KIBERNUM
IT Academy

Módulo 6 – Aprendizaje de Máquina No Supervisado

Demo Orange

Especialización en Ciencia de Datos

Objetivos

- Utilizar los conceptos básicos de aprendizaje de máquinas no supervisado.
- Conocer los distintos tipos de algoritmos.
- Diferenciar entre supervisado y no supervisado.
- Casos de uso.

KIBERNUM

Contenido

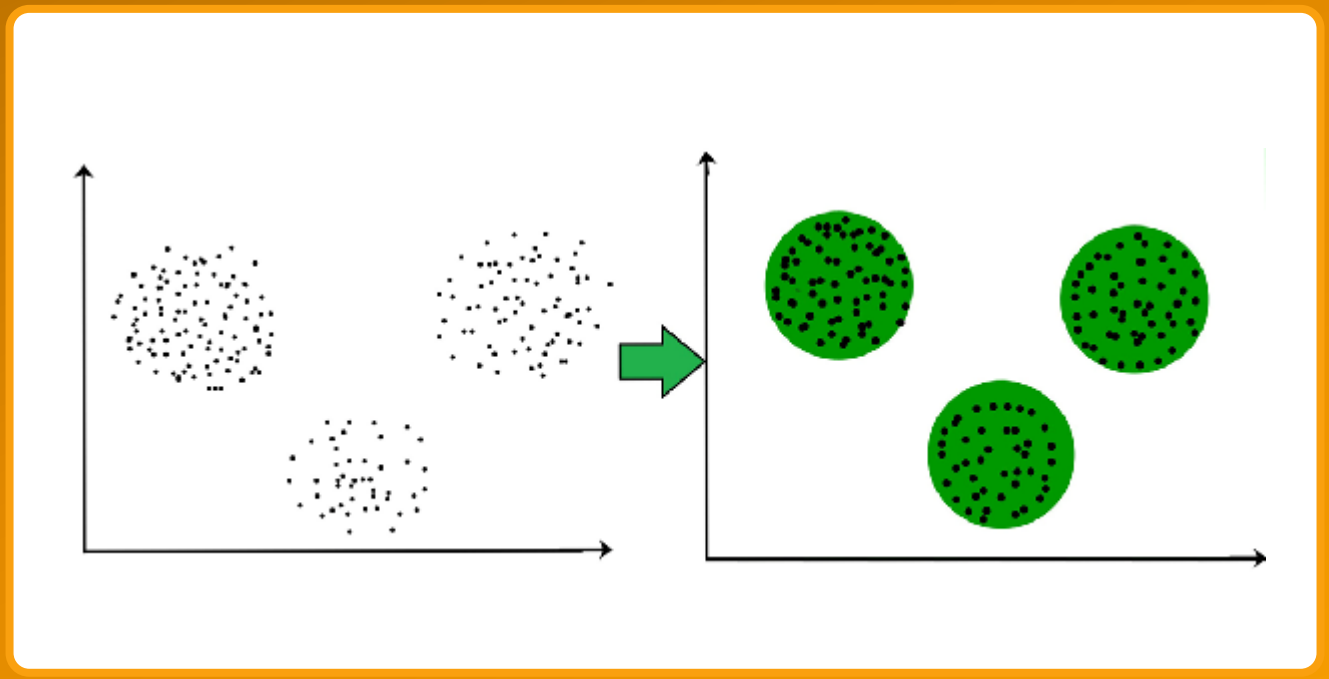
- Demostración Clusterización.
- Demostración Reducción Dimensionalidad.
- Demostración Detección Anomalías.

KIBERNUM

Demostración

Clusterización

Clustering, también conocido como agrupamiento, es una técnica de aprendizaje no supervisado en la que se utilizan algoritmos para identificar grupos o clústeres de objetos o datos similares. El objetivo principal del clustering es agrupar objetos similares juntos y separar objetos diferentes en grupos distintos, sin tener una clasificación previa de los datos.



El algoritmo de clustering, funciona al analizar las características de los datos, buscando patrones y similitudes en los valores. Estos patrones se utilizan para agrupar los objetos o datos en clusters o grupos.

Corresponde a técnicas de Machine Learning no-supervisado en donde a partir de datos no etiquetados, se le asigna una etiqueta.

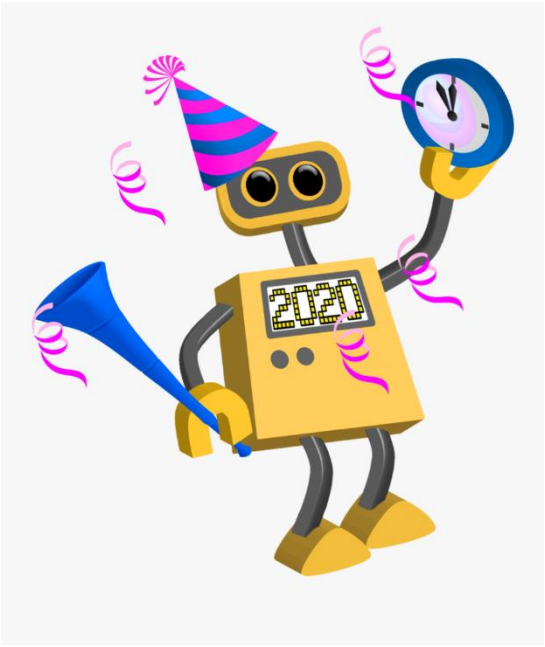
Reducción de Dimensionalidad

La reducción de dimensionalidad se utiliza para abordar varios problemas en el análisis de datos, incluyendo la complejidad computacional, la visualización de datos y la mejora del rendimiento de los modelos de aprendizaje automático.

Al reducir la cantidad de variables, es posible **disminuir la complejidad computacional de ciertas tareas**, como la clasificación y la agrupación de datos. Además, la reducción de dimensionalidad puede **ayudar a visualizar datos en espacios de menor dimensión**, lo que puede ayudar a comprender mejor las relaciones entre las variables. Finalmente, la reducción de dimensionalidad puede mejorar el rendimiento de los modelos de aprendizaje automático, al **reducir el ruido en los datos y mejorar la generalización**.

Menor Complejidad

Simplicidad



Mejor Generalización

Mejor Visualización

KIBERNUM

Caso Wines

El siguiente dataset, corresponde a mediciones de propiedades químicas de vinos procedentes de 3 cultivos distintos de un área específica de Italia. El dataset, contiene los resultados de 178 mediciones de 13 variables químicas medidas para cada muestra. A continuación, una muestra:

Alcohol	Malic Acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	80/OD315 of diluted ı	Proline
14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.040	3.92	1065
13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.050	3.40	1050
13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.030	3.17	1185
14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.8	0.860	3.45	1480
13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.040	2.93	735
14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.050	2.85	1450

`Alcohol`
Alcohol

`Malic`
Malic acid

`Ash`
Ash

`Alcalinity`
Alcalinity of ash

`Magnesium`
Magnesium

`Phenols`
Total phenols

`Flavanoids`
Flavanoids

`Nonflavanoids`
Nonflavanoid phenols

`Proanthocyanins`
Proanthocyanins

`Color`
Color intensity.

