CICLO
II-2020

PROYECTO FASE 3

MATERIA:

Datawarehouse y Mineria de Datos

DOCENTE:

Ing. Carlos Filiberto Alfaro Castro

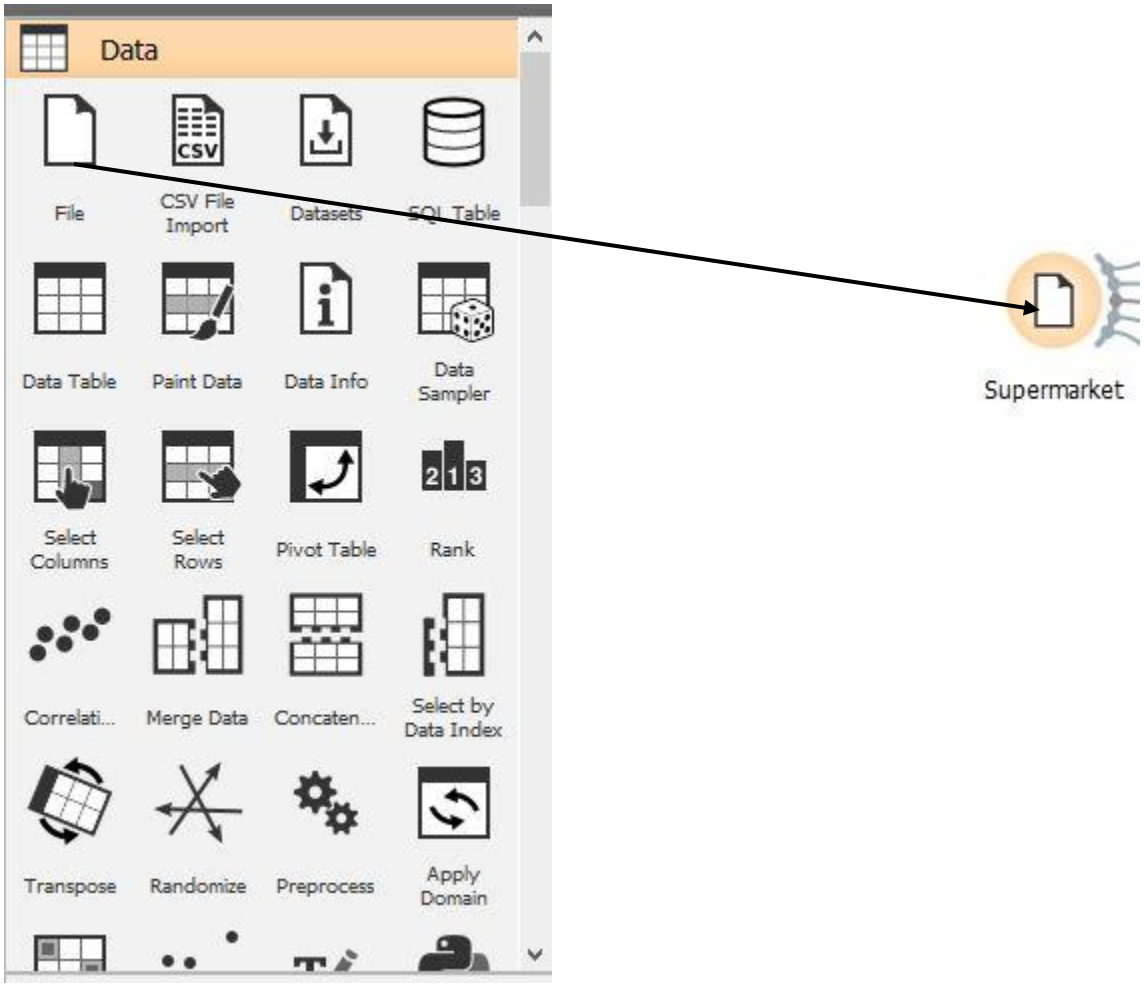
INTEGRANTES:

Carlos Moisés Pérez Cabrera PC190261
Victoria Margarita Sura Jiménez SJ190060
Byron Enrique Aguillón Amaya AA180117
Carlos David Herrera Guardado HG190072
Edgardo Aníbal Zepeda López ZL180073

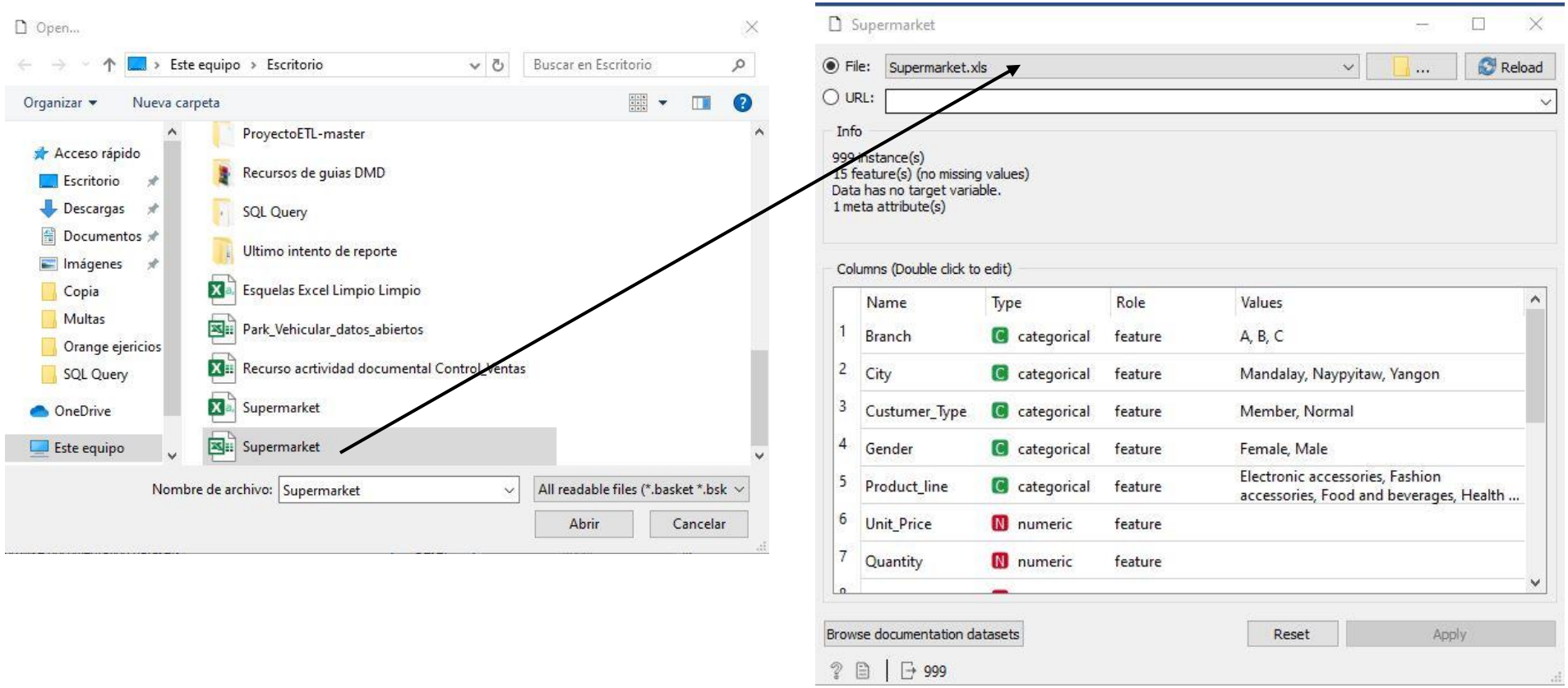
ORANGE

ARBOL DE DECISIÓN (ORANGE DATA MINING).

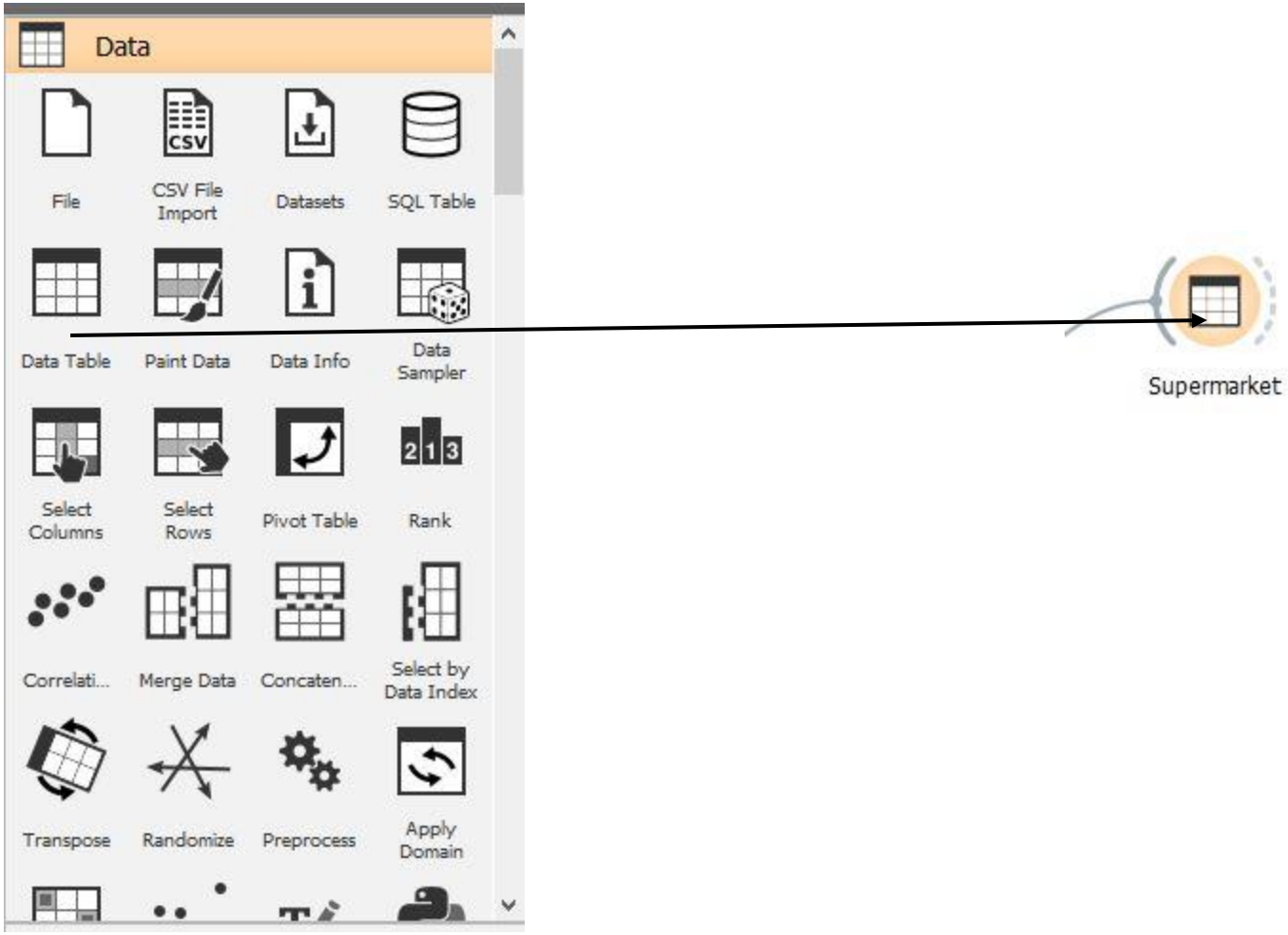
. Se nos muestra un panel que se llama data, arrastramos el complemento fila a lienzo en blanco.



