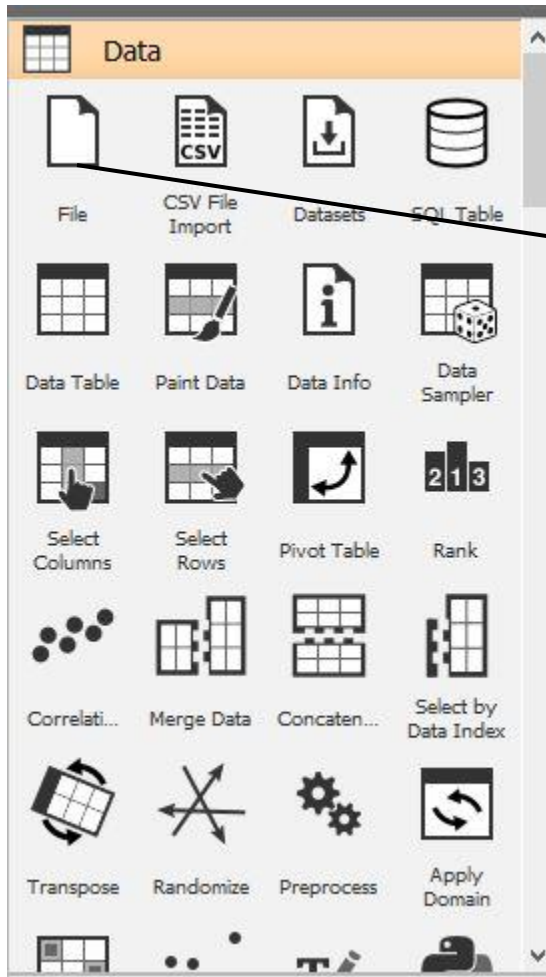
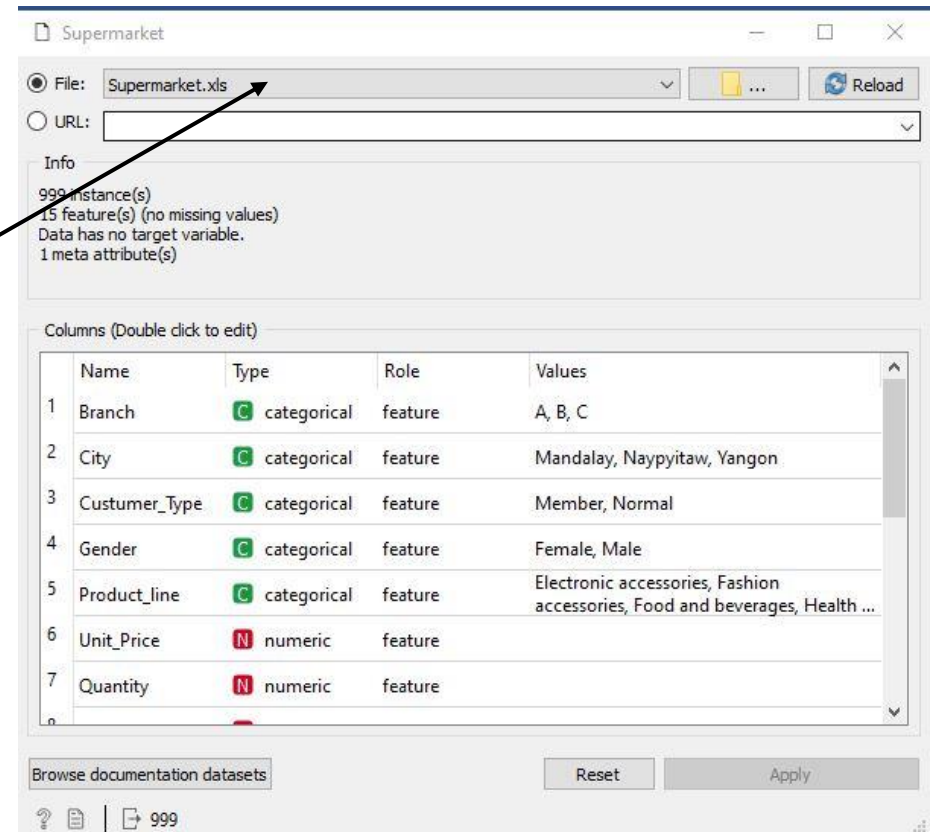
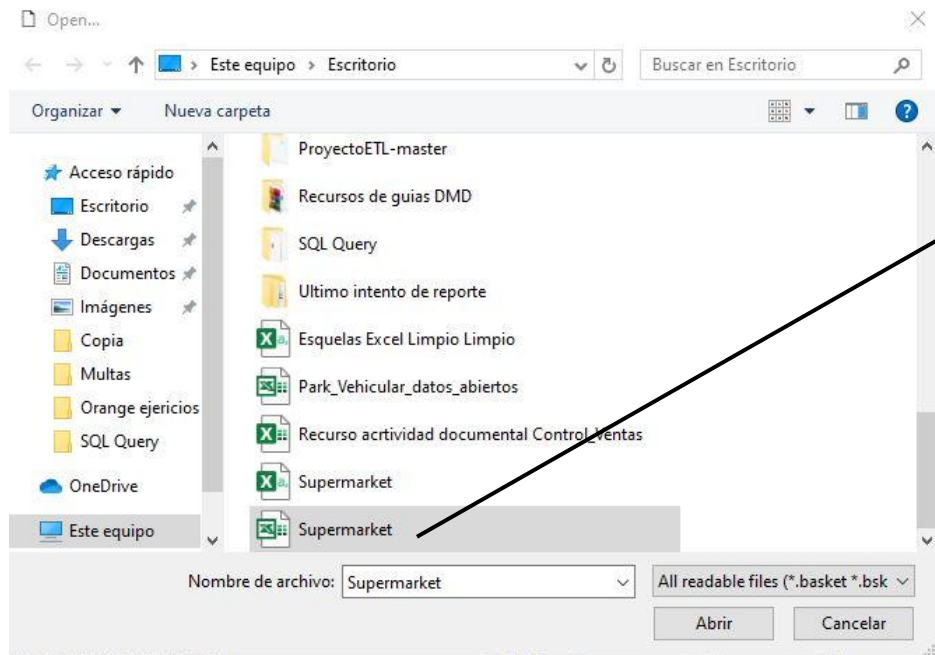


ARBOL DE DECISION (ORANGE DATA MINING).