`Hue`
Hue

`Dilution`
D280/OD315 of diluted wines.

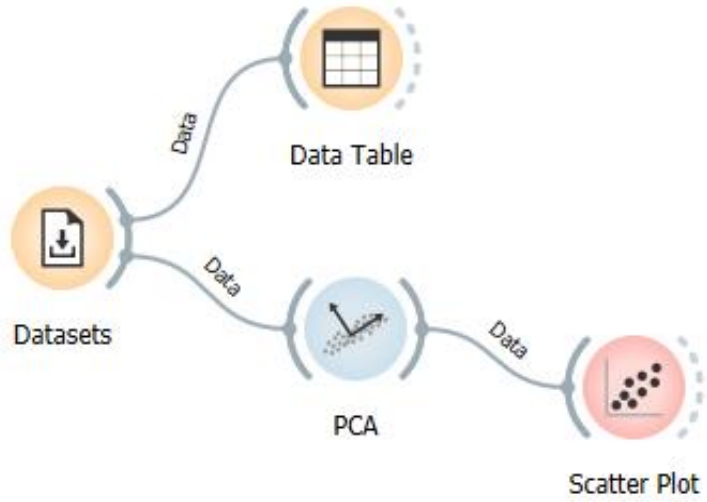
`Proline`
Proline

Asuncion, A. & Newman, D.J. (2007). *UCI Machine Learning Repository* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Caso Wines

Bajo el supuesto que cada cultivo tuvo un tratamiento homogéneo pero característico, es posible pensar que cada cultivo tenga mediciones similares. De esta forma, se podría agrupar dichos datos y determinar la cantidad de cultivos y a qué cultivo pertenece cada medición.

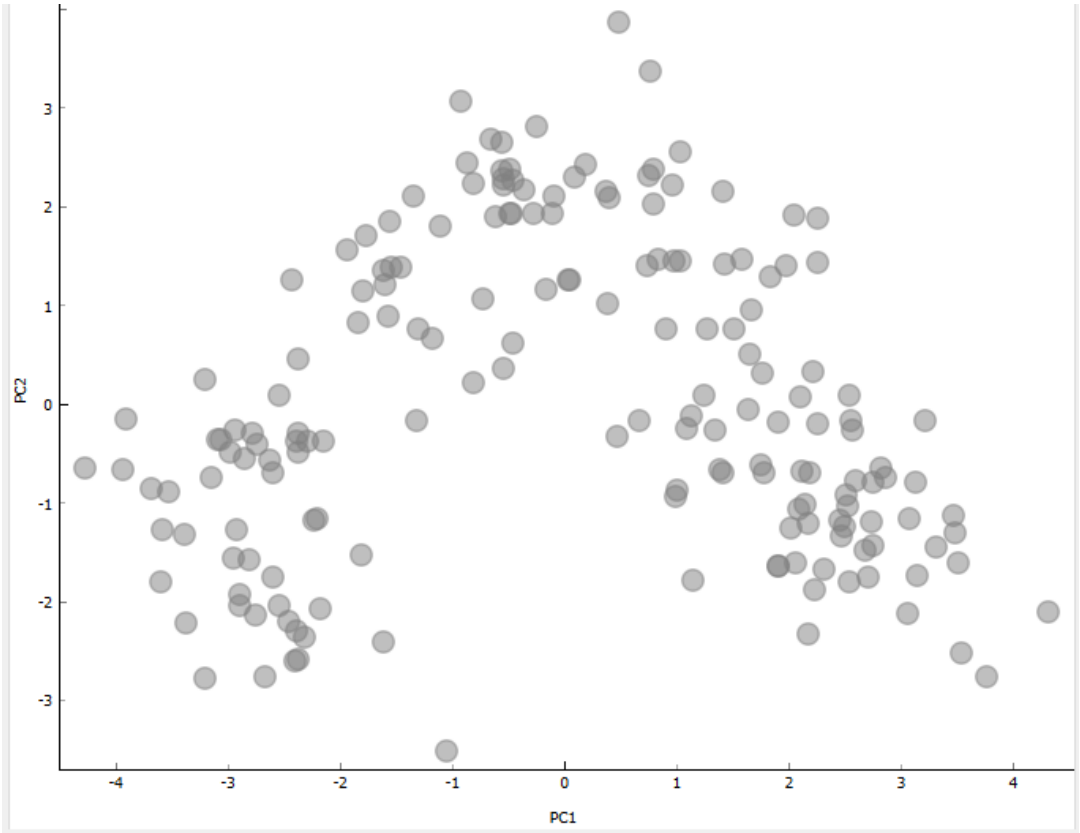
El principal problema, es que 13 features son difíciles de explorar. Es por esto, que primero **realizaremos una tarea de Reducción de Dimensionalidad de los datos**. La idea es disminuir la dimensionalidad tratando de perder la menor información manteniendo la esencia de los datos. Perderemos algo de exactitud, pero ganaremos en visualización. Lo anterior, es muy simple de realizar en Orange.



KIBERNUM

Caso Wines

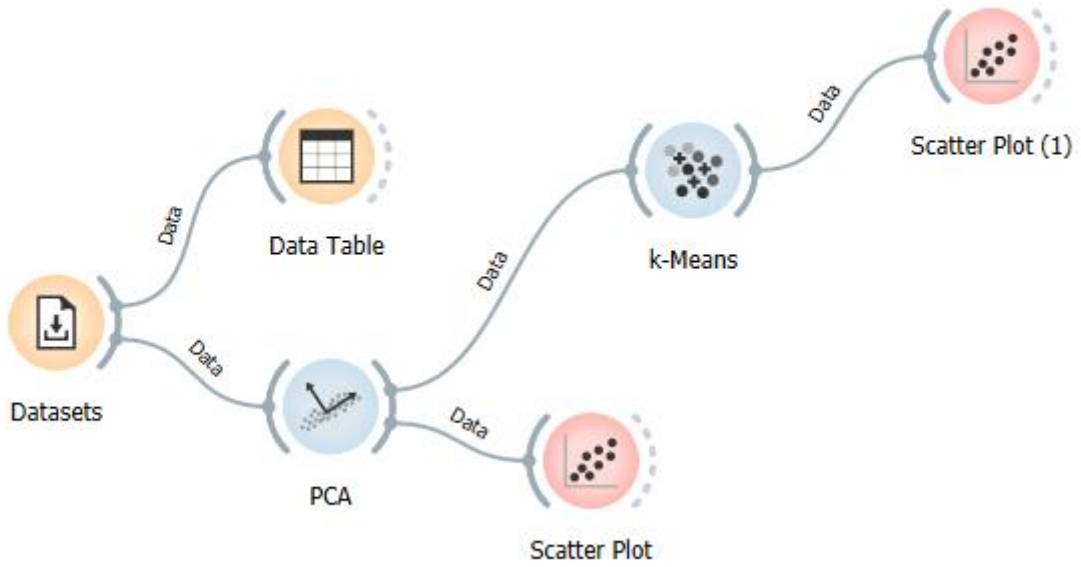
El siguiente diagrama, nos muestra nuestro set de datos original de 13 dimensiones, proyectado en un espacio de sólo 2 dimensiones (componentes principales). Nótese que se aprecian grupos de datos similares, es decir, datos próximos unos de otros.