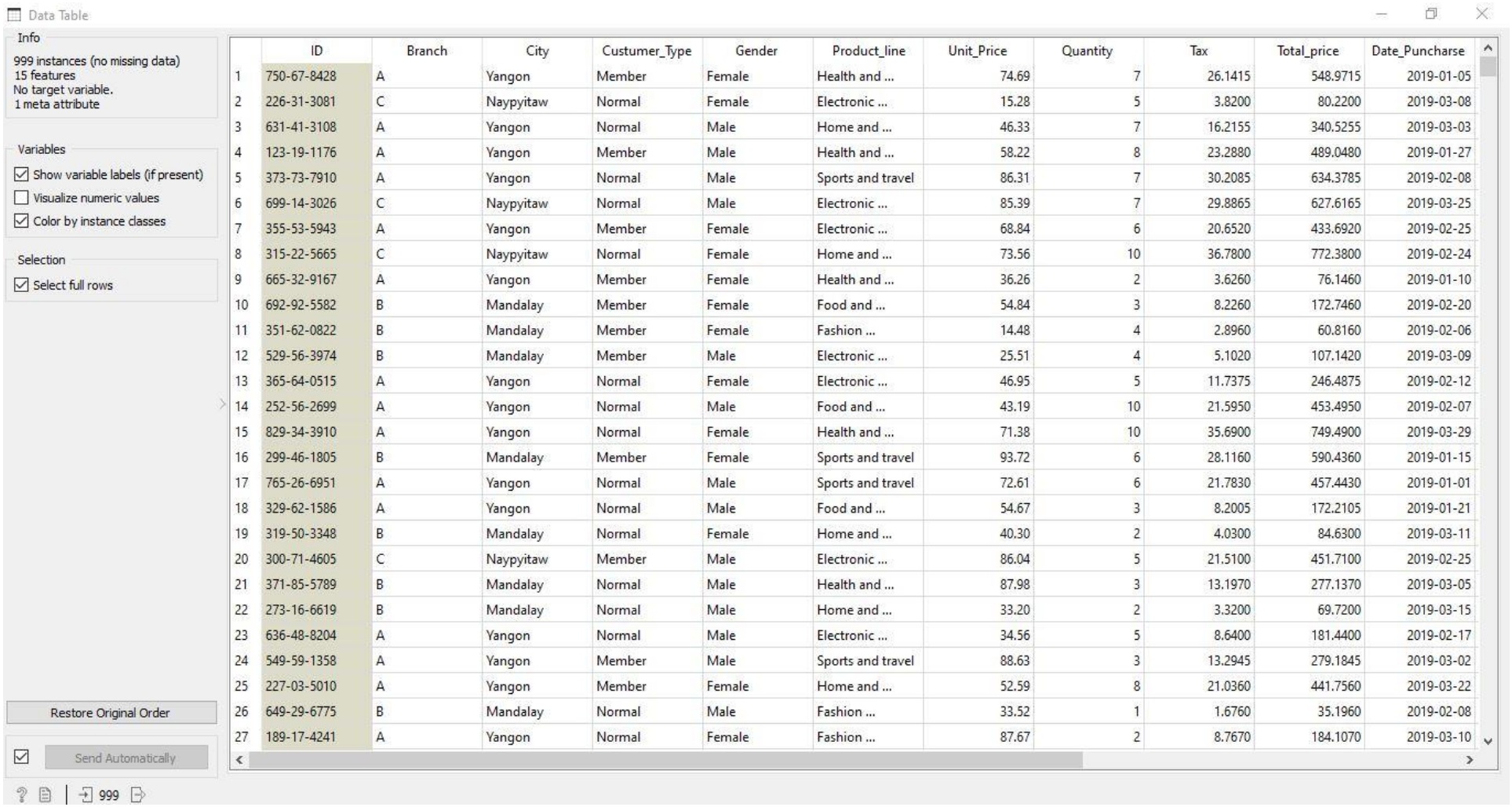
. Dentro del file nosotros buscamos el archivo que queramos usar y en este caso elegimos Supermarket.csv o xls. Y le damos aceptar, nos quedara una distribución en las columnas ya categorizada por el tipo de dato, sin problemas alguno al poder usar el archivo. También aparecerán los roles y los valores que contiene cada columna en la tabla.



. Luego arrastramos un data table, al lienzo y luego lo unimos al archivo que estamos utilizando, así para que nos puedan aparecer mas a detalle lo que se encuentra en el archivo y todos sus datos.

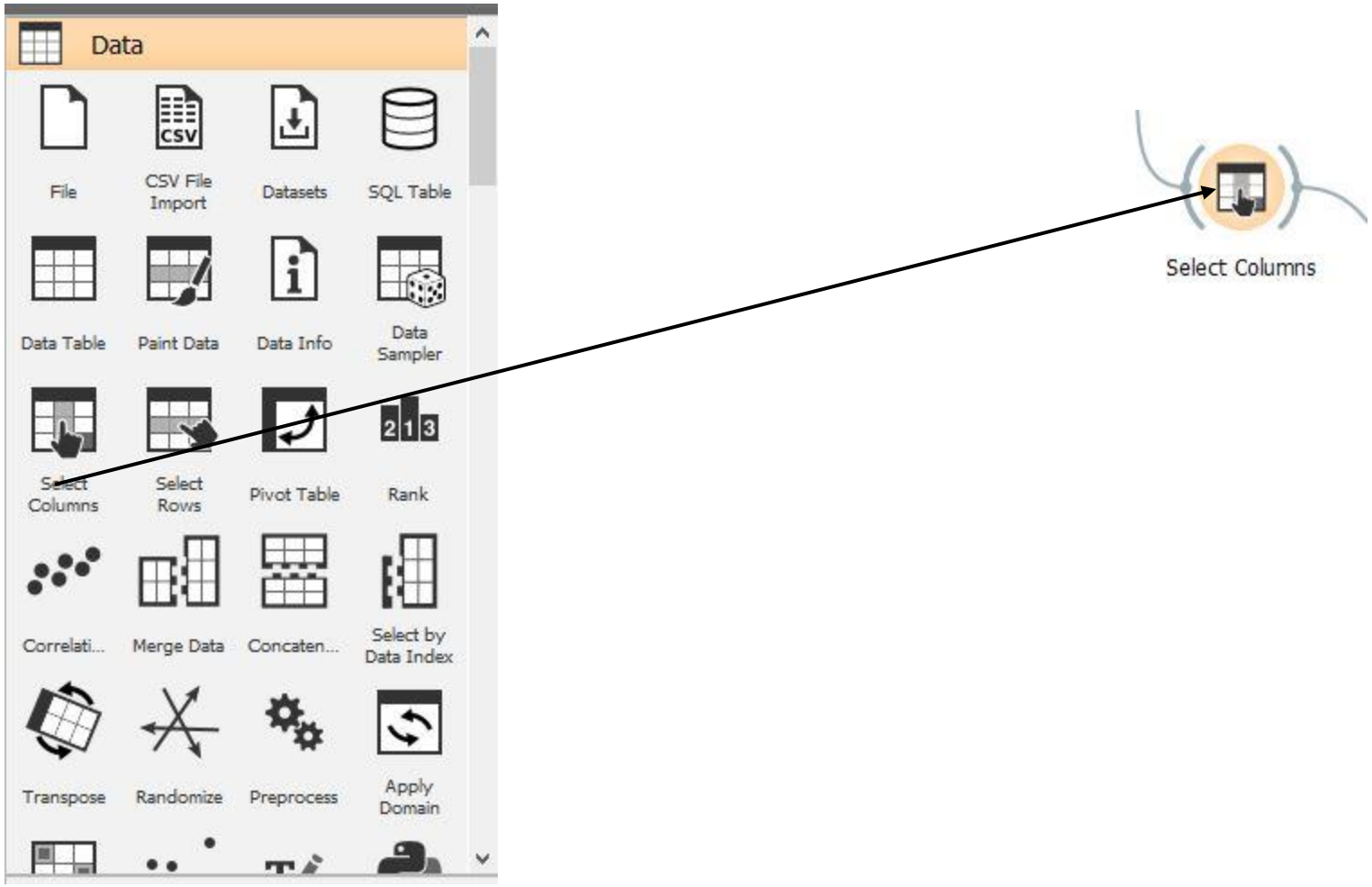


. Y como se puede observar en la tabla toda la información que contiene el archivo, donde podemos verificar más a detalle en el panel izquierdo donde si se desea seleccionar todas las filas de la tabla. Y en las variables donde si se pueden mostrar las variables que se presentan. En el apartado del color por instancia de clases es mención a cuando se vaya a realizar un análisis ya se para un tipo de algoritmo. Por eso se deja seleccionado. También se puede dejar la opción en mostrar datos automáticamente, o si se prefiere hacerlo manualmente solo quitamos la opción.

