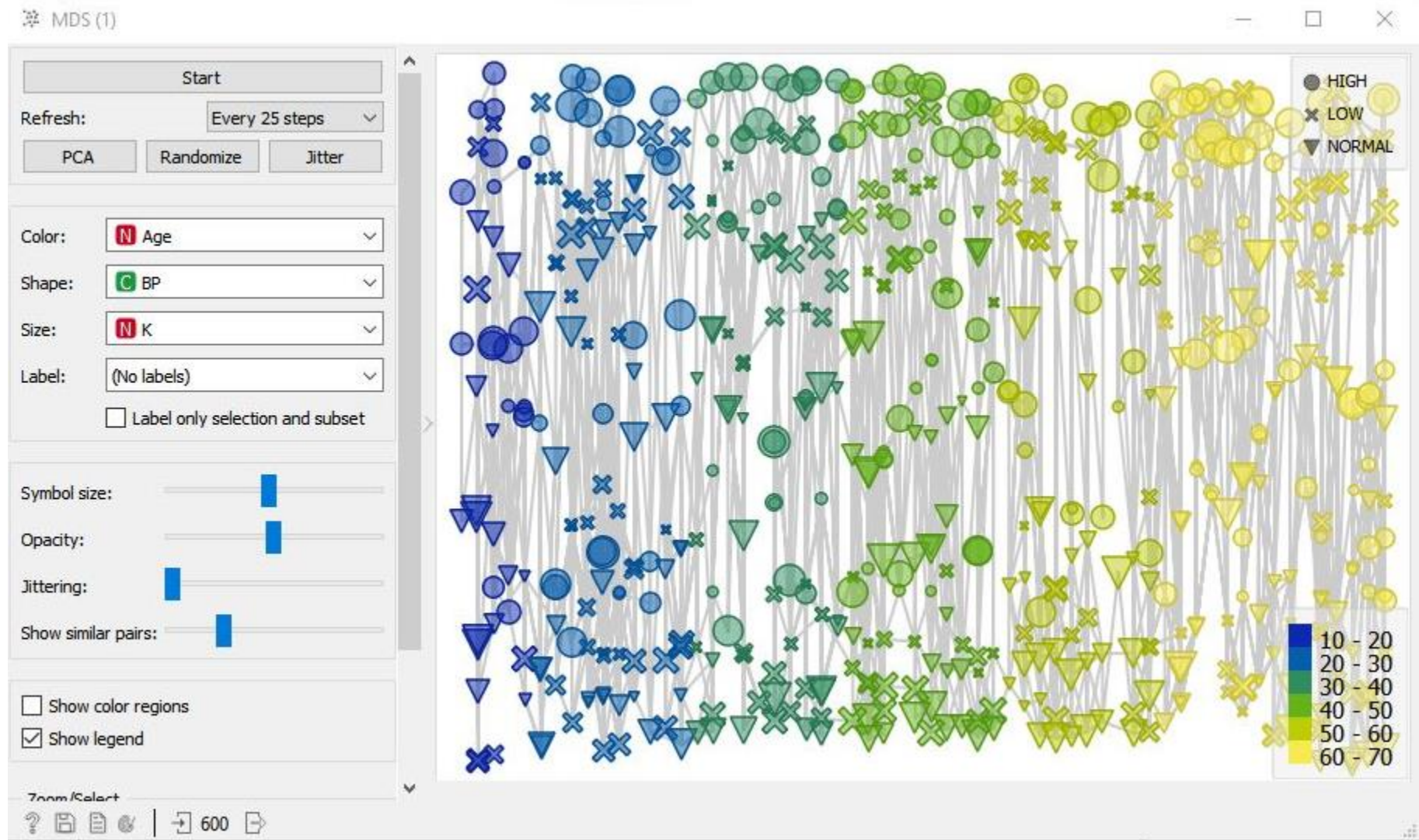
. Este seria el scale plot donde podemos ver que en el eje X esta designado la edad y eje Y K (ketamina). Donde da paso al filtrado de datos que nos permite visualizar mejor el análisis del algoritmo en este caso en color elegimos Sexo en Shape BP (Blood Pressure) y en tamaño la edad, esto nos permite visualizar como K-means va generando los datos y el funcionamiento del algoritmo. Esto quiere decir es la muestra de datos que el algoritmo a generado entre las personas de dicha edad cuanto de presión en la sangre tienen. Y que grado de Ketamina tienen.



. En Box plot se a tomado los datos de el tipo de droga, la edad y Sexo. En este rango podemos denotar que tipo de droga es mas utilizada por ambos sexos. Y su escala de edades.



. MDS se han tomado los datos de la edad de las personas y su escala de presión de la sangre y es así como los datos se van mostrando y generando en el algoritmo. Y también la escala en este caso es la ketamina.



. Ya como parte opcional, si se requiere nuevamente agregar un modulo de visualización de distribución de datos para verlo gráficamente. Aquí podemos ver como seria la frecuencia del tipo de droga.