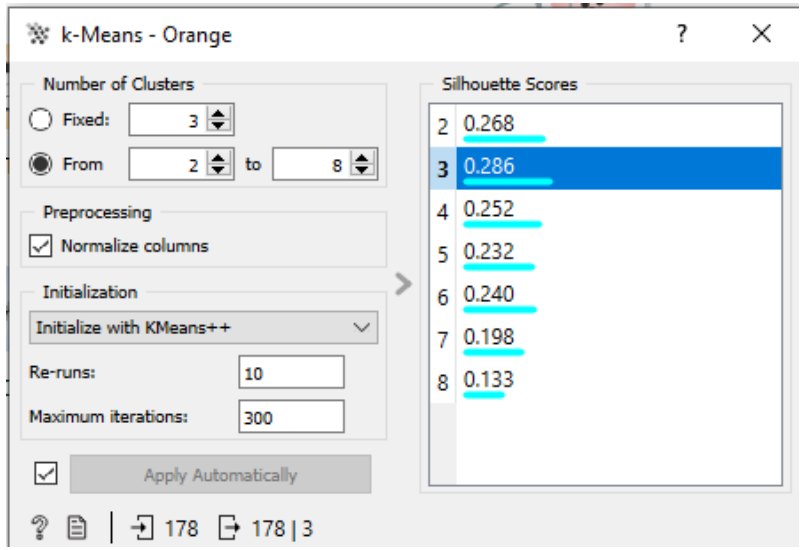


KIBERNUM

Caso Wines

Ahora ejecutaremos una tarea de clusterización, para que el algoritmo pueda encontrar cuáles son los grupos o clusters con instancias consideradas similares. Para esto, incorporaremos un algoritmo clusterizador llamado K-Means, dejaremos que haga su trabajo y veremos el resultado.