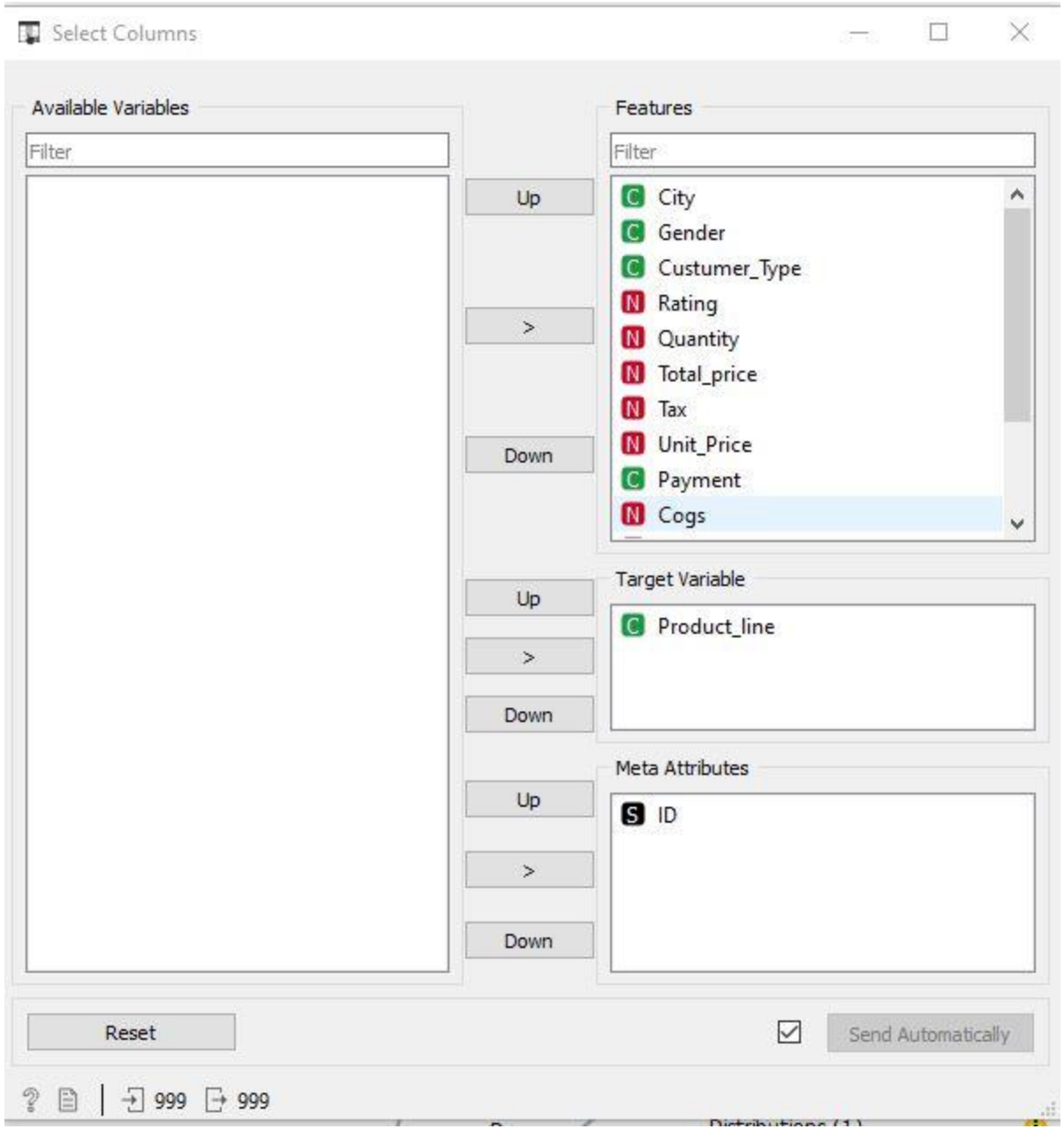


	ID	Branch	City	Custumer_Type	Gender	Product_line	Unit_Price	Quantity	Tax	Total_price	Date_Puncharse
1	750-67-8428	A	Yangon	Member	Female	Health and ...	74.69	7	26.1415	548.9715	2019-01-05
2	226-31-3081	C	Naypyitaw	Normal	Female	Electronic ...	15.28	5	3.8200	80.2200	2019-03-08
3	631-41-3108	A	Yangon	Normal	Male	Home and ...	46.33	7	16.2155	340.5255	2019-03-03
4	123-19-1176	A	Yangon	Member	Male	Health and ...	58.22	8	23.2880	489.0480	2019-01-27
5	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2019-02-08
6	699-14-3026	C	Naypyitaw	Normal	Male	Electronic ...	85.39	7	29.8865	627.6165	2019-03-25
7	355-53-5943	A	Yangon	Member	Female	Electronic ...	68.84	6	20.6520	433.6920	2019-02-25
8	315-22-5665	C	Naypyitaw	Normal	Female	Home and ...	73.56	10	36.7800	772.3800	2019-02-24
9	665-32-9167	A	Yangon	Member	Female	Health and ...	36.26	2	3.6260	76.1460	2019-01-10
10	692-92-5582	B	Mandalay	Member	Female	Food and ...	54.84	3	8.2260	172.7460	2019-02-20
11	351-62-0822	B	Mandalay	Member	Female	Fashion ...	14.48	4	2.8960	60.8160	2019-02-06
12	529-56-3974	B	Mandalay	Member	Male	Electronic ...	25.51	4	5.1020	107.1420	2019-03-09
13	365-64-0515	A	Yangon	Normal	Female	Electronic ...	46.95	5	11.7375	246.4875	2019-02-12
14	252-56-2699	A	Yangon	Normal	Male	Food and ...	43.19	10	21.5950	453.4950	2019-02-07
15	829-34-3910	A	Yangon	Normal	Female	Health and ...	71.38	10	35.6900	749.4900	2019-03-29
16	299-46-1805	B	Mandalay	Member	Female	Sports and travel	93.72	6	28.1160	590.4360	2019-01-15
17	765-26-6951	A	Yangon	Normal	Male	Sports and travel	72.61	6	21.7830	457.4430	2019-01-01
18	329-62-1586	A	Yangon	Normal	Male	Food and ...	54.67	3	8.2005	172.2105	2019-01-21
19	319-50-3348	B	Mandalay	Normal	Female	Home and ...	40.30	2	4.0300	84.6300	2019-03-11
20	300-71-4605	C	Naypyitaw	Member	Male	Electronic ...	86.04	5	21.5100	451.7100	2019-02-25
21	371-85-5789	B	Mandalay	Normal	Male	Health and ...	87.98	3	13.1970	277.1370	2019-03-05
22	273-16-6619	B	Mandalay	Normal	Male	Home and ...	33.20	2	3.3200	69.7200	2019-03-15
23	636-48-8204	A	Yangon	Normal	Male	Electronic ...	34.56	5	8.6400	181.4400	2019-02-17
24	549-59-1358	A	Yangon	Member	Male	Sports and travel	88.63	3	13.2945	279.1845	2019-03-02
25	227-03-5010	A	Yangon	Member	Female	Home and ...	52.59	8	21.0360	441.7560	2019-03-22
26	649-29-6775	B	Mandalay	Normal	Male	Fashion ...	33.52	1	1.6760	35.1960	2019-02-08
27	189-17-4241	A	Yangon	Normal	Female	Fashion ...	87.67	2	8.7670	184.1070	2019-03-10

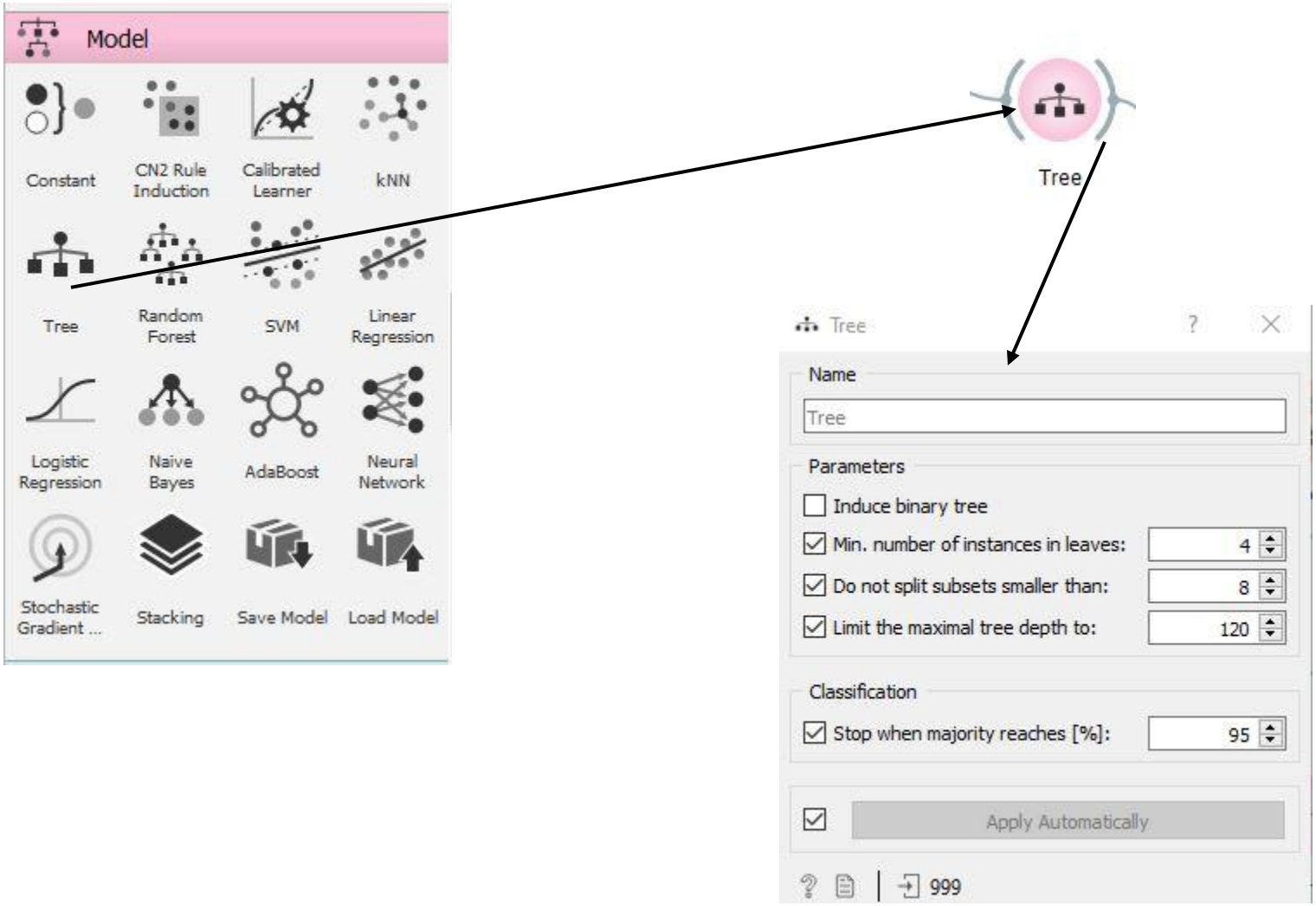
. Luego volvemos al panel y seleccionamos el complemento de seleccionar columnas, luego lo arrastramos al lienzo.



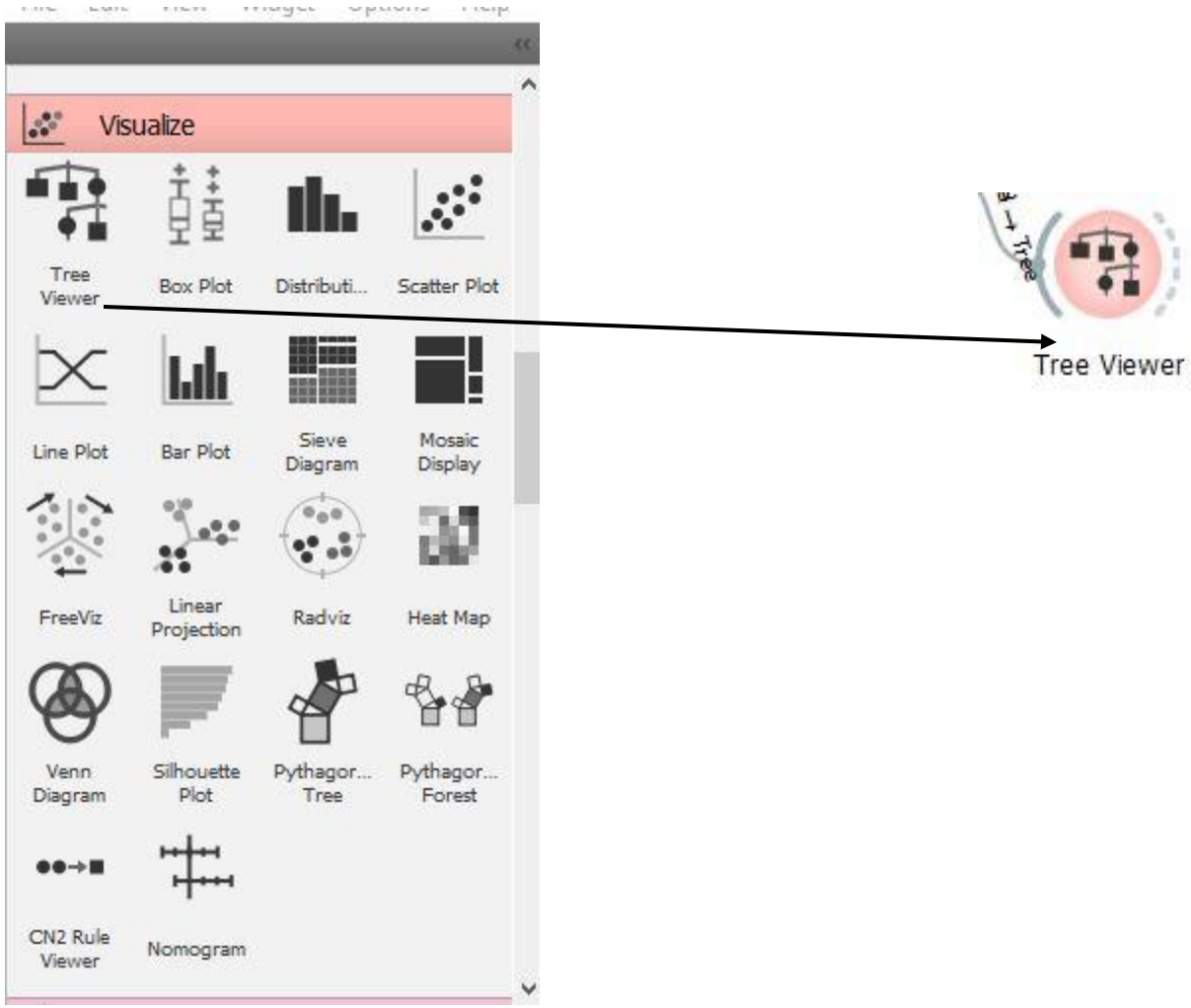
. Luego seleccionamos el complemento, aquí se puede mostrar las opciones donde queremos arrastrar los campos y en qué parte, en este caso tenemos: Variables disponibles, tenemos el apartado de funciones, la variable objetivo y por ultimo los atributos. Por defecto en los atributos deja en categoría tipo texto. En el caso de funciones y variable objetivo se debe seleccionar manualmente el análisis lo cual como variable objetivo dejaremos product_line y todas las funciones habilitadas. Para mayor precisión en el árbol de decisión.



. Luego vamos al apartado de modelo, arrastramos al lienzo el complemento tree para crear el árbol de decisiones y después lo unimos con el select column. Damos click en el complemento de tree y se nos mostrara las opciones y configuraciones del árbol dejamos las opciones que ya están predeterminadas, exceptuando la parte de inducir binario al árbol, esa se deja sin enmarcar.