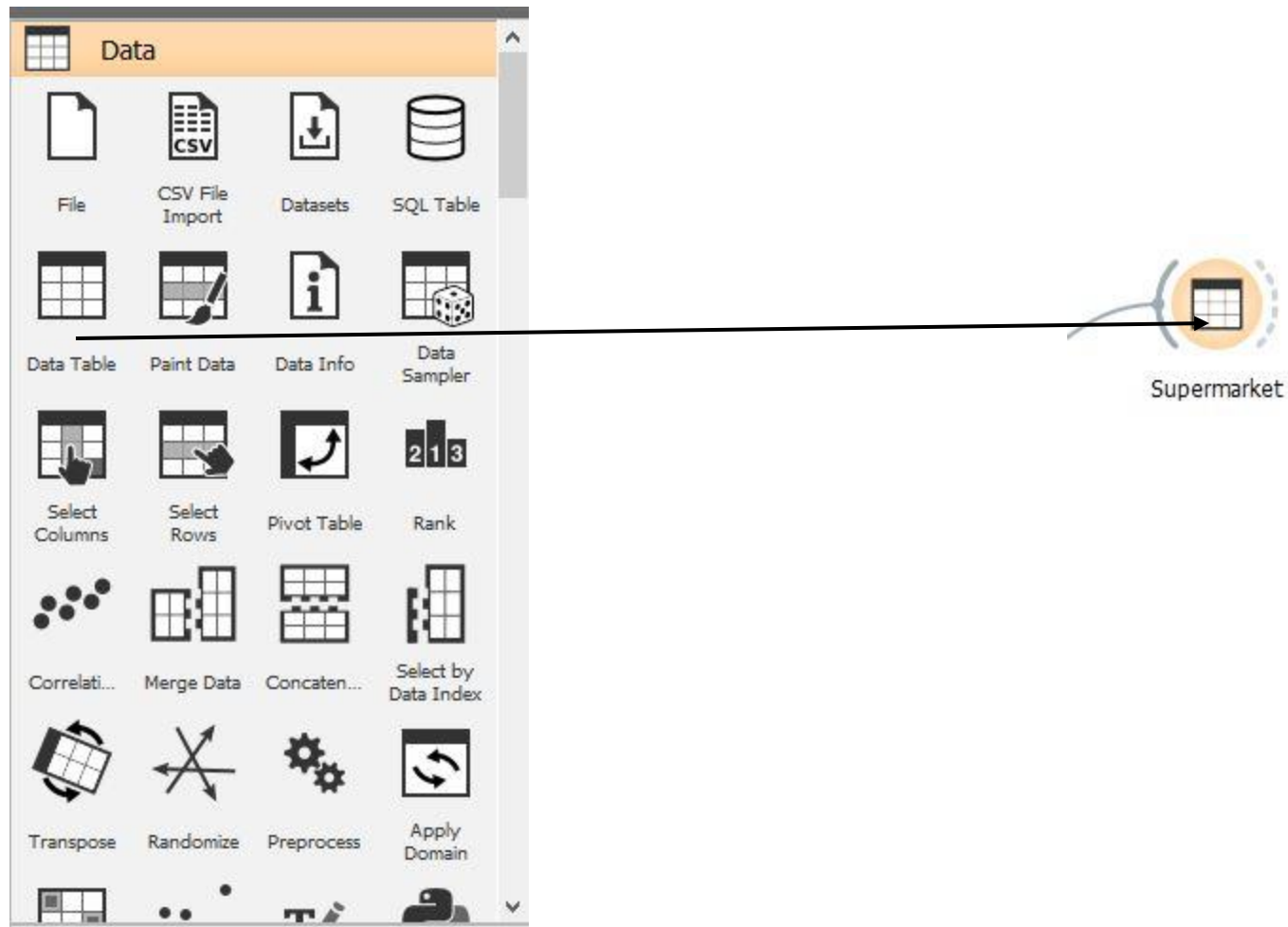
. Se nos muestra un panel que se llama data, arrastramos el complemento fila a lienzo en blanco.



. 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.



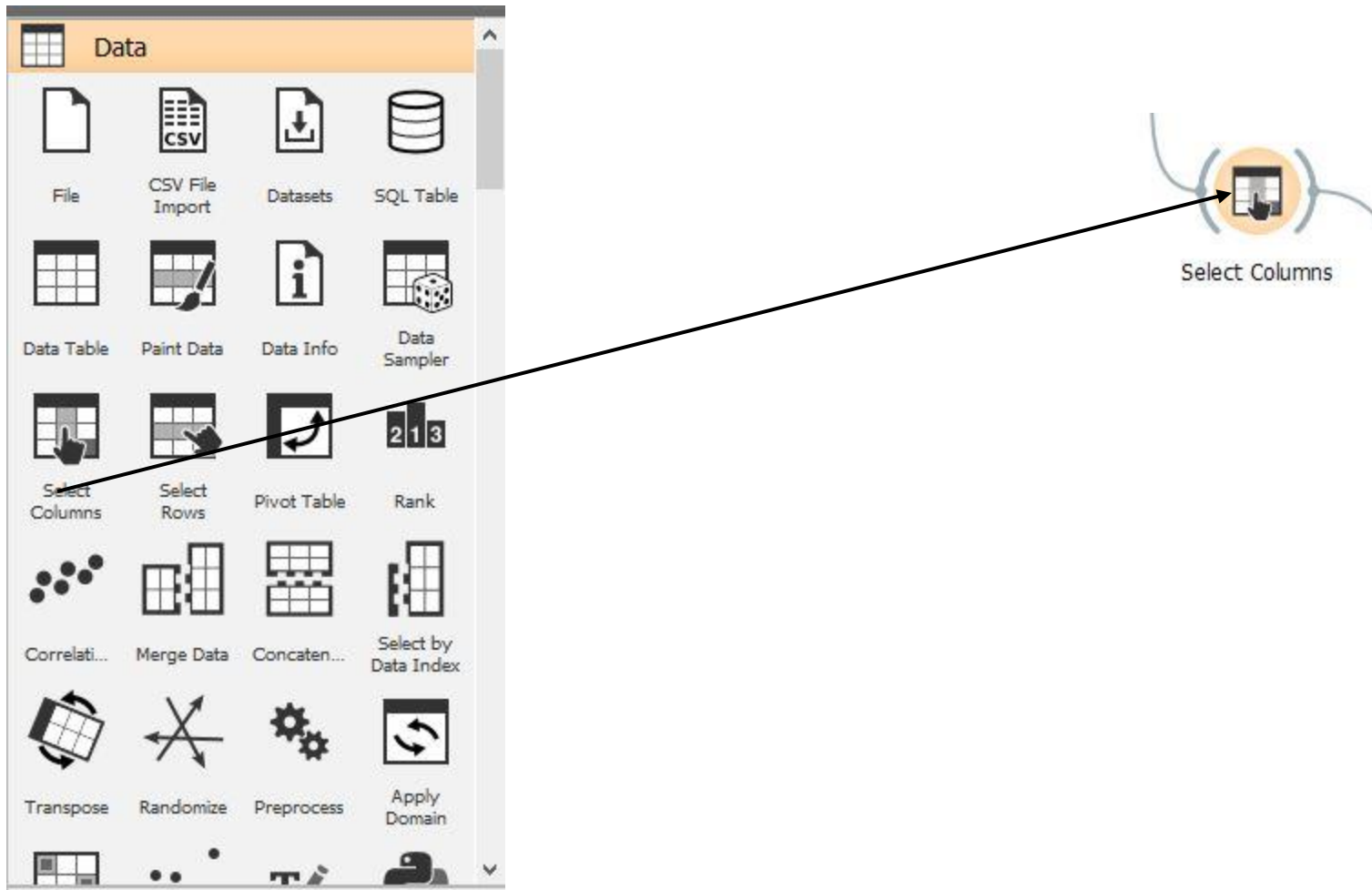
. Luego arrastramos un data table, al lienzo y luego lo unimos al archivo que estamos utilizando, así para que nos puedan aparecer mas a detalle lo que se encuentra en el archivo y todos sus datos.