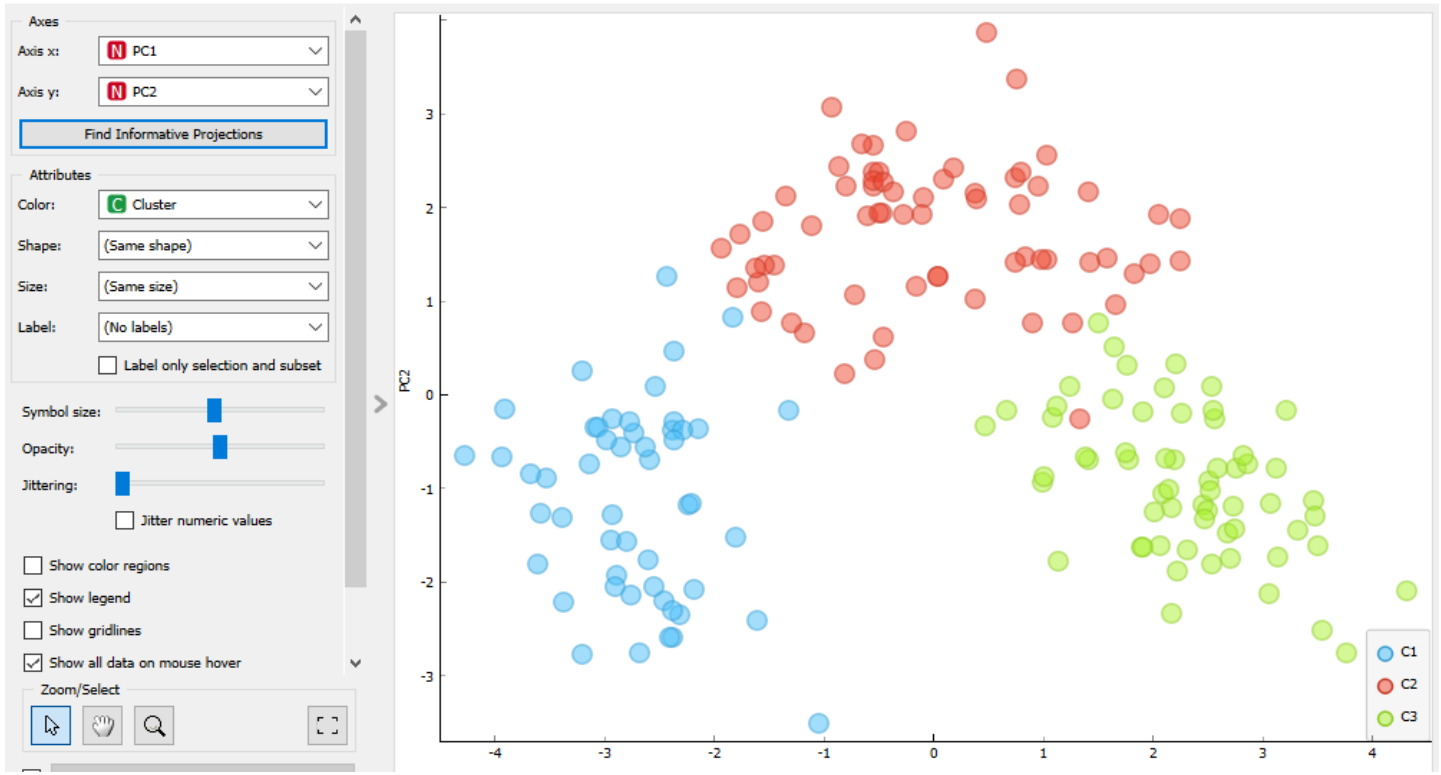




KIBERNUM

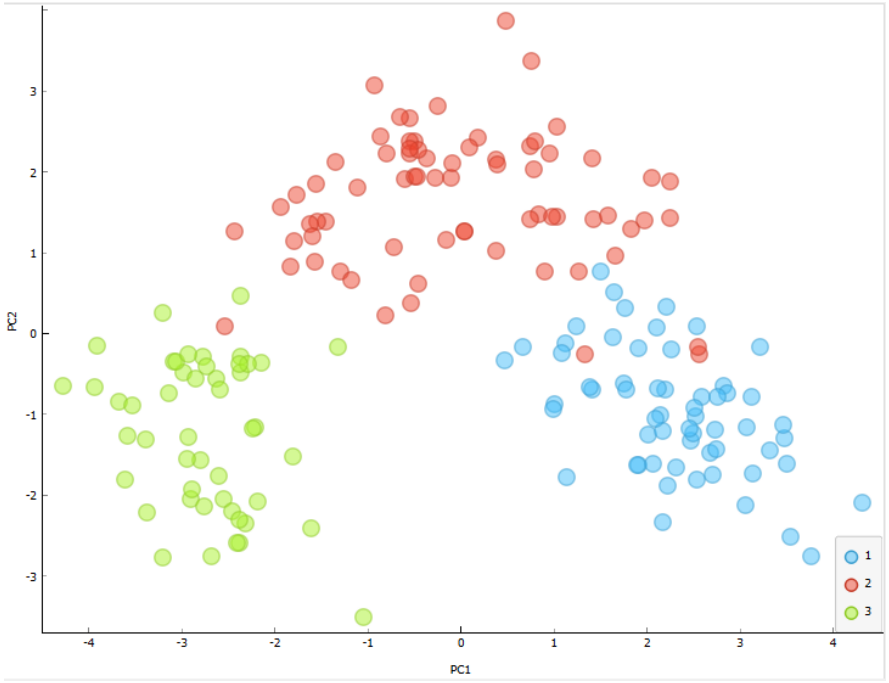
Caso Wines

Ahora visualizaremos los clusters encontrados por K-Means. Nótese que fueron identificados 3 clusters, por lo tanto, podríamos pensar que los datos corresponden a 3 cultivos distintos. Cada cultivo estaría representado en un cluster identificado con un color distintivo.

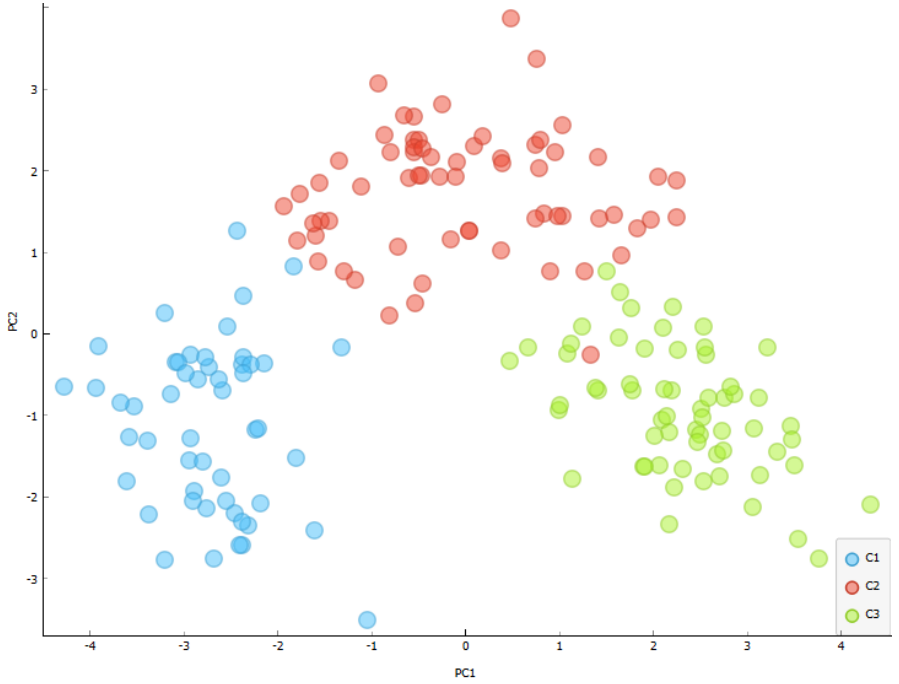


Caso Wines

Pues bien, en aprendizaje no supervisado, no podemos comparar contra las etiquetas reales puesto que éstas no son conocidas. Sin embargo, en el dataset utilizado sí se poseían las etiquetas de cada cultivo, por lo tanto, vamos a comparar qué tan certero fue el procedimiento no supervisado con la realidad. Como se aprecia, hay una gran similitud entre los clusters encontrados de forma no supervisada y los valores reales.



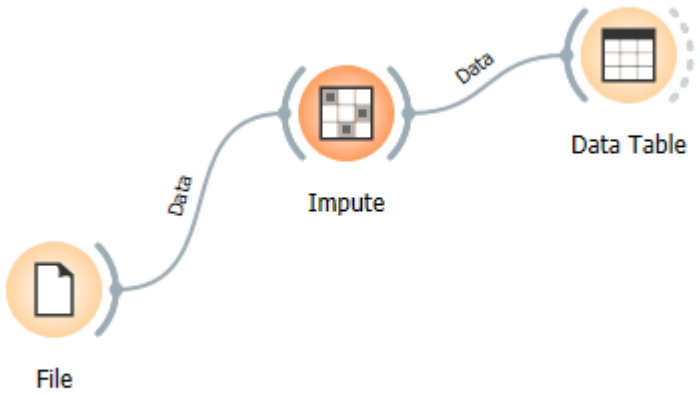
Cultivos reales

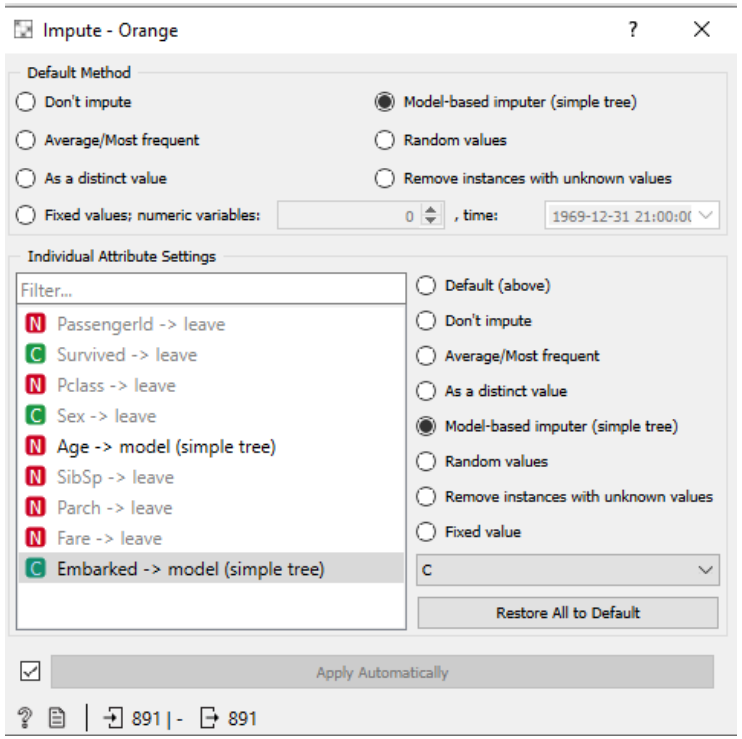


Clusters encontrados

Caso Titanic

Ahora haremos un poco de aprendizaje no supervisado sobre el dataset del Titanic. Para eso, vamos a leer el archivo de Titanic que hemos trabajado en los módulos anteriores. Realizaremos algunas imputaciones en las columnas de Age y Embarked con el imputador algorítmico.