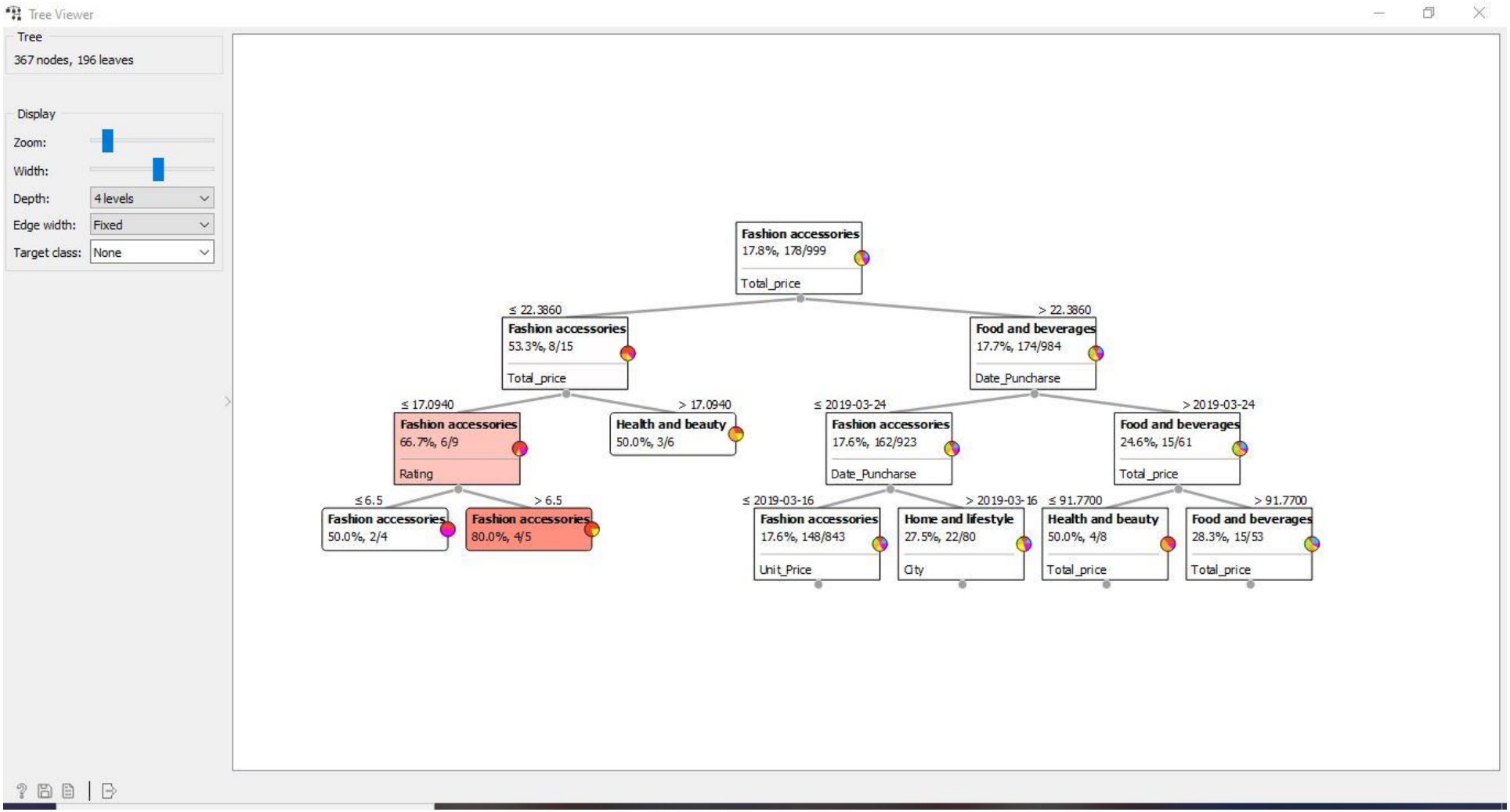


. Ahora en el panel, nos vamos al apartado visualizar, donde arrastramos el complemento de tree viewer al lienzo y lo unimos desde el complemento anterior que era tree.

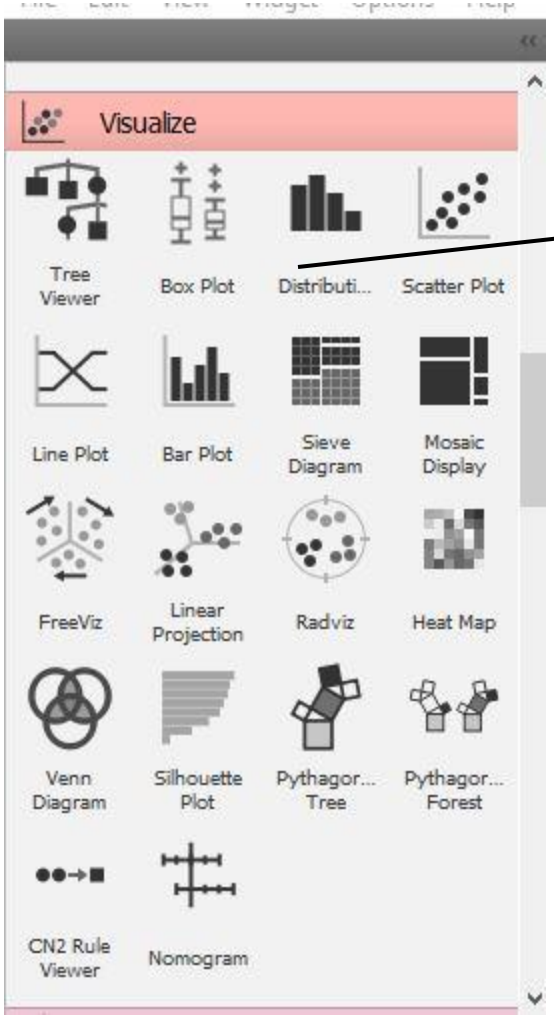


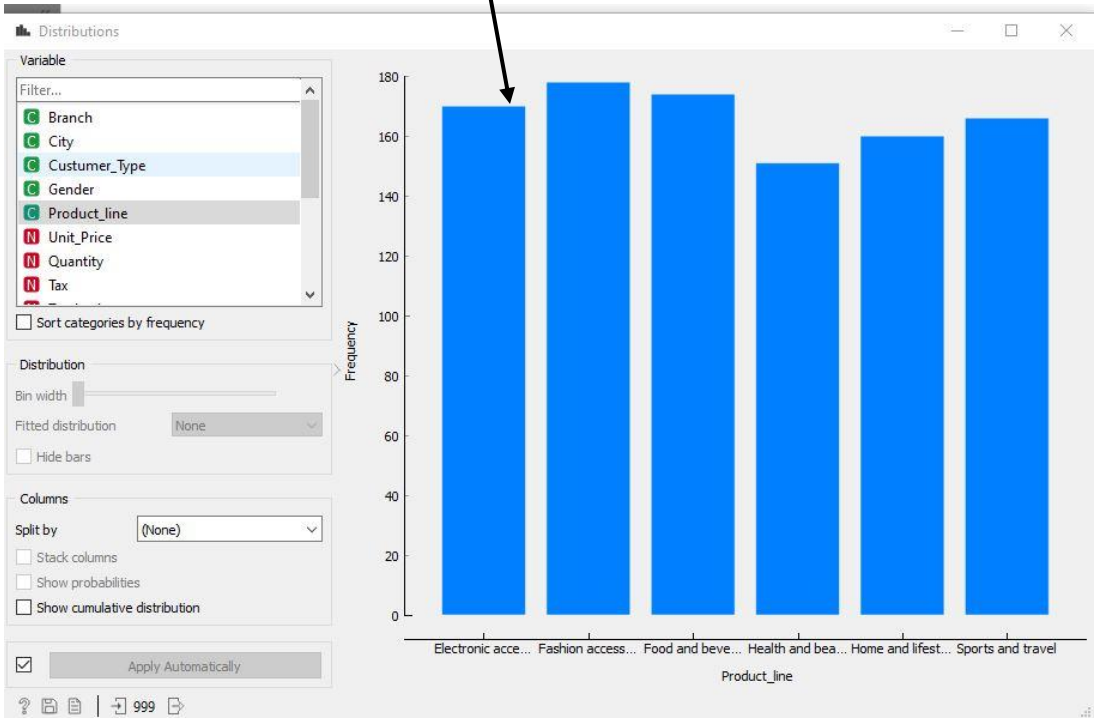
. Y luego damos click en el complemento de visualizar el árbol, y este es el resultado. Como anteriormente, se definió en la parte de seleccionar columnas, aquí lo podemos observar ya de manera que el análisis y decisiones que a tomado el árbol, en total son 9 niveles en que se demuestra, pero como ejemplo se tomaran

.



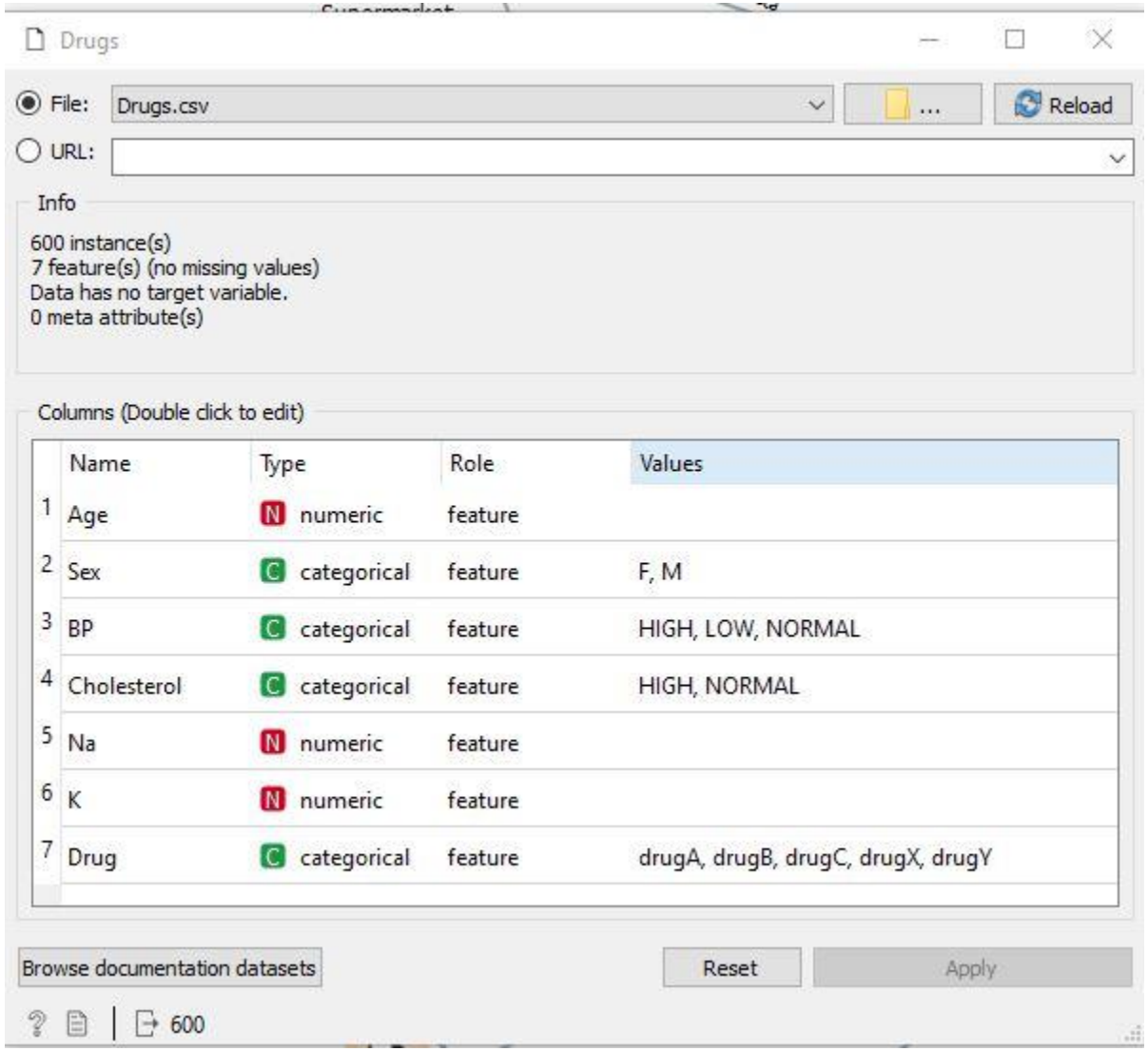
. Como punto opcional, se puede visualizar los datos en una distribución eso quiere decir, como un análisis gráfico. Donde se observa los datos de cada campo y como es demostrado en una grafica de barras. Solo arrastramos el complemento de distribución al lienzo y lo unimos directamente con el complemento file donde alberga el dataset que estamos utilizando.