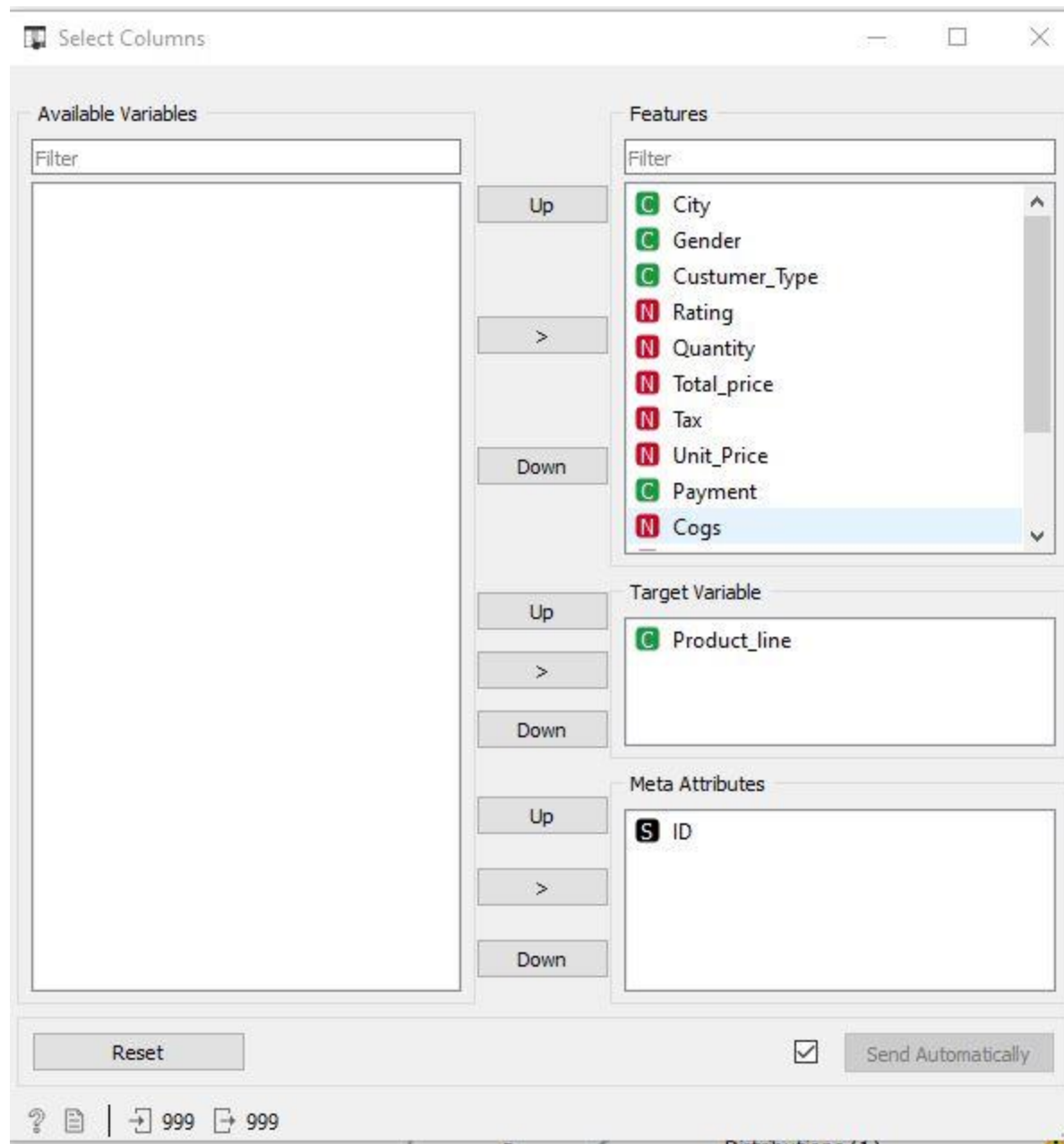
. Y como se puede observar en la tabla toda la información que contiene el archivo, donde podemos verificar mas a detalle en el panel izquierdo donde si se desea seleccionar todas las filas de la tabla. Y en las variables donde si se pueden mostrar las variables que se presentan. En el apartado del color 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.