KIBERNUM

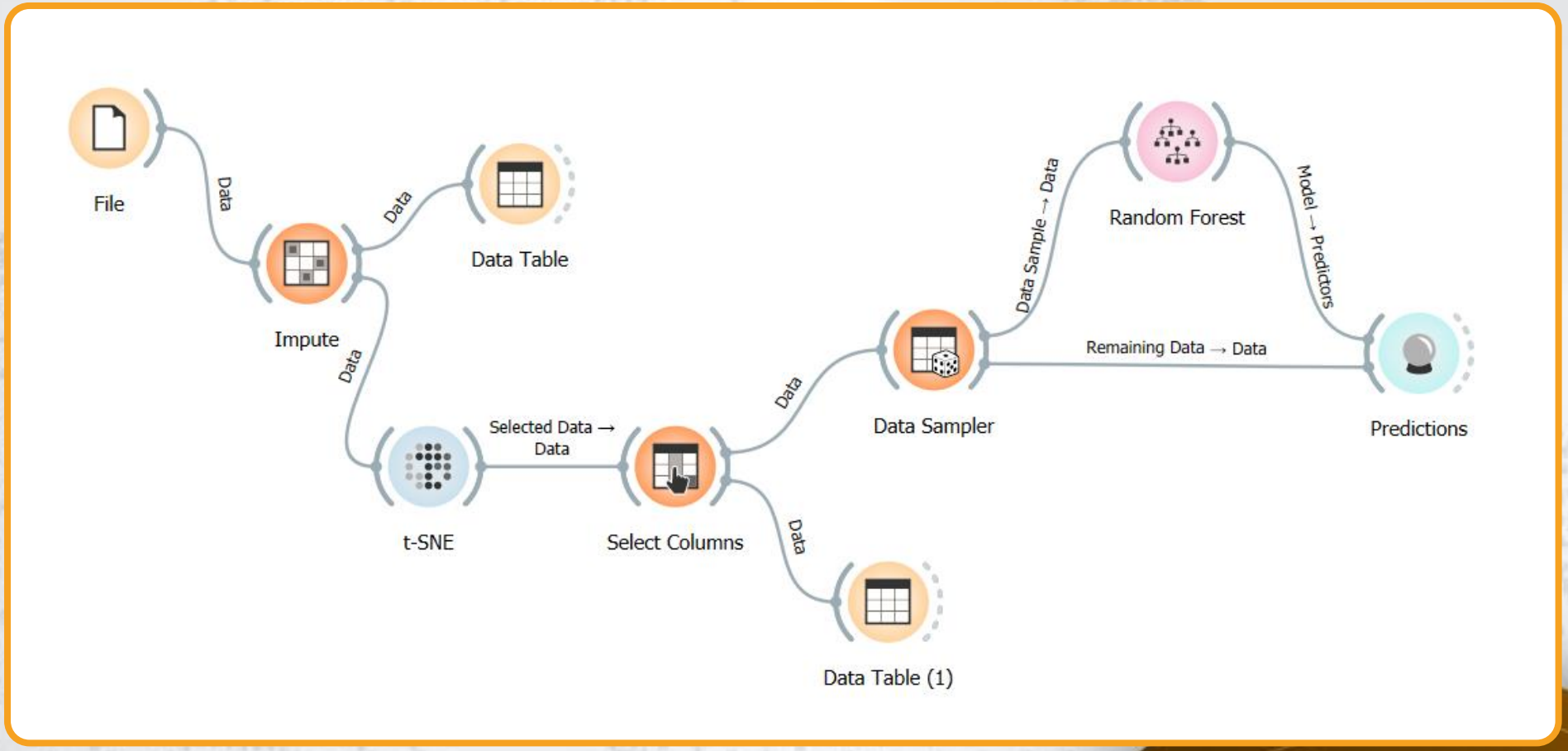
Caso Titanic

Realizaremos un proceso de disminución de dimensionalidad con el algoritmo t-SNE y visualizaremos los datos en un espacio de dos dimensiones. En este caso, al hacer la proyección a dos dimensiones, se puede ver que ahora es más fácil diferenciar a los sobrevivientes de los no sobrevivientes.