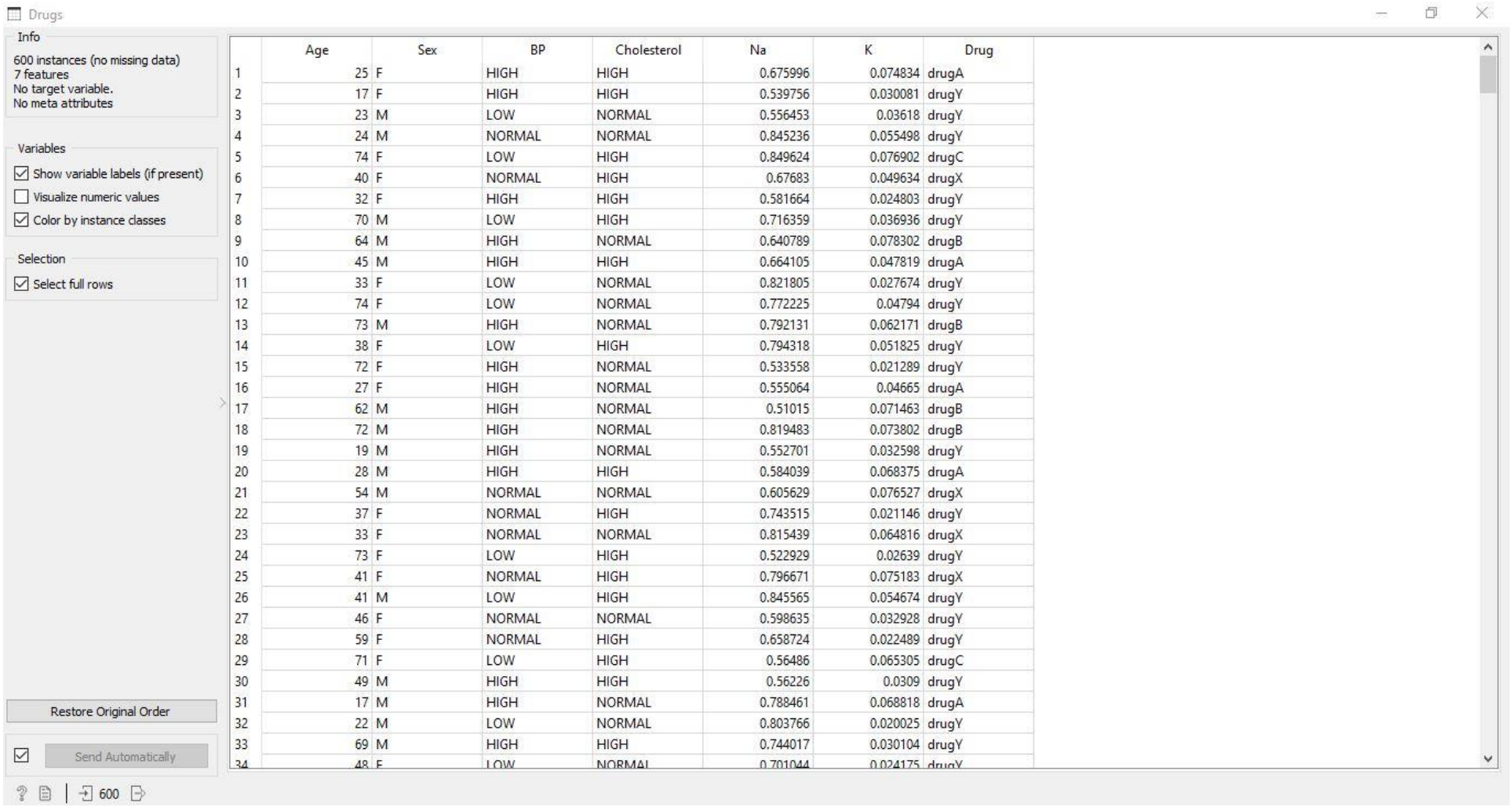


K-MEANS (ORANGE DATAMINING).

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

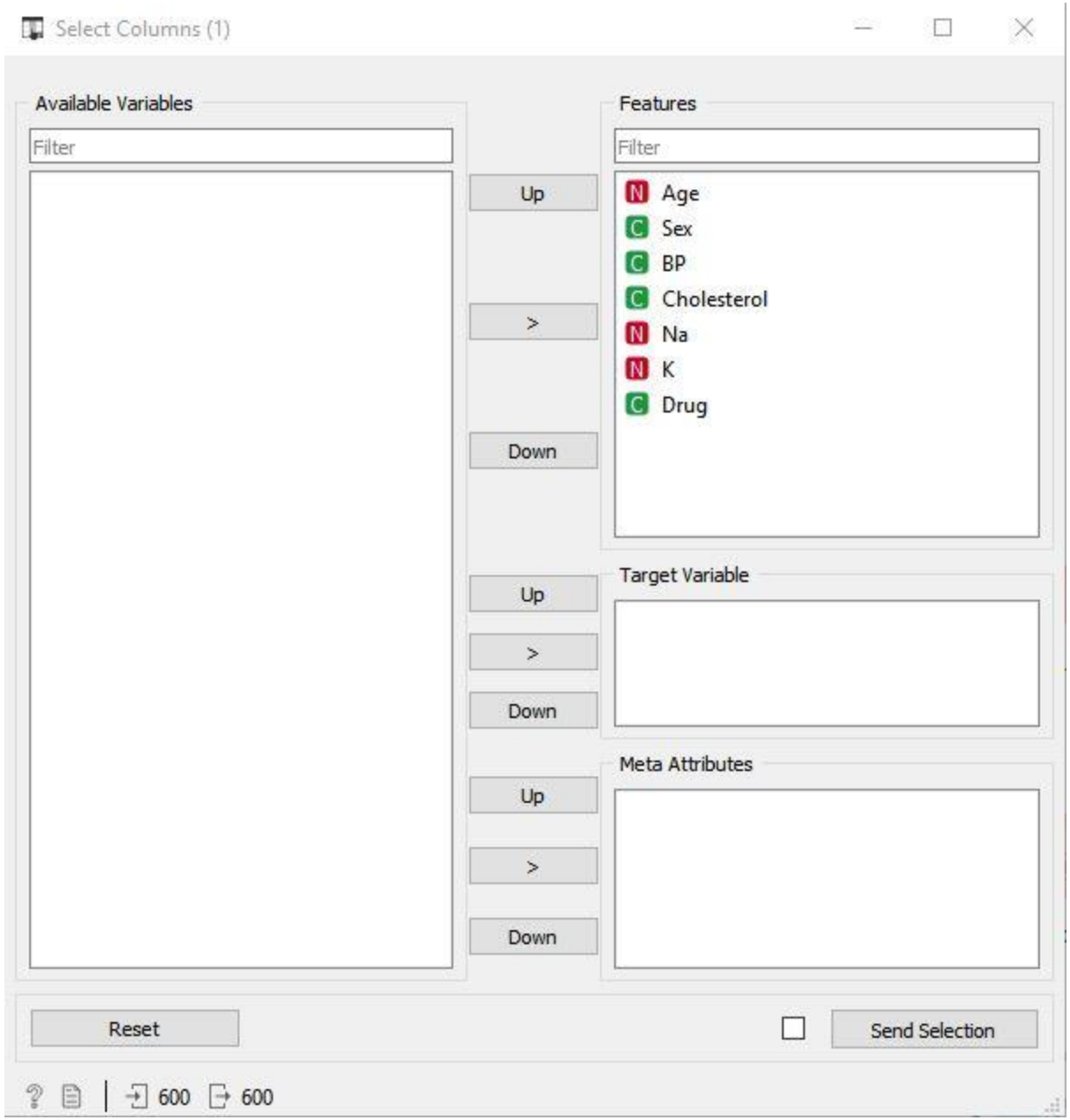


. Tomamos también un data table para poder ver mas detalladamente los datos que contiene el origen.

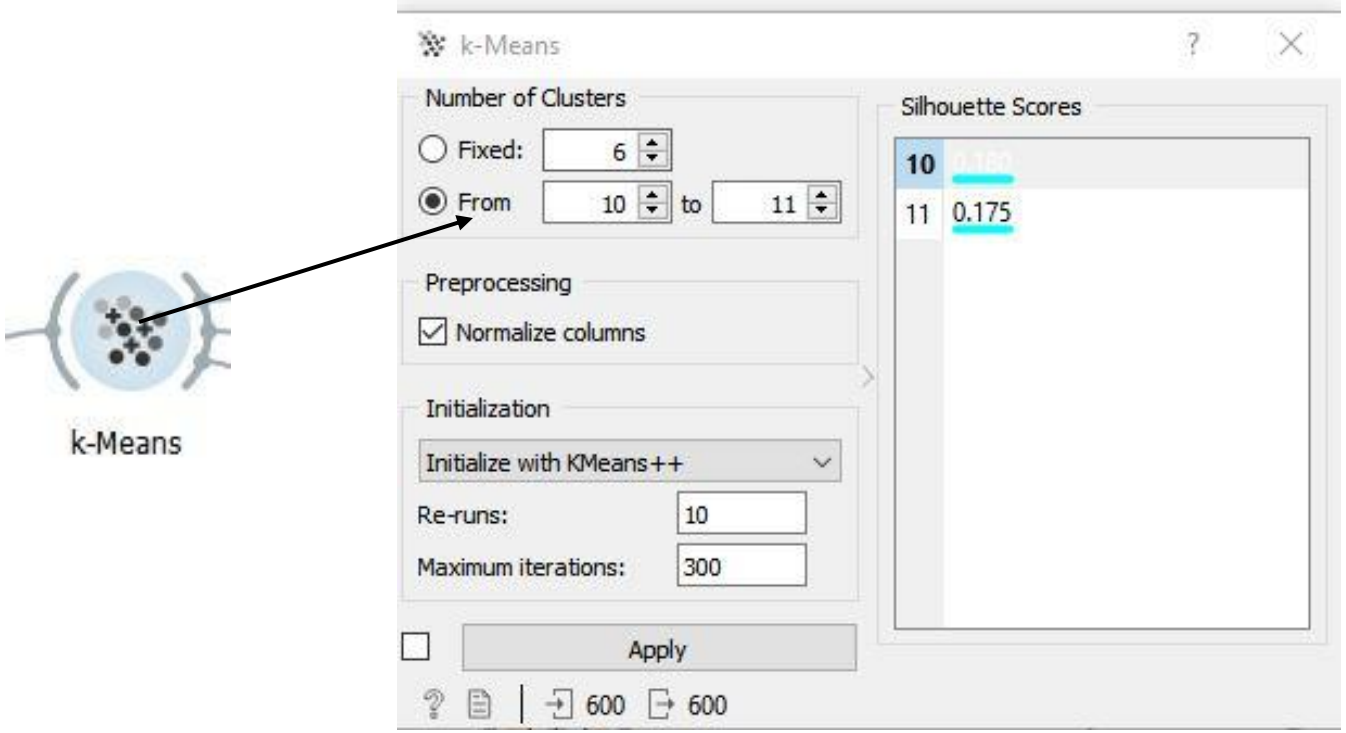


	Age	Sex	BP	Cholesterol	Na	K	Drug
1	25	F	HIGH	HIGH	0.675996	0.074834	drugA
2	17	F	HIGH	HIGH	0.539756	0.030081	drugY
3	23	M	LOW	NORMAL	0.556453	0.03618	drugY
4	24	M	NORMAL	NORMAL	0.845236	0.055498	drugY
5	74	F	LOW	HIGH	0.849624	0.076902	drugC
6	40	F	NORMAL	HIGH	0.67683	0.049634	drugX
7	32	F	HIGH	HIGH	0.581664	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