Data Table											
Info											
999 instances (no missing data) 15 features No target variable. 1 meta attribute											
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											
ID	Branch	City	Customer_Type	Gender	Product_line	Unit_Price	Quantity	Tax	Total_price	Date_Purchase	
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

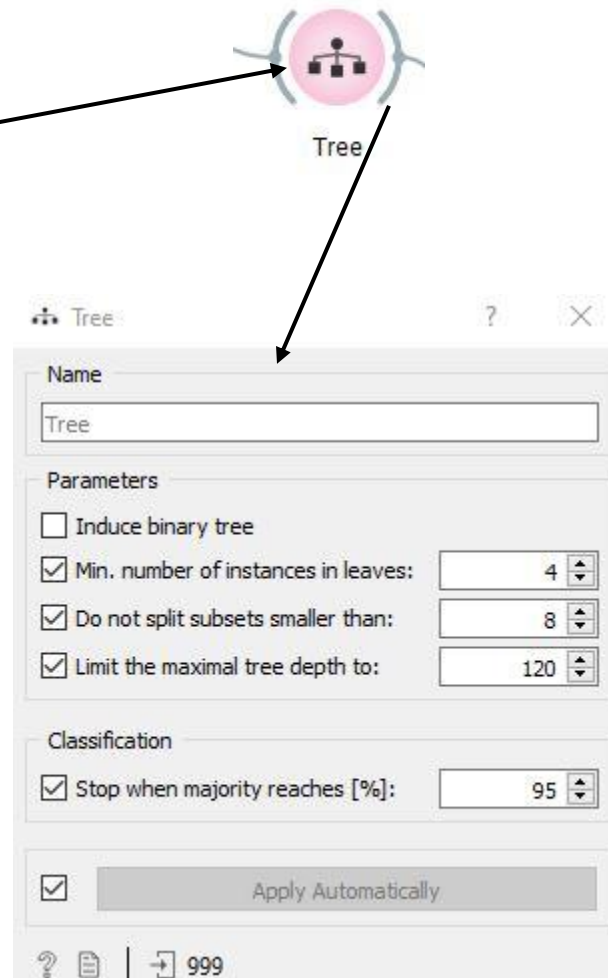
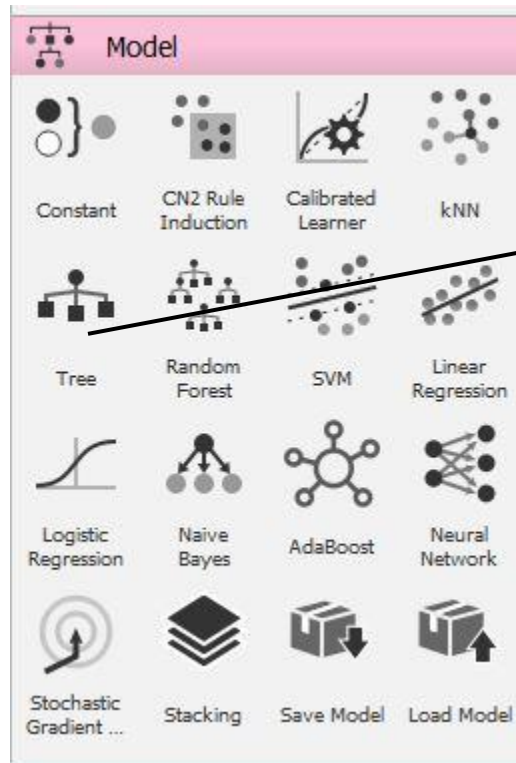
. Luego volvemos al panel y seleccionamos el complemento de seleccionar columnas, luego lo arrastramos al lienzo.



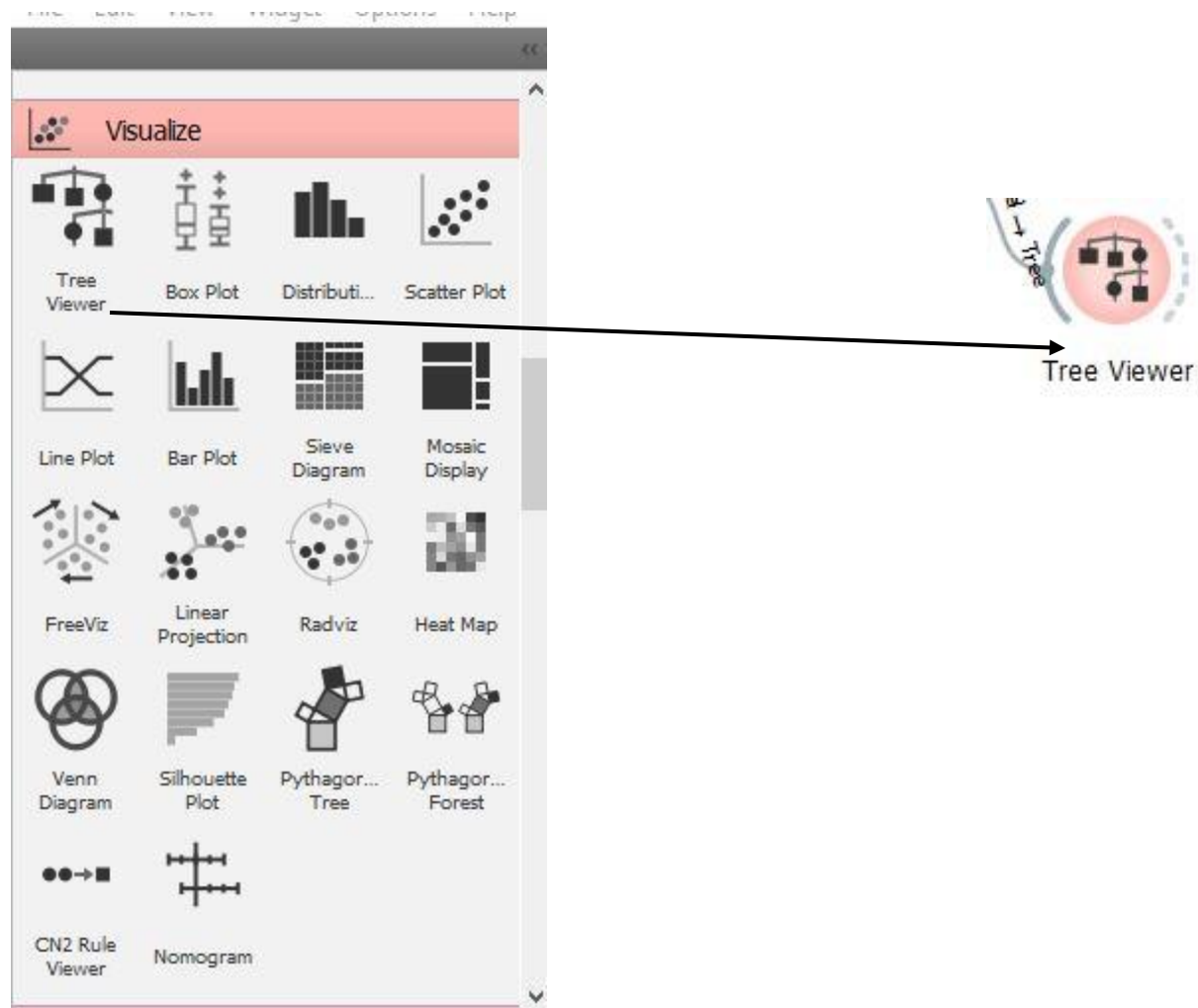
. Luego seleccionamos el complemento, aquí se puede mostrar las opciones donde queremos arrastrar los campos y en qué parte, en este caso tenemos: Variables disponibles, tenemos el apartado de funciones, la variable objetivo y por ultimo los atributos. Por defecto en los atributos deja en categoría tipo 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.