KIBERNUM

Caso Titanic

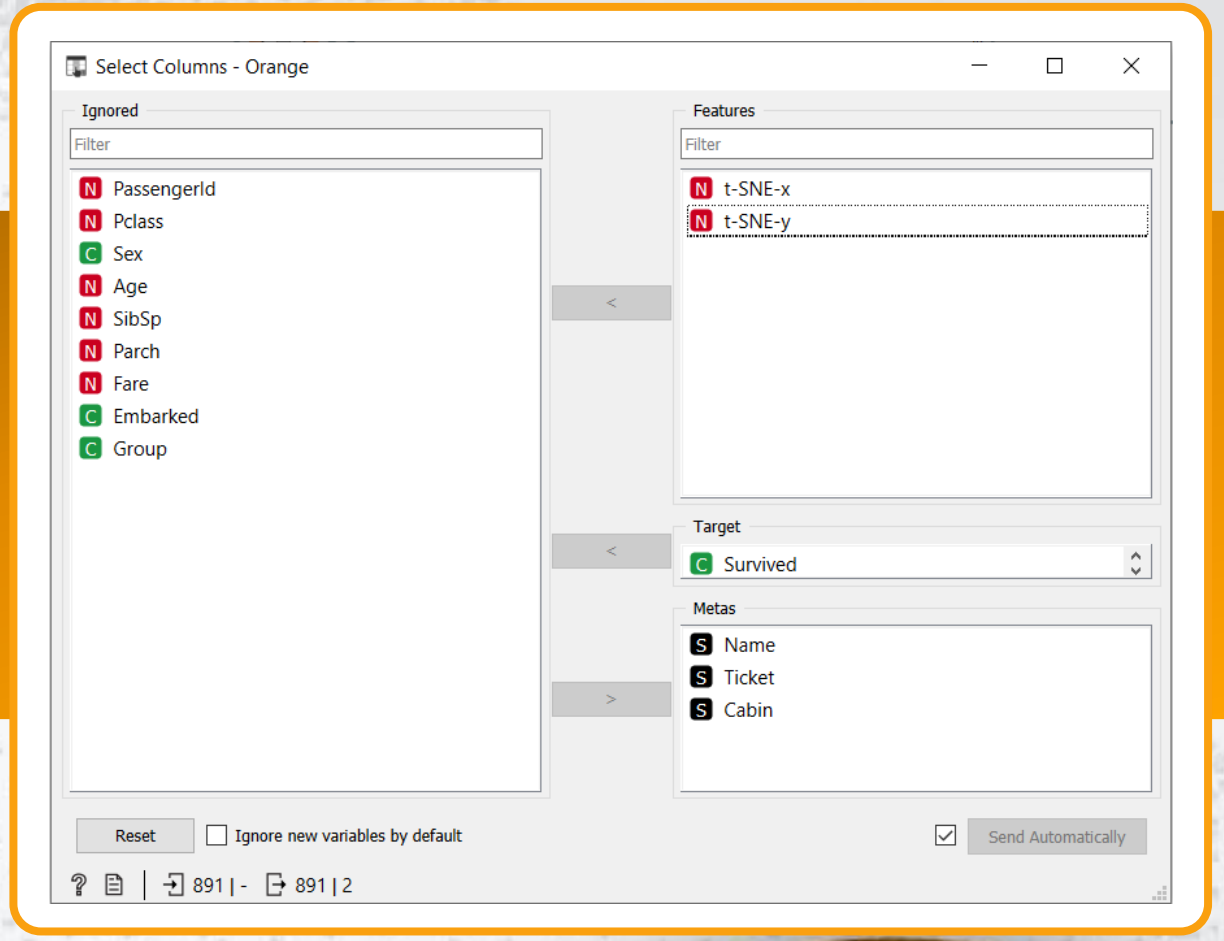
Lo que haremos ahora, será aplicar una tarea de clasificación con un algoritmo de aprendizaje supervisado de acuerdo al siguiente flujo.



KIBERNUM

Caso Titanic

- Nótese que agregamos un paso de selección de columnas en donde hemos dejado para el modelo supervisado los valores en el nuevo espacio de dos dimensiones reducido por t-SNE. La variable objetivo sigue siendo Survived.



KIBERNUM

Caso Titanic

Estos son los resultados obtenidos luego de incorporar la tarea no supervisada de reducción de dimensionalidad. Accuracy=99,3%

Predictions - Orange

Show probabilities for: Classes in data | Restore Original Order

	Random Forest	Survived	Name	Ticket	Cabin	t-SNE-x	t-SN
1	1.00 : 0.00 → 0	0	Taussig, Mr. Emil	110413	E67	-13.0251	-10.7555
2	1.00 : 0.00 → 0	0	Andersson, Mrs...	347082	?	6.64736	5.52732
3	0.00 : 1.00 → 1	0	Allison, Miss. H...	113781	C22 C26	15.5689	0.441987
4	0.25 : 0.75 → 1	1	Sundman, Mr. J...	STON/O 2. 310...	?	-1.34421	-21.5365
5	1.00 : 0.00 → 0	0	Bateman, Rev. R...	S.O.P. 1166	?	-15.5944	-6.7635
6	1.00 : 0.00 → 0	0	Morley, Mr. He...	250655	?	-24.5649	-7.5068
7	1.00 : 0.00 → 0	0	Guggenheim, ...	PC 17593	B82 B84	11.7678	-26.3745
8	0.00 : 1.00 → 1	1	Thayer, Mr. Joh...	17421	C70	15.02	-17.7792
9	1.00 : 0.00 → 0	0	Somerton, Mr. ...	A.5. 18509	?	-21.1737	5.10572
10	1.00 : 0.00 → 0	0	Ford, Miss. Robi...	W./C. 6608	?	2.32149	10.8756
11	0.00 : 1.00 → 1	1	Andersson, Mr. ...	350043	?	0.653264	-20.9332
12	1.00 : 0.00 → 0	0	Pasic, Mr. Jakob	315097	?	-29.7351	3.10679
13	1.00 : 0.00 → 0	0	Lam, Mr. Len	1601	?	-28.4668	-5.16282

Show perfomance scores | Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.993	0.993	0.993	0.993	0.993

267 | 267 | 1×267

Actual \ Predicted	0	1	Σ
0	167	2	169
1	0	98	98
Σ	167	100	267

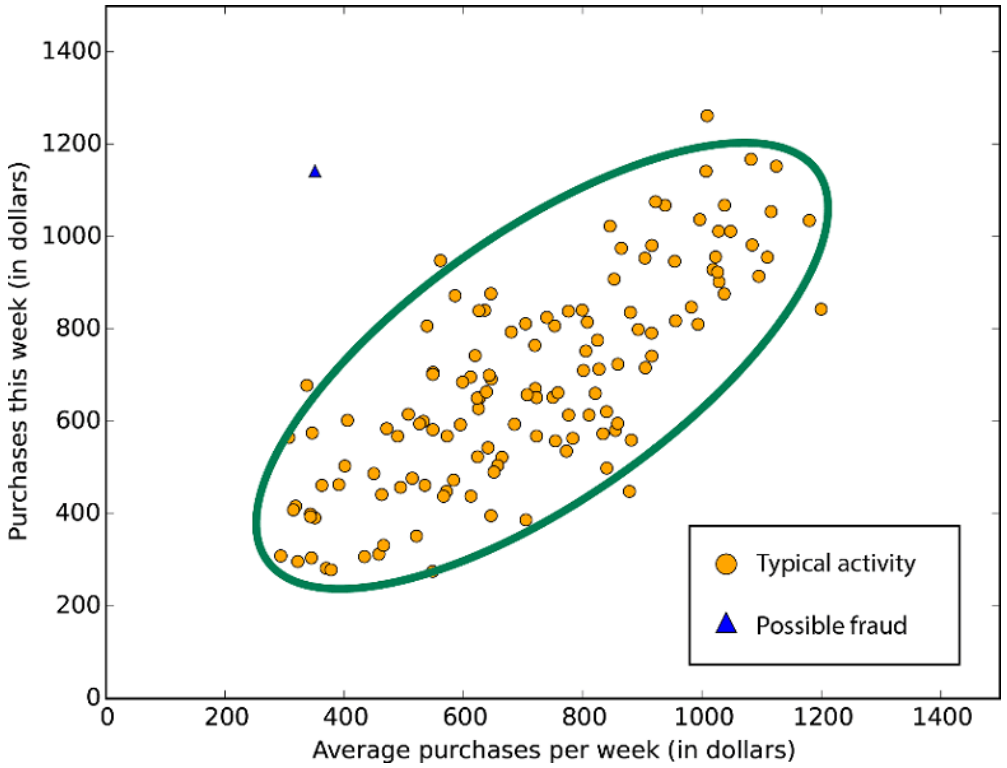
KIBERNUM

Detección de Anomalías

La detección de anomalías es un proceso de identificación de patrones inusuales o atípicos en datos, que difieren significativamente de la mayoría de los demás datos en un conjunto. Las anomalías también se conocen como "outliers" o valores atípicos.