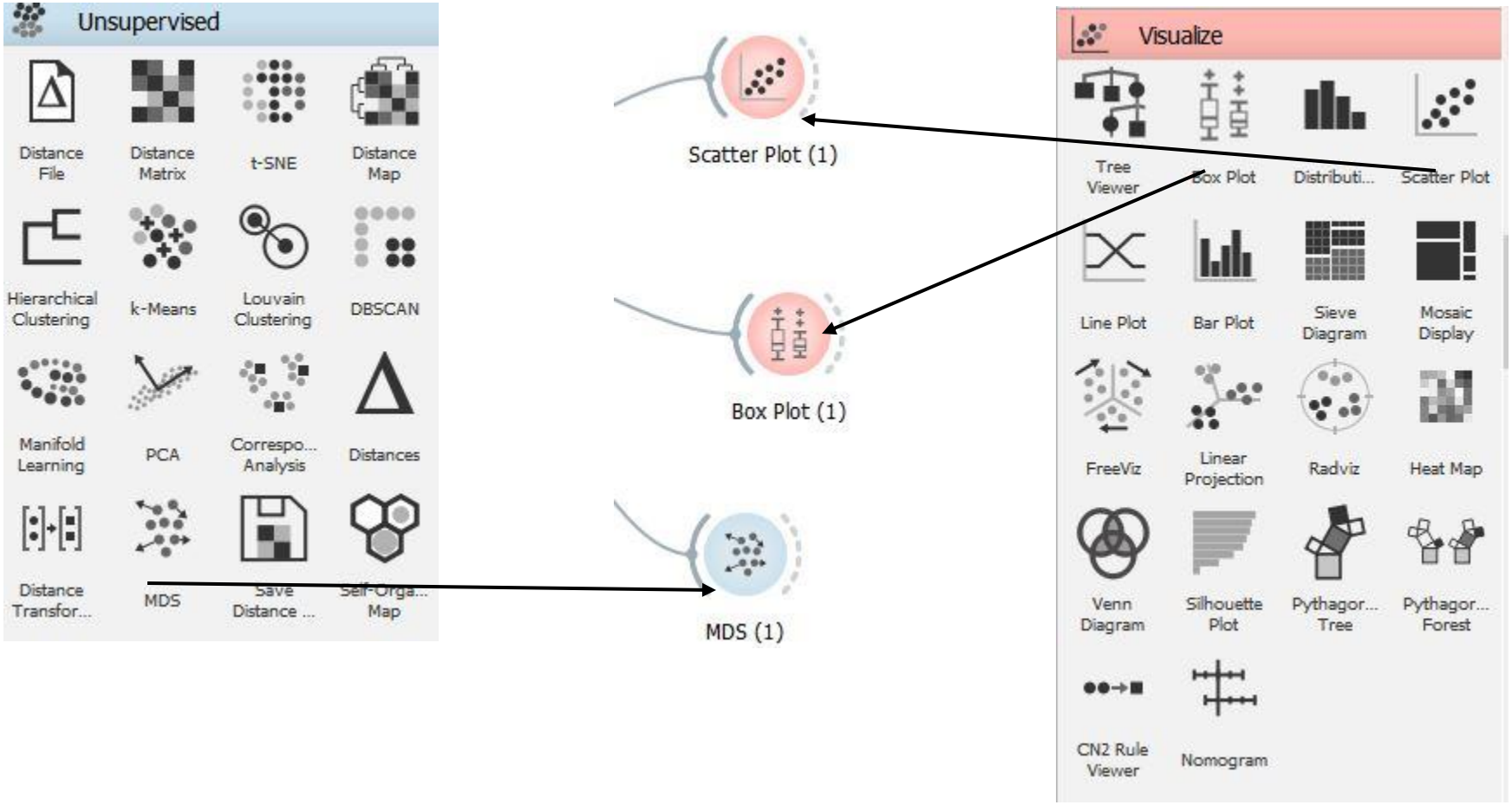
. . Cuando tengamos un select column ya al haberlo unido desde el origen de datos, en este caso solo tomaremos todos los campos de la tabla y no será necesario tener una variable objetivo, ya que en este caso que se usará K-means leerá todo lo que esta en el origen y en el módulo de select column.



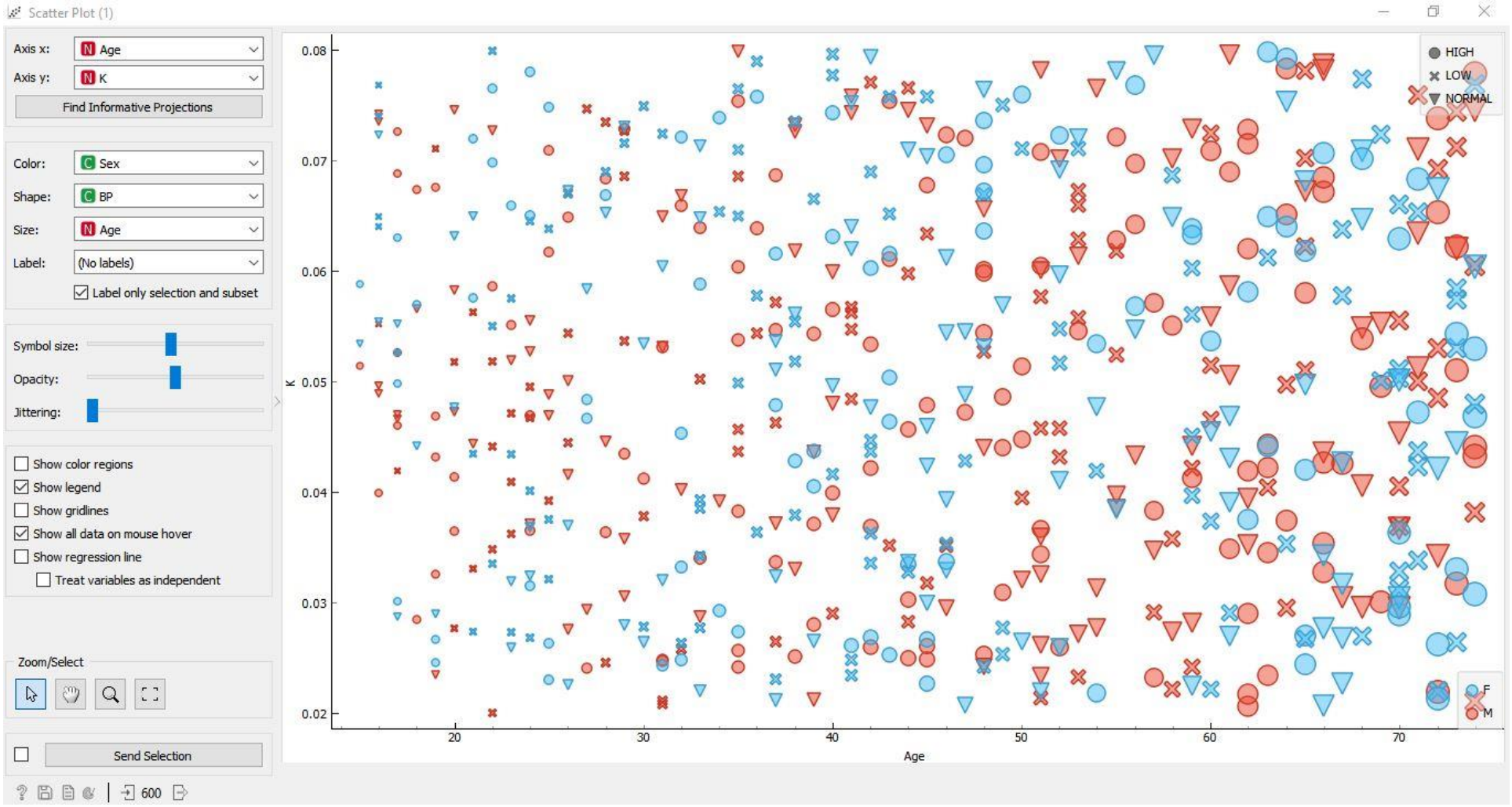
. Luego unimos desde select column hasta el módulo de K-means. Para poder agregar este modulo que esta en oculto, solo arrastramos la línea para conectar, y nos desplegara un menú con todos los modulos y buscamos el que dice K-means. Luego damos click y se nos mostrara un formulario donde podremos configurar 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.



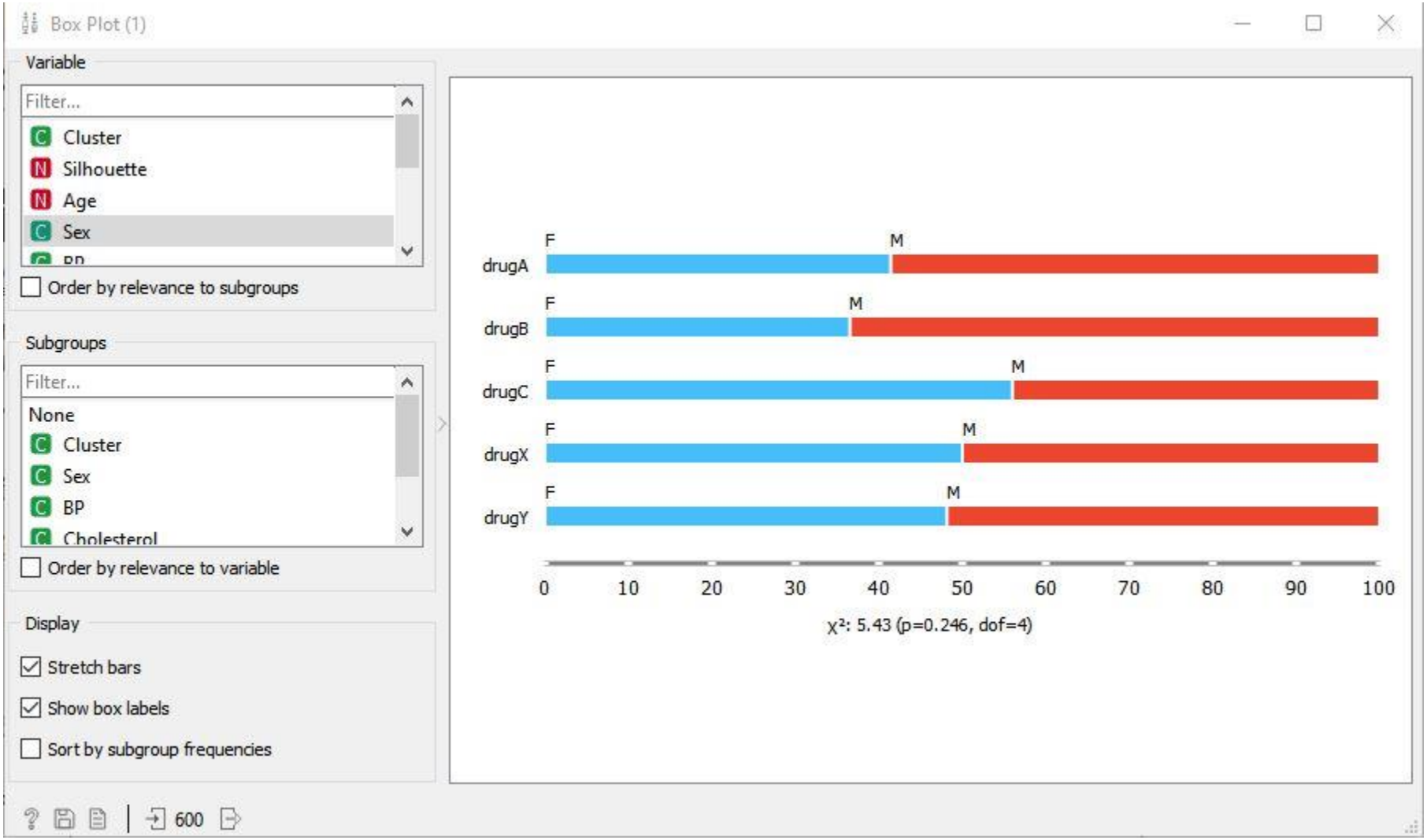
. Ahora vamos al panel de Unsupervised y Visualize. Elegimos los modulos de visualización de Scatter Plot, Box Plot y MDS. ¿Por qué de estos módulos de visualización? Ya que al conectarlos a K-means ya que este algoritmo trabaja con este tipo de visualización de tipo cluster Esto nos permite ver más 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.