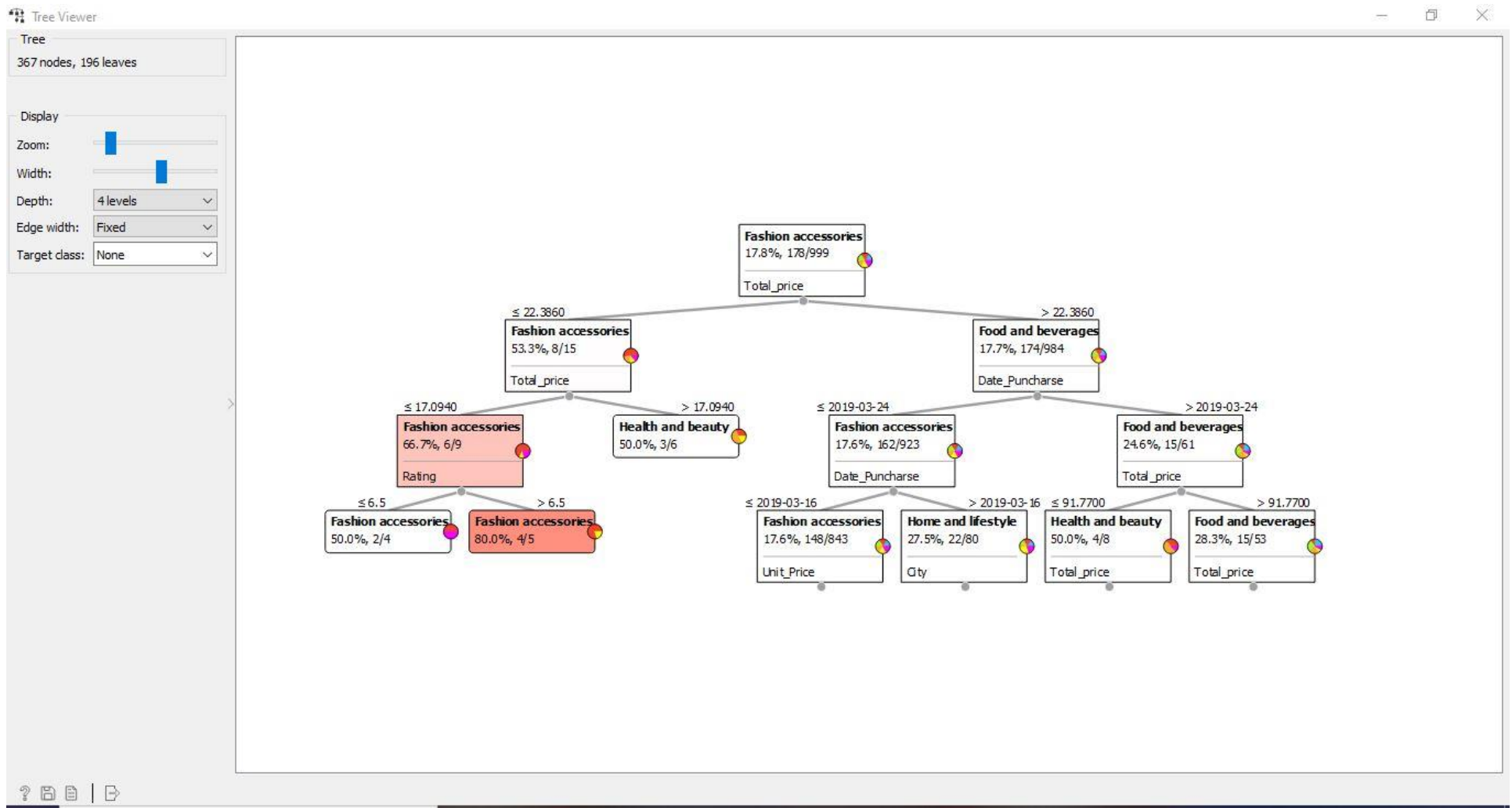
. Luego vamos al apartado de modelo, arrastramos al lienzo el complemento tree para crear el árbol de decisiones y después lo unimos con el select column. Damos click en el complemento de tree y se nos mostrara las opciones y configuraciones del árbol dejamos las opciones que ya están predeterminadas,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