La detección de anomalías es un campo importante en la minería de datos y el aprendizaje automático, y se utiliza en diversas aplicaciones, como la detección de fraudes en transacciones financieras, el monitoreo de sistemas informáticos para detectar intrusiones, la detección de fallas en equipos industriales y la vigilancia de la salud en pacientes para identificar patologías.

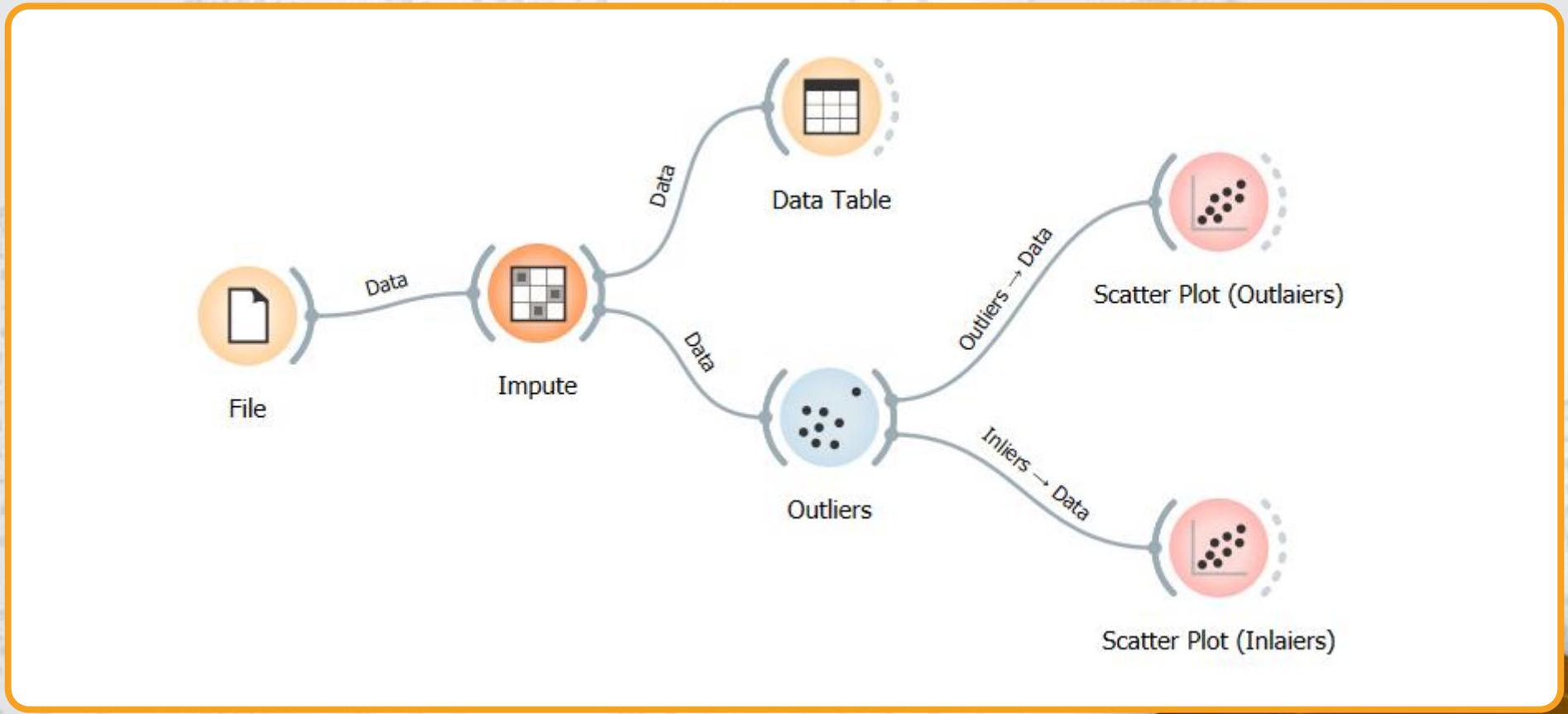
Los métodos de detección de anomalías pueden ser supervisados o no supervisados. En el enfoque no supervisado, se busca identificar patrones inusuales sin la ayuda de etiquetas de clase previas. En el enfoque supervisado, se entrena un modelo con ejemplos etiquetados de datos normales y luego, se utilizan estos modelos para identificar datos anómalos.



KIBERNUM

Detección de Anomalías

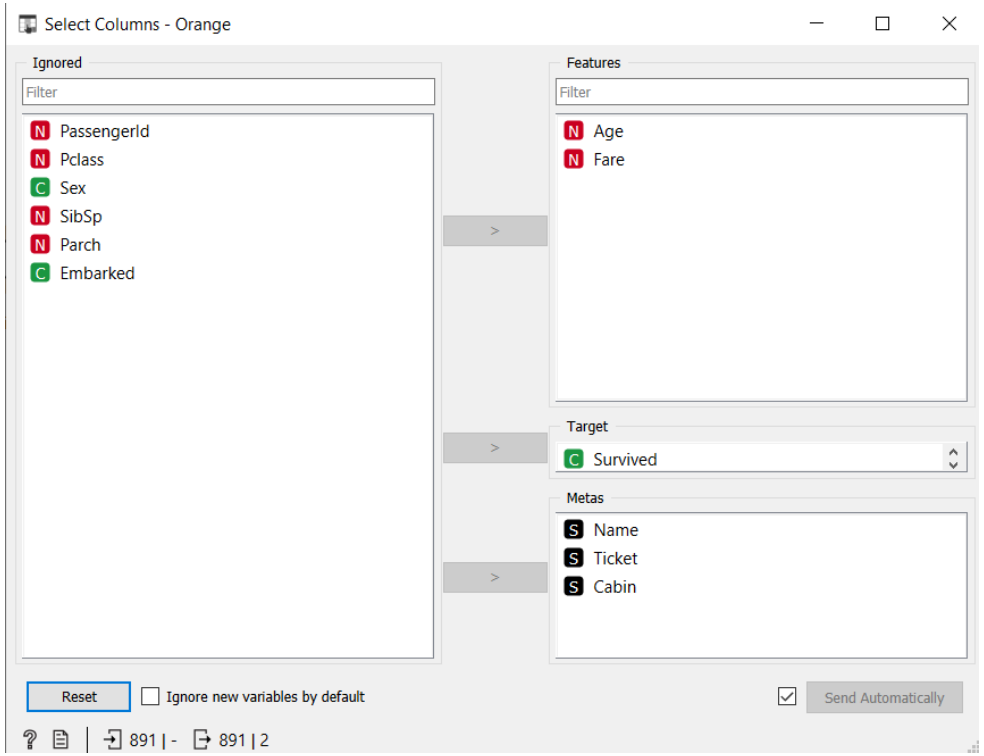
Ahora, analizaremos el dataset Titanic para detectar outliers en los datos. A continuación, se presenta el flujo en donde se ha realizado una tarea de detección de outliers.