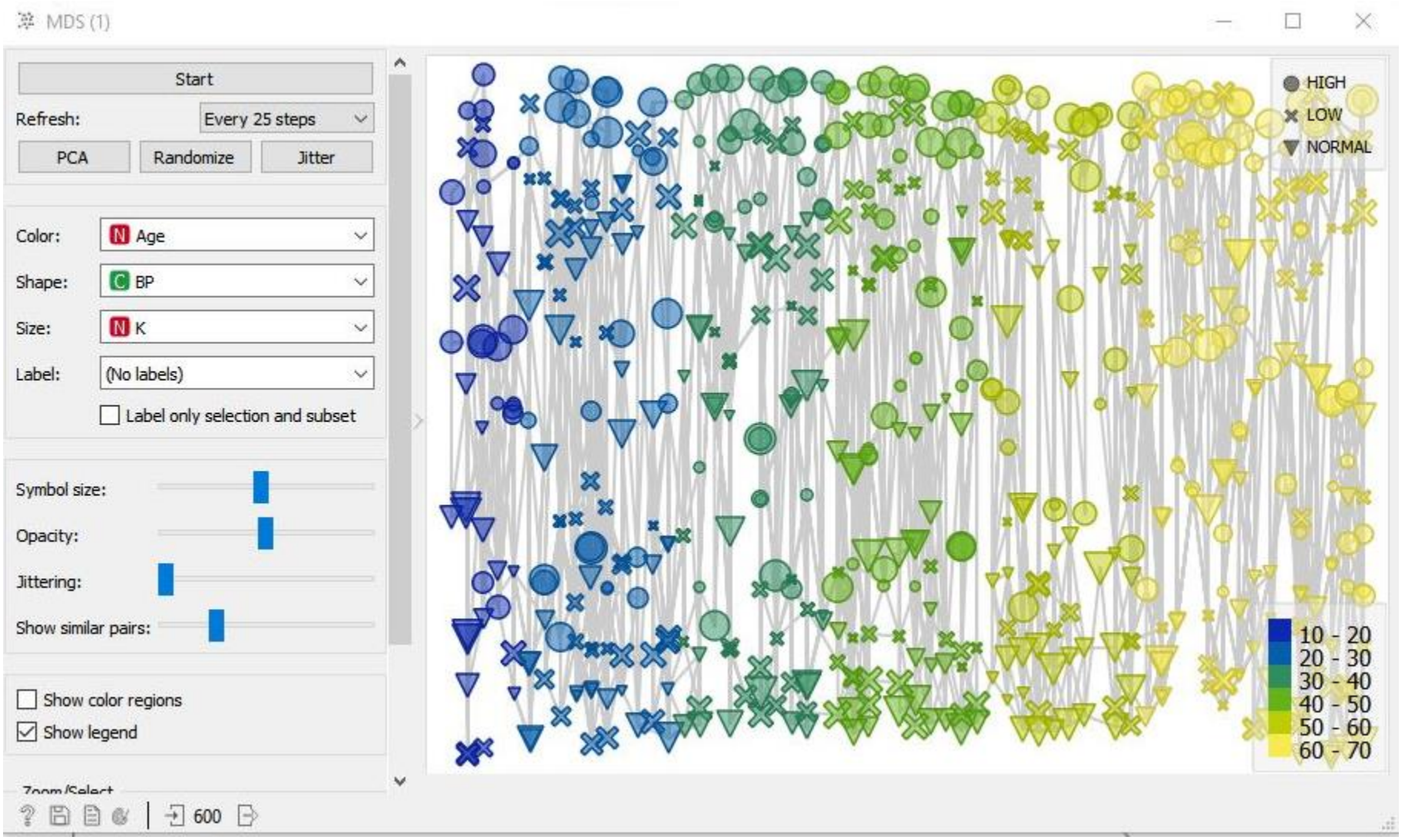
. Este seria el scale plot donde podemos ver que en el eje X esta designado la edad y eje Y K (ketamina). Donde da paso al filtrado de datos que nos permite visualizar mejor el análisis del algoritmo en este caso en color elegimos Sexo en Shape BP (Blood Pressure) y en tamaño la edad, esto nos permite 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.



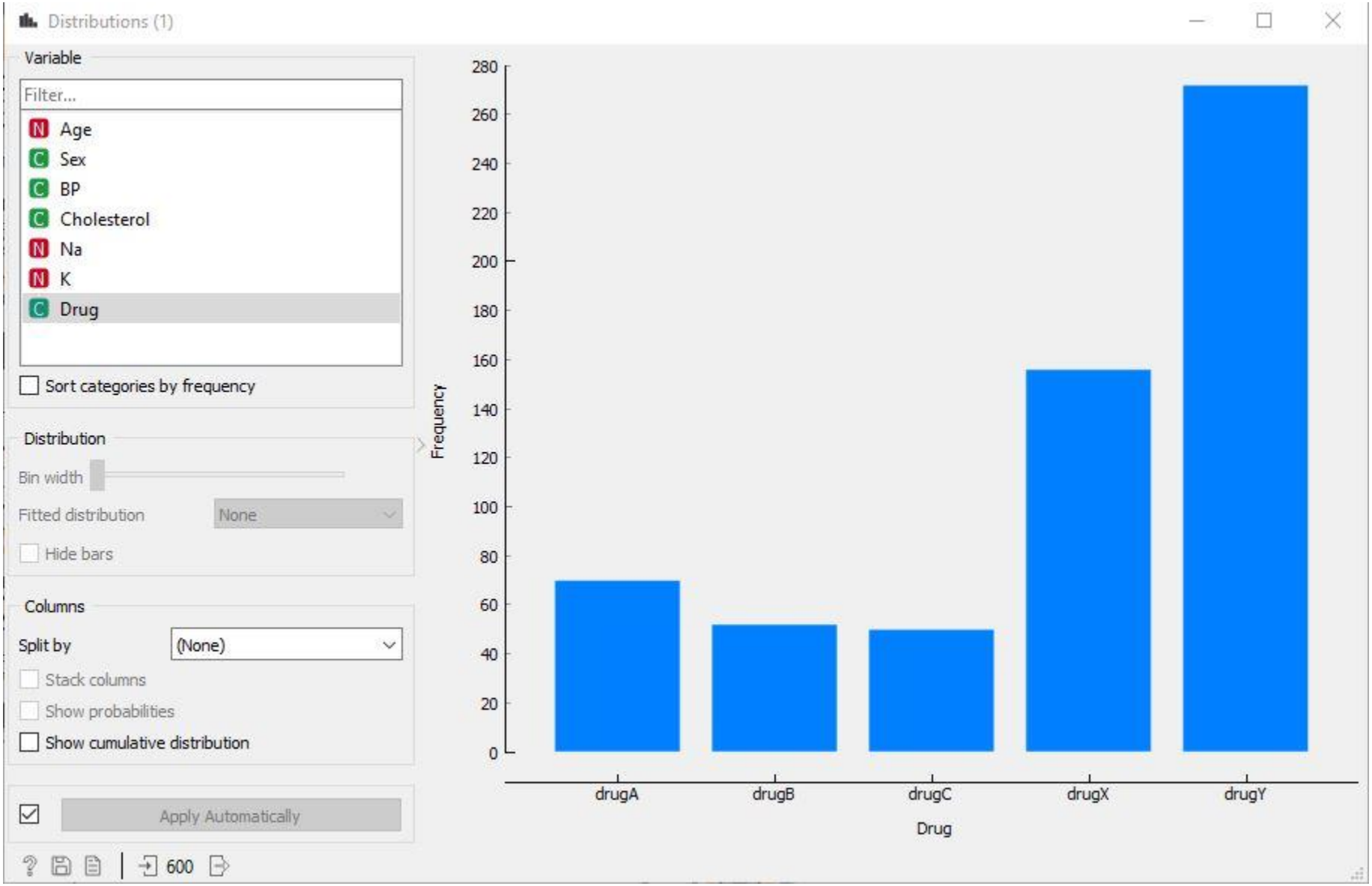
. En Box plot se a tomado los datos de el tipo de droga, la edad y Sexo. En este rango podemos denotar que tipo de droga es mas utilizada por ambos sexos. Y su escala de edades.



. 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.

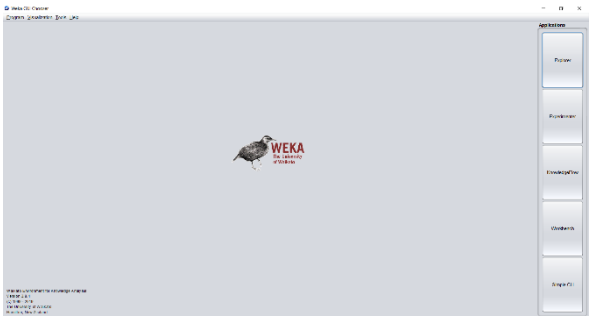


. Ya como parte opcional, si se requiere nuevamente agregar un modulo de visualización de distribución de datos para verlo gráficamente. Aquí podemos ver como seria la frecuencia del tipo de droga.



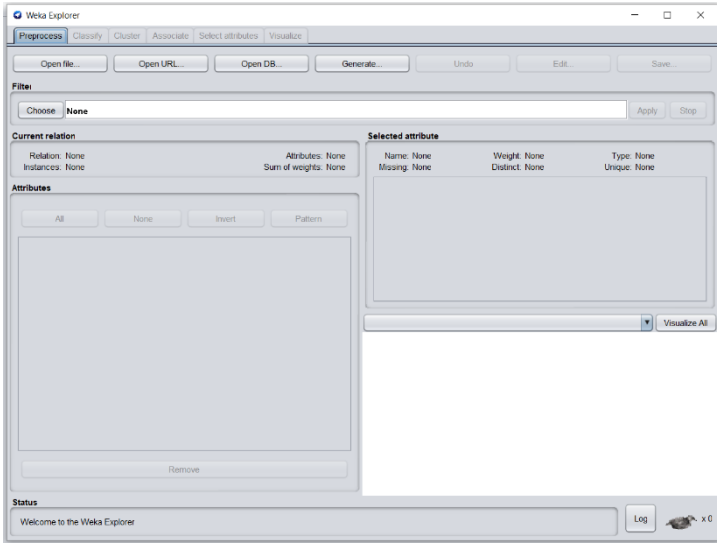
WEKA

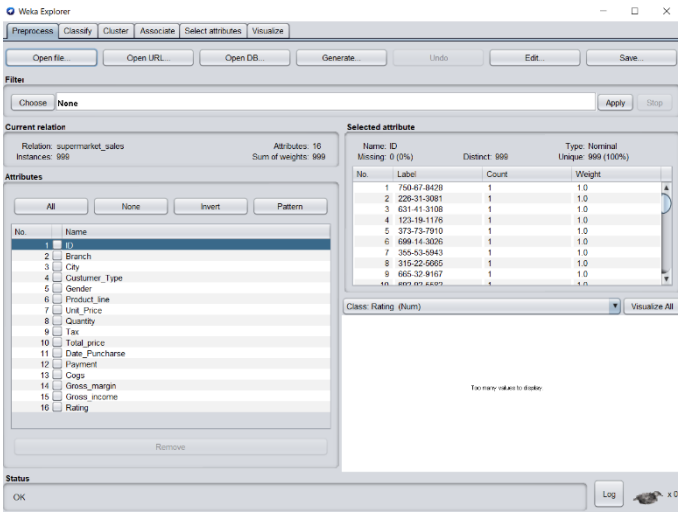
ARBOL DE DECISIÓN (WEKA).



Primeramente abrimos el programa de WEKA y damos click en exponer eso nos abrirá la siguiente ventana.

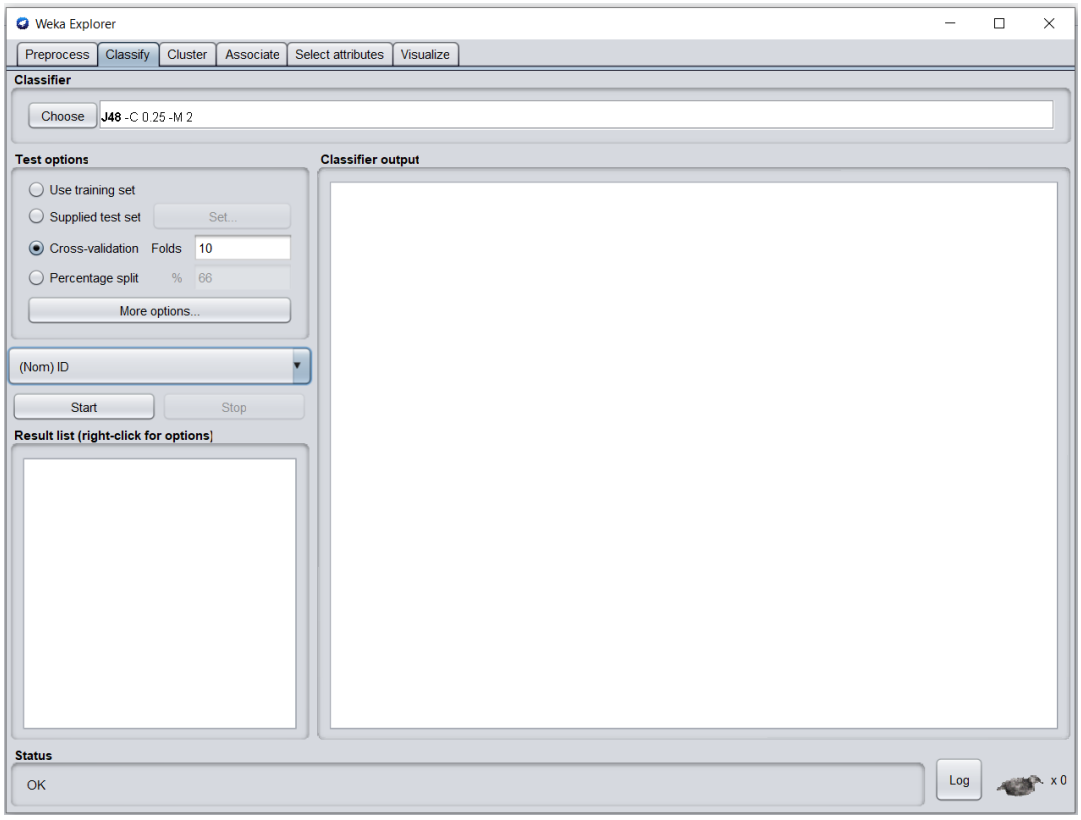
Luego damos clic en el botón Open file no Hola buscamos él chivo en la carpeta en donde las guardamos en este caso el de Superman y nos aparecerán la siguiente ventana.