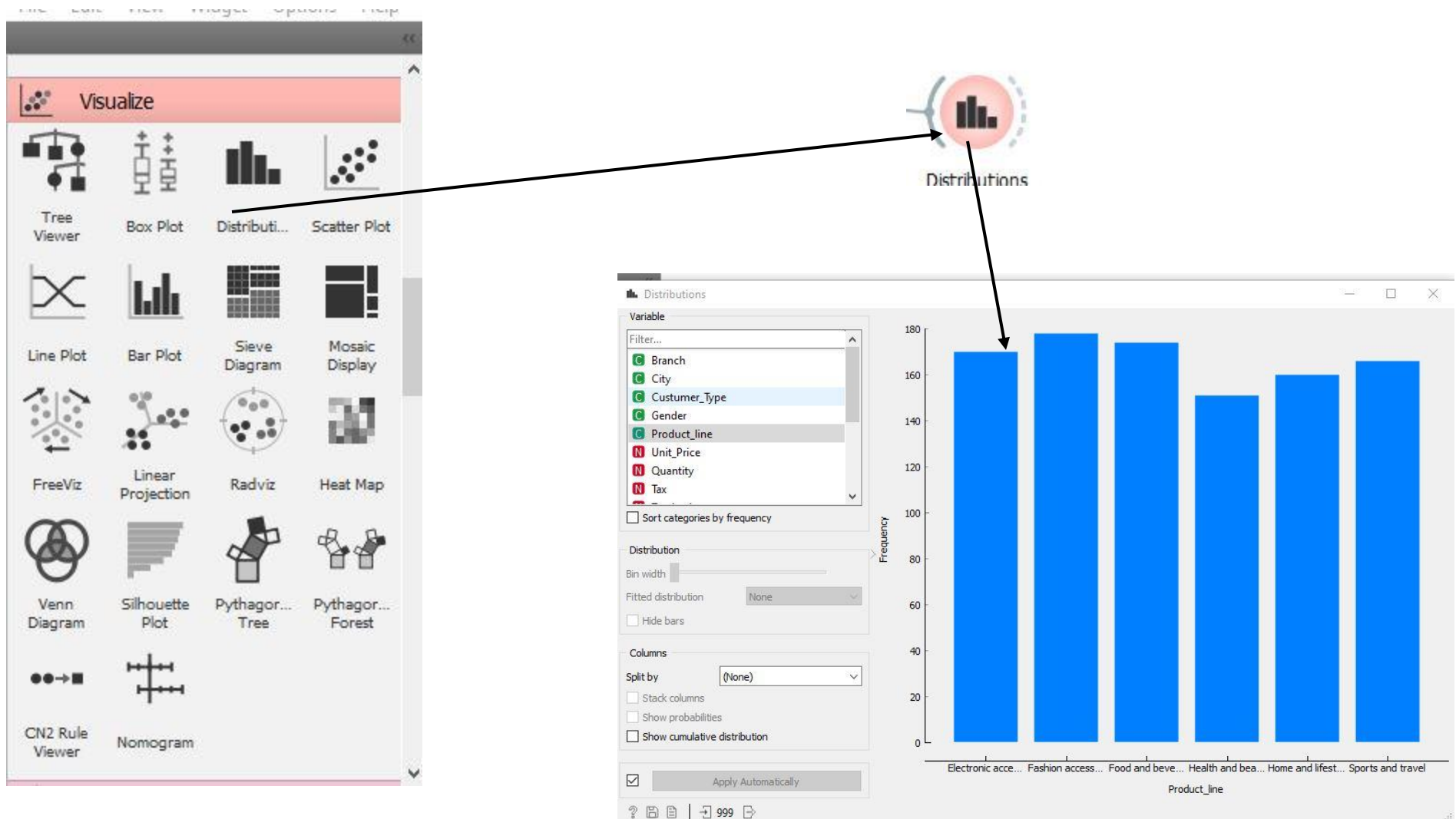
. Ahora en el panel, nos vamos al apartado visualizar, donde arrastramos el complemento de tree viewer al lienzo y lo unimos desde el complemento anterior que era tree.



. 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.



REGLAS DE ASOCIACIÓN (ORANGE DATA MINING).

. Serían los mismos pasos solo que en este caso el dataset que vamos a utilizar un archivo csv de Drugs, y luego lo cargamos y nos aparecerá, el nombre de la columna, el tipo de categoría que esta anexado al tipo de dato que pertenece, el rol que toma y por ultimo los valores y datos que contiene.

The screenshot shows the 'Data' widget in Orange Data Mining. The 'File' tab is selected, showing 'Drugs.csv'. The 'Info' section indicates 600 instances, 7 features, and no target variable. The 'Columns' section displays a table with 7 columns: Name, Type, Role, and Values. The 'Values' column is highlighted. The table lists features: Age (numeric), Sex (categorical), BP (categorical), Cholesterol (categorical), Na (numeric), K (numeric), and Drug (categorical).

	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

Buttons: Browse documentation datasets, Reset, Apply. Status bar: 600 instances.

. Luego sería lo mismo arrastramos un data table al lienzo y lo unimos al dataset origen, y se nos mostrara lo que contiene más detalladamente los datos.

Drugs

Info

600 instances (no missing data)
7 features
No target variable.
No meta attributes