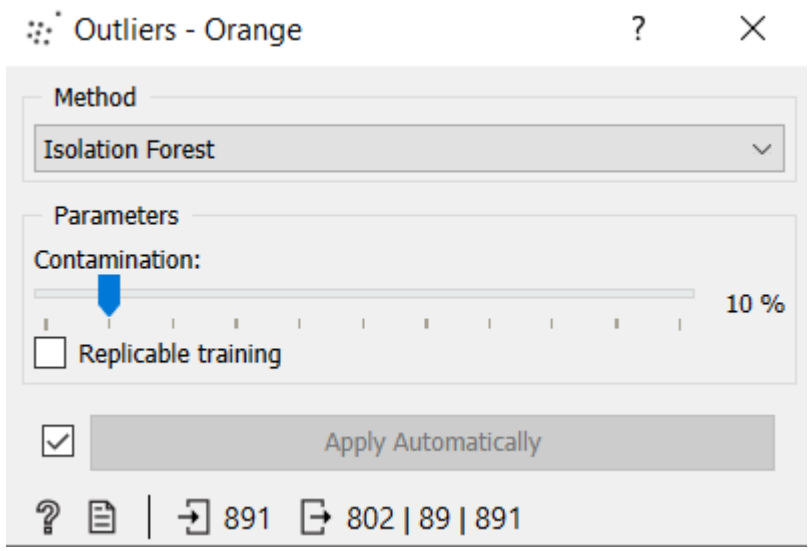


KIBERNUM

Detección de Anomalías

Para este ejemplo, hemos seleccionado sólo las columnas Age y Fare. Y se ha utilizado el algoritmo Isolation Forest para la detección de anomalías.

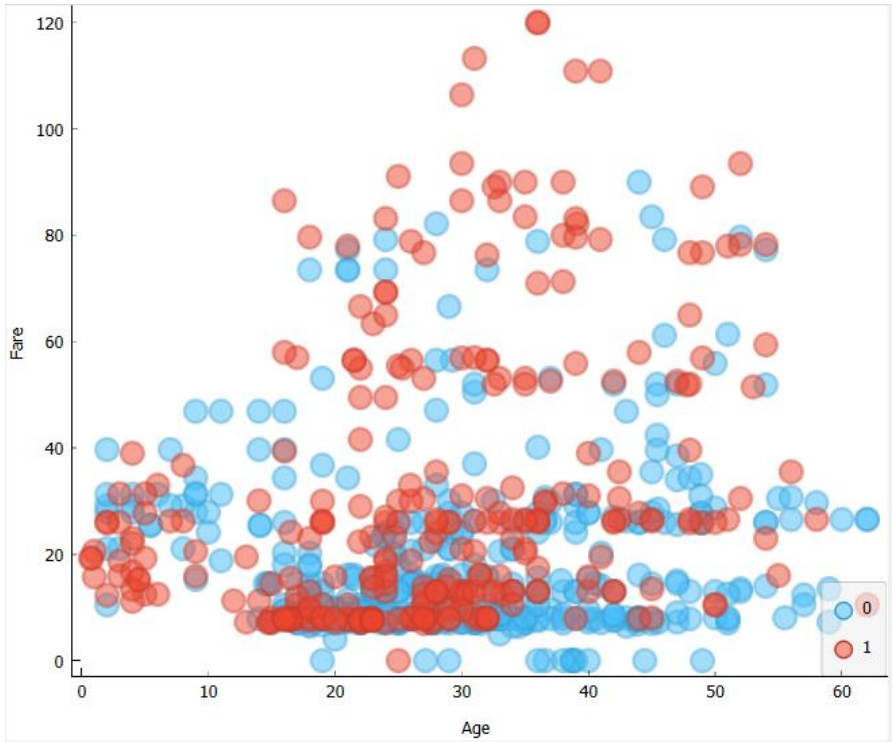




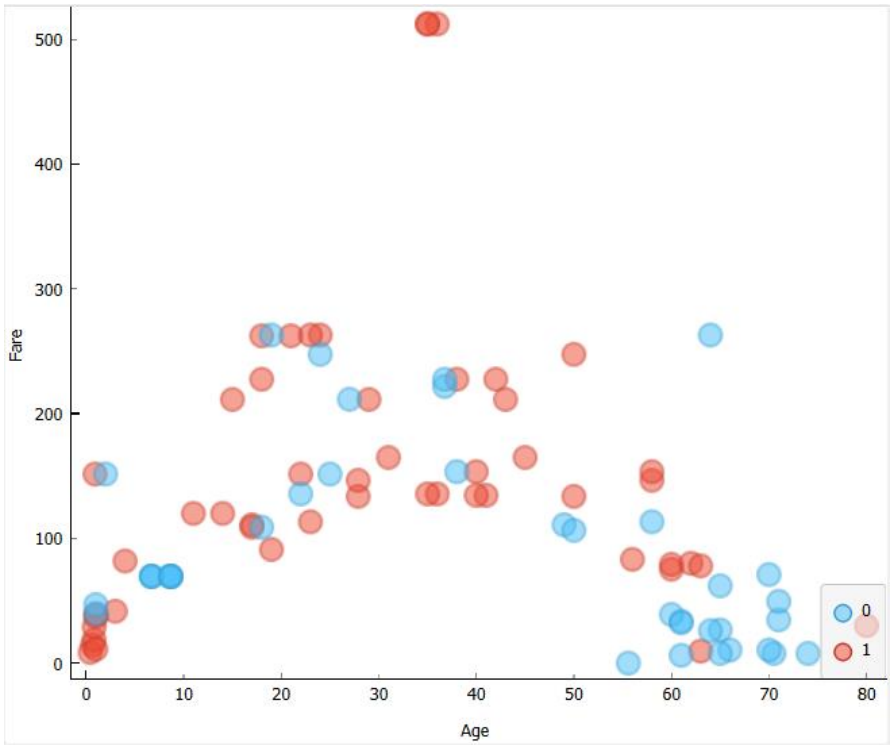
KIBERNUM

Detección de Anomalías

Acá se puede apreciar los puntos que son considerados inlaiers y los que son considerados outlaiers de acuerdo al algoritmo.



Inliers



Outliers

KIBERNUM

Dudas y consultas

KIBERNUM

Fin Presentación