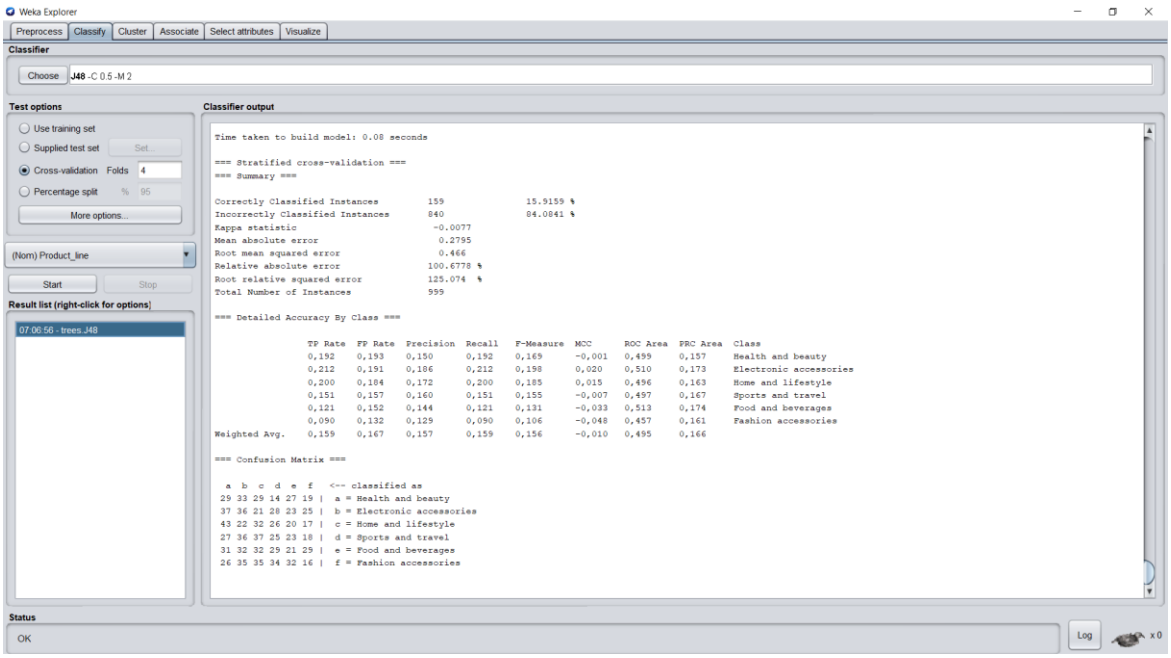


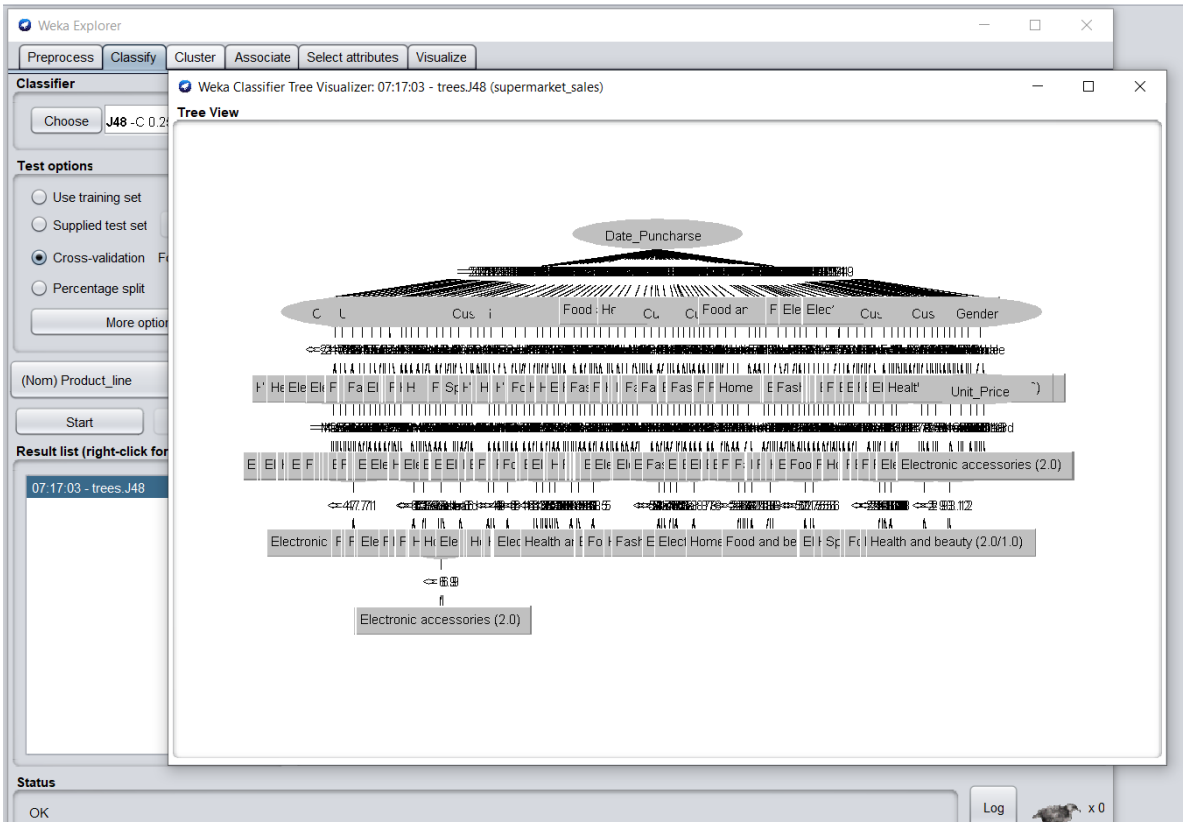
En las pestañas de arriba se activará Hola buscamos la opción de clasificar

Buscamos en shows la opción en la carpeta tree el archivo J48 correspondiente al árbol de decisión y seleccionamos en este caso Producto line.



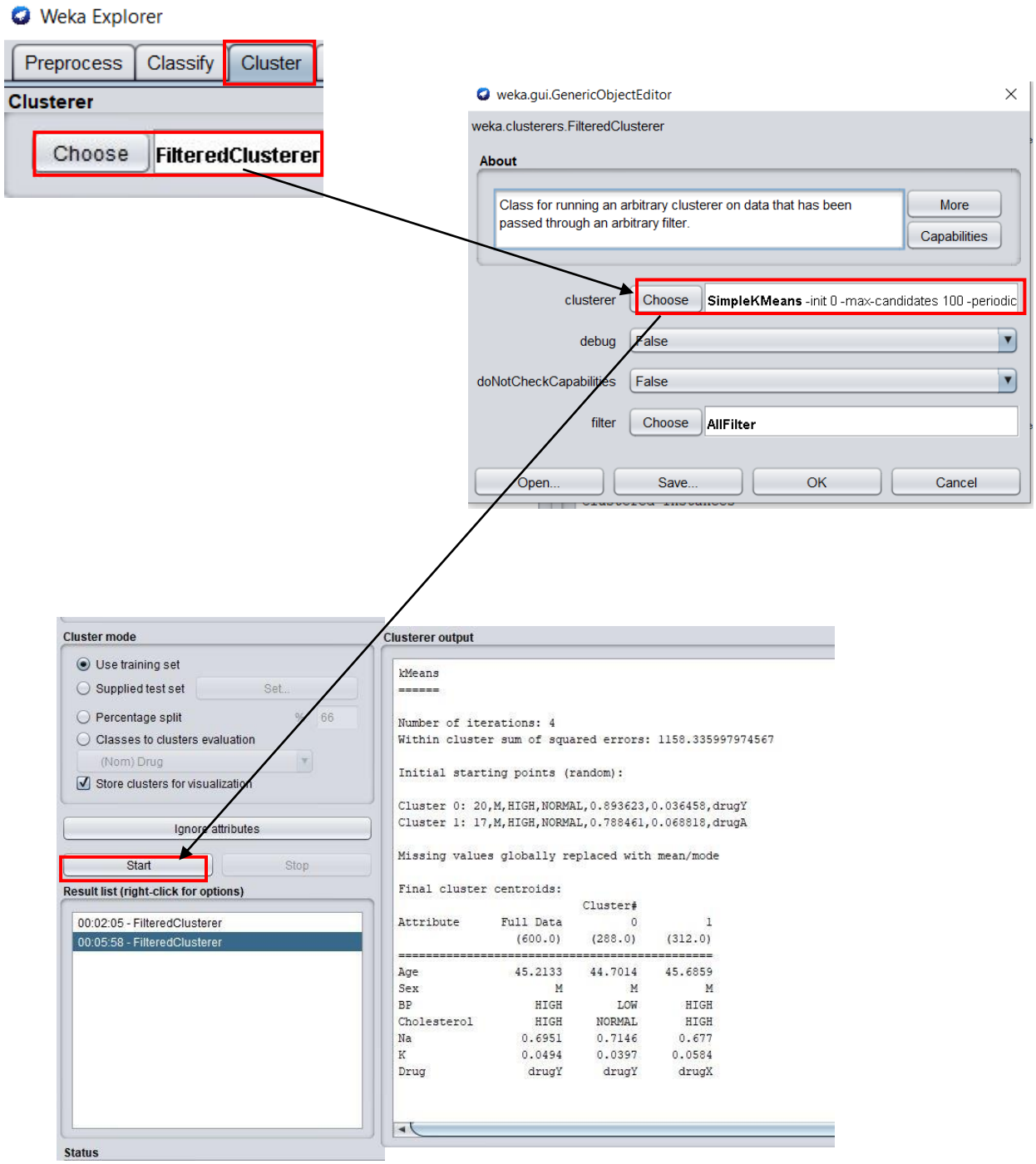
Y nos aparecerá esto, damos clic derecho en el archivo que nos aparece si le damos en visualización de árbol para que nos dé una imagen gráfica de este.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.5 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
4
Percentage split
%
95
More options...
(Nom) Product_line
Start
Stop
Result list (right-click for options)
07:06:56 - trees.J48
Classifier output
Time taken to build model: 0.08 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 159 15.9159 %
Incorrectly Classified Instances 840 84.0841 %
Kappa statistic -0.0077
Mean absolute error 0.2795
Root mean squared error 0.466
Relative absolute error 100.6778 %
Root relative squared error 125.074 %
Total Number of Instances 999
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0,192 0,193 0,150 0,192 0,169 -0,001 0,499 0,157 Health and beauty
0,212 0,191 0,186 0,212 0,198 0,020 0,510 0,173 Electronic accessories
0,200 0,184 0,172 0,200 0,185 0,015 0,496 0,163 Home and lifestyle
0,151 0,157 0,160 0,151 0,155 -0,007 0,497 0,167 Sports and travel
0,121 0,152 0,144 0,121 0,131 -0,033 0,513 0,174 Food and beverages
0,090 0,132 0,129 0,090 0,106 -0,048 0,457 0,161 Fashion accessories
Weighted Avg. 0,159 0,167 0,157 0,159 0,156 -0,010 0,495 0,166
=== Confusion Matrix ===
a b c d e f <-- classified as
29 33 29 14 27 19 | a = Health and beauty
37 36 21 28 23 25 | b = Electronic accessories
43 22 32 26 20 17 | c = Home and lifestyle
27 36 37 25 23 18 | d = Sports and travel
31 32 32 29 21 25 | e = Food and beverages
26 35 35 34 32 16 | f = Fashion accessories
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
Test options
Use training set
Supplied test set
Cross-validation
Percentage split
More optio
(Nom) Product_line
Start
Result list (right-click for
07:17:03 - trees.J48
Status
OK
Log
x 0
Weka Classifier Tree Visualizer: 07:17:03 - trees.J48 (supermarket_sales)
Tree View
Date_Puncharse
Unit_Price
Electronic accessories (2.0)
Health and beauty (2.0/1.0)
Electronic accessories (2.0)

K-MEANS (WEKA).

Buscamos la pestaña Cluster y en Choose Buscamos el archivo FilteredCluster, al dar clic en la letra negra nos aparecerá una ventana en donde seleccionaremos SimpleKMeans, damos ok, y luego en start, clic derecho en el resultado obtenido



Y nos dará esto.