Variables

☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection

☒ Select full rows

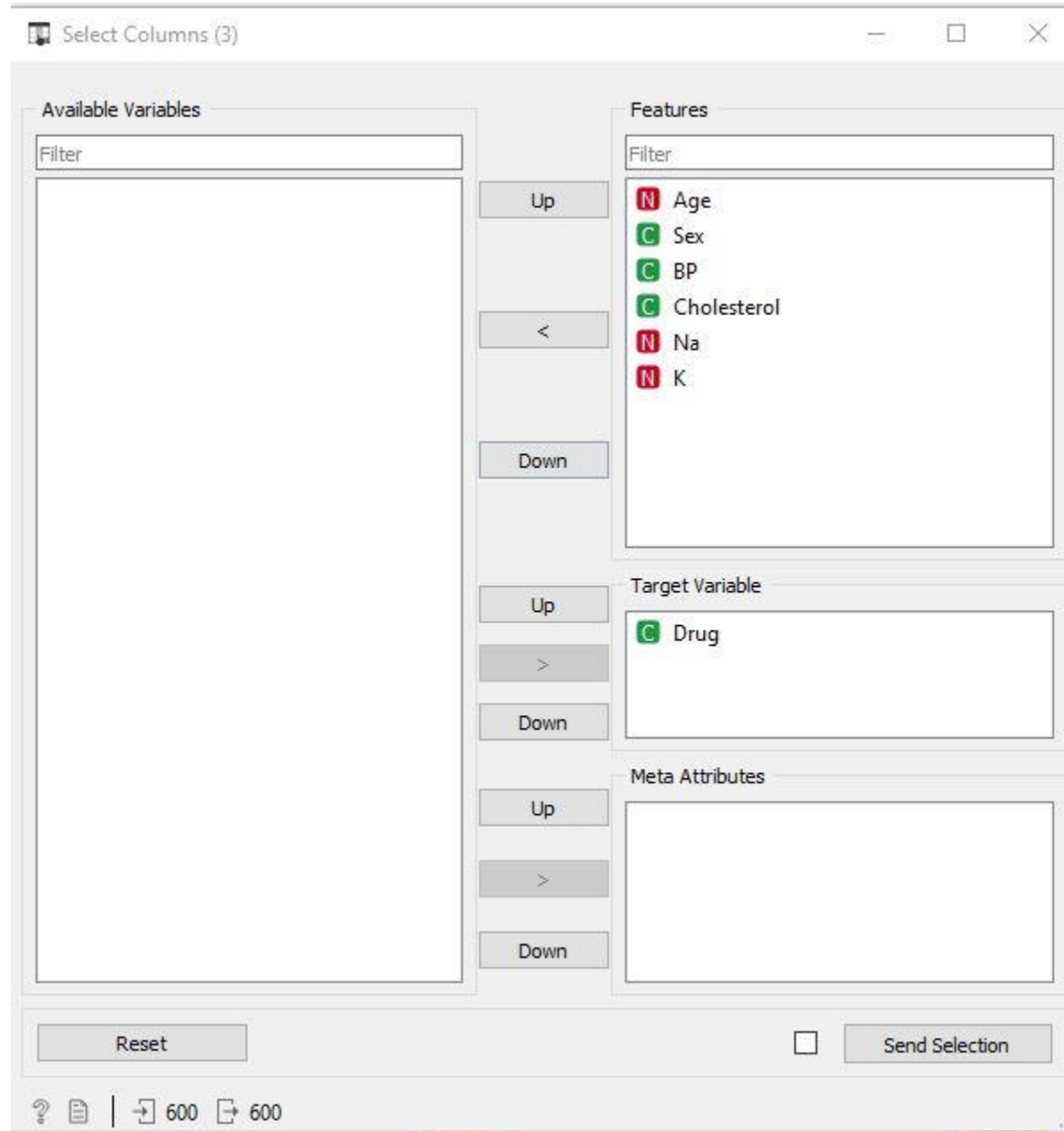
Restore Original Order

☒ Send Automatically

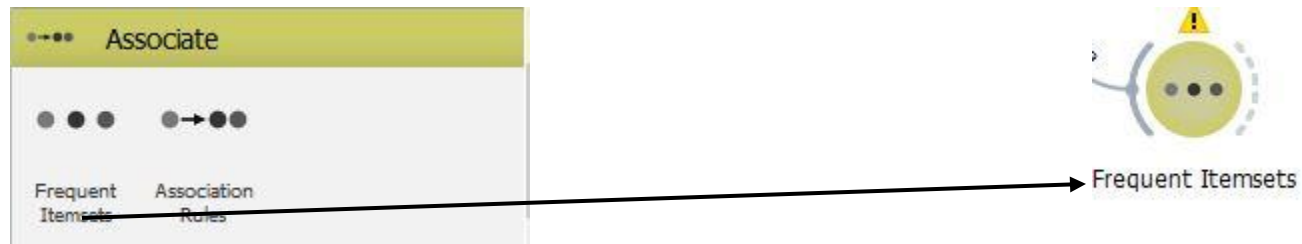
600

	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

. Luego arrastramos al lienzo un select column, luego lo unimos a la data set directamente. Y como se puede observar no hay variables disponibles, como en el anterior en este caso directamente se pueden usar las funciones de cada campo, y tomando en cuenta el análisis usamos el campo de drug como variable objetivo. Y siempre dejar el atributo por defecto. Pero en este caso no tiene un atributo por defecto eso quiere decir que no tiene un contenido de tipo texto.



. Regresamos al panel y luego, debemos arrastrar el módulo de Frequent Itemsets y unirlo con el select column, este módulo permite recopilar todos los datos que se incluyen como ítems, donde tiene un determinado valor cada ítem y de cuantos pertenecen a la data set, es un referente para poder crear la regla de asociación y hacer un análisis para el algoritmo. Se puede observar una advertencia en el módulo de itemsets, en este caso el dataset origen carga con variables continuas, eso quiere decir que esta cargando cada dato de la table de dichas variables.



. Al dar click en el módulo de itemsets, lo que muestra es la información de cada item de la tabla dividiéndolo por support y su porcentaje de análisis, donde del lado izquierdo se puede manipular de cuanto quiere que sea el mínimo de support y el máximo de ítems. En este caso dejaremos el mínimo de support en un 10%(se refiere a la cantidad de datos que se van a mostrar del origen de datos) y en numero máximo de ítemset en 10,000 y luego le damos click en Find Itemsets.

*** Frequent Itemsets

Info

Number of itemsets: 43
Selected itemsets: 0
Selected examples: 0

Expand all
Collapse all

Find itemsets

Minimal support: 10%
Max. number of itemsets: 10000

☐ Find Itemsets

Filter itemsets

Contains:

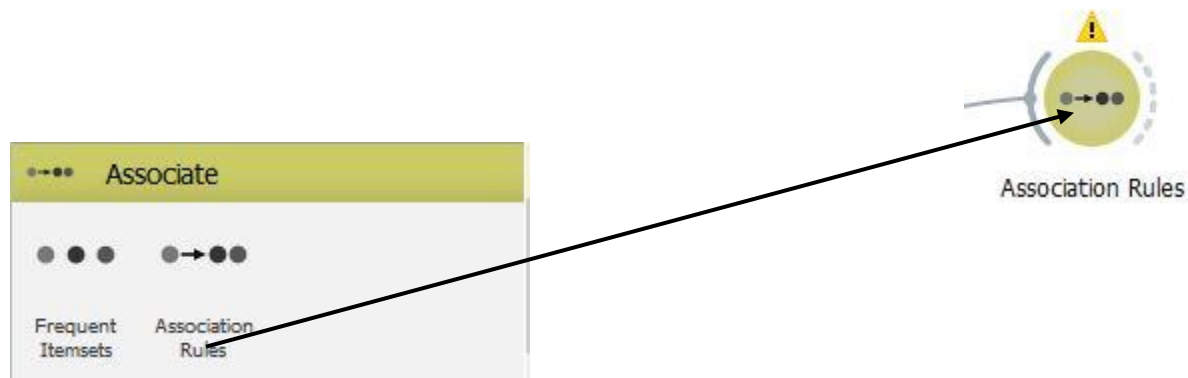
Min. items: 1 Max. items: 999

☒ Apply these filters in search

☒ Send Selection Automatically

Itemsets	Support	%
Sex=F	285	47.5
BP=HIGH	89	14.83
BP=LOW	102	17
BP=NORMAL	94	15.67
Cholesterol=HIGH	147	24.5
Drug=drugY	64	10.67
Cholesterol=NORMAL	138	23
Drug=drugY	67	11.17
Drug=drugX	78	13
Drug=drugY	131	21.83
Sex=M	315	52.5
BP=HIGH	125	20.83
Cholesterol=HIGH	67	11.17
BP=LOW	93	15.5
BP=NORMAL	97	16.17
Cholesterol=HIGH	162	27
Drug=drugY	67	11.17
Cholesterol=NORMAL	153	25.5
Drug=drugY	74	12.33
Drug=drugX	78	13
Drug=drugY	141	23.5
BP=HIGH	214	35.67
Cholesterol=HIGH	117	19.5
Cholesterol=NORMAL	97	16.17
Drug=drugA	70	11.67
Drug=drugY	92	15.33
BP=LOW	195	32.5
Cholesterol=HIGH	100	16.67
Cholesterol=NORMAL	95	15.83
Drug=drugY	95	15.83
BP=NORMAL	191	31.83
Cholesterol=HIGH	92	15.33
Cholesterol=NORMAL	99	16.5
Drug=drugX	106	17.67
Drug=drugY	85	14.17
Cholesterol=HIGH	309	51.5
Drug=drugY	131	21.83
Cholesterol=NORMAL	291	48.5

. Volvemos al panel y arrastramos al lienzo las reglas de asociación, y lo conectamos directamente al dataset de origen de esquelas. También podemos observa que hay una advertencia, donde también aparece que las variables son continuas, pero en este caso no afectaría las reglas como tal, ya que directamente toma los datos del origen para cada variable.



. Y al final podemos observar las reglas de asociación, en este caso el support mínimo se tomó de un 7%, mínimo de confidence de un 34%. Y el número máximo de reglas se puede observar que hay un valor muy grande esto quiere decir que es el valor de ítems que hay en el dataset y que cuantos va tomar para poder crear las reglas de asociación.

Info

Number of rules: 52
Filtered rules: 52
Selected rules: 0
Selected examples: 0

Find association rules

Minimal support: 7%
Minimal confidence: 34%
Max. number of rules: 10000
☐ Induce classification (itemset → class) rules

☐ Find Rules

Filter rules

Antecedent

Contains:
Min. items: 1 Max. items: 999

Consequent

Contains:
Min. items: 1 Max. items: 999
☒ Apply these filters in search

☒ Send Selection Automatically

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
0.270	0.524	0.515	1.019	0.999	-0.000	Cholesterol=HIGH	Sex=M
0.270	0.514	0.525	0.981	0.999	-0.000	Sex=M	Cholesterol=HIGH
0.255	0.526	0.485	1.082	1.001	0.000	Cholesterol=NORMAL	Sex=M
0.255	0.486	0.525	0.924	1.001	0.000	Sex=M	Cholesterol=NORMAL
0.245	0.476	0.515	0.922	1.002	0.000	Cholesterol=HIGH	Sex=F
0.245	0.516	0.475	1.084	1.002	0.000	Sex=F	Cholesterol=HIGH
0.230	0.474	0.485	0.979	0.998	-0.000	Cholesterol=NORMAL	Sex=F
0.230	0.484	0.475	1.021	0.998	-0.000	Sex=F	Cholesterol=NORMAL
0.208	0.584	0.357	1.472	1.113	0.021	BP=HIGH	Sex=M
0.208	0.397	0.525	0.679	1.113	0.021	Sex=M	BP=HIGH
0.195	0.379	0.515	0.693	1.062	0.011	Cholesterol=HIGH	BP=HIGH
0.195	0.547	0.357	1.444	1.062	0.011	BP=HIGH	Cholesterol=HIGH
0.170	0.523	0.325	1.462	1.101	0.016	BP=LOW	Sex=F
0.170	0.358	0.475	0.684	1.101	0.016	Sex=F	BP=LOW
0.167	0.513	0.325	1.585	0.996	-0.001	BP=LOW	Cholesterol=HIGH
0.165	0.340	0.485	0.656	1.069	0.011	Cholesterol=NORMAL	BP=NORMAL
0.165	0.518	0.318	1.524	1.069	0.011	BP=NORMAL	Cholesterol=NORMAL
0.162	0.508	0.318	1.649	0.967	-0.005	BP=NORMAL	Sex=M
0.162	0.453	0.357	1.360	0.935	-0.011	BP=HIGH	Cholesterol=NORMAL
0.158	0.487	0.325	1.492	1.004	0.001	BP=LOW	Cholesterol=NORMAL
0.157	0.492	0.318	1.492	1.036	0.005	BP=NORMAL	Sex=F
0.155	0.477	0.325	1.615	0.908	-0.016	BP=LOW	Sex=M
0.153	0.482	0.318	1.618	0.935	-0.011	BP=NORMAL	Cholesterol=HIGH
0.148	0.416	0.357	1.332	0.876	-0.021	BP=HIGH	Sex=F
0.112	0.573	0.195	2.692	1.091	0.009	BP=HIGH, Cholesterol=HIGH	Sex=M
0.112	0.414	0.270	1.321	1.160	0.015	Sex=M, Cholesterol=HIGH	BP=HIGH
0.112	0.536	0.208	2.472	1.041	0.004	Sex=M, BP=HIGH	Cholesterol=HIGH
0.097	0.598	0.162	3.247	1.139	0.012	BP=HIGH, Cholesterol=NORMAL	Sex=M
0.097	0.379	0.255	1.399	1.063	0.006	Sex=M, Cholesterol=NORMAL	BP=HIGH
0.097	0.464	0.208	2.328	0.957	-0.004	Sex=M, BP=HIGH	Cholesterol=NORMAL
0.085	0.510	0.167	2.850	1.074	0.006	BP=LOW, Cholesterol=HIGH	Sex=F

Data has continuous attributes which will be skipped.

. Como dato extra en Orange datamining se puede buscar el consecuente y el antecedente de los datos, la búsqueda mas precisa para que se pueda desarrollar mejor el algoritmo en específico. Un ejemplo serio, que en el antecedente sea el colesterol que estamos buscando y el consecuente normal, y entre el algoritmo toma los datos mas develados entre el colesterol y el BP en Normal.

Info

Number of rules: 2
Filtered rules: 2
Selected rules: 0
Selected examples: 0

Find association rules

Minimal support: 7%
Minimal confidence: 34%
Max. number of rules: 10000
☐ Induce classification (itemset → class) rules
☐ Find Rules

Filter rules

Antecedent
Contains: Cholesterol
Min. items: 1 Max. items: 999
Consequent
Contains: NORMAL
Min. items: 1 Max. items: 999
☒ Apply these filters in search

☒ Send Selection Automatically

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
0.165	0.340	0.485	0.656	1.069	0.011	Cholesterol=NORMAL →	BP=NORMAL
0.080	0.348	0.230	1.384	1.093	0.007	Sex=F, Cholesterol=NORMAL →	BP=NORMAL

Data has continuous attributes which will be skipped.

. También como parte opcional podemos arrastrar al diagrama, una distribución de datos y un ejemplo en este caso nos puede mostrar la frecuencia de Drug, que mostraría los tipos de droga en el dataset de Drugs.

