Módulo 8

8.2 ¿Cómo resolver un problema con KNIME?

Por **María Martínez Rojas**
Profesora Titular en CA, Universidad de Granada

Por **José Manuel Soto Hidalgo**
Profesor Titular en ICAR, Universidad de Granada

1. INTRODUCCIÓN

En módulos anteriores se han mostrado ejemplos de cómo resolver un problema de Ciencia de Datos desde la perspectiva de las dos ramas principales: aprendizaje supervisado y aprendizaje no supervisado (Módulos 4, 5 y 6). Con el principal objetivo de mostrar la capacidad que ofrece la herramienta KNIME para resolver un problema de ciencia de datos, en esta cápsula se diseñará un flujo de datos que representa de manera general el ciclo de vida de Ciencia de Datos (ver Módulo 3): lectura de datos, manipulación de datos, exploración de datos, análisis de datos, medidas de calidad y exportación e informes.

El flujo de datos consta de cinco partes bien diferenciadas que marcan las principales etapas del ciclo de vida. Se comienza con nodos que permiten la lectura de datos y su preparación o manipulación. Posteriormente se muestran nodos de exploración visual de los datos para realizar un análisis previo de los datos y se realiza el análisis en sí basado en la generación de modelos y su validación. Finalmente se extraen conclusiones mediante informes o exportación de resultados.

La Figura 1 muestra el flujo de datos a desarrollar en esta cápsula. En concreto, se va a construir un árbol de decisión con el conjunto de datos iris, se va a leer datos a partir de un .csv, se van a manipular los datos y visualizar, así como se van a validar los modelos generados con el árbol de decisión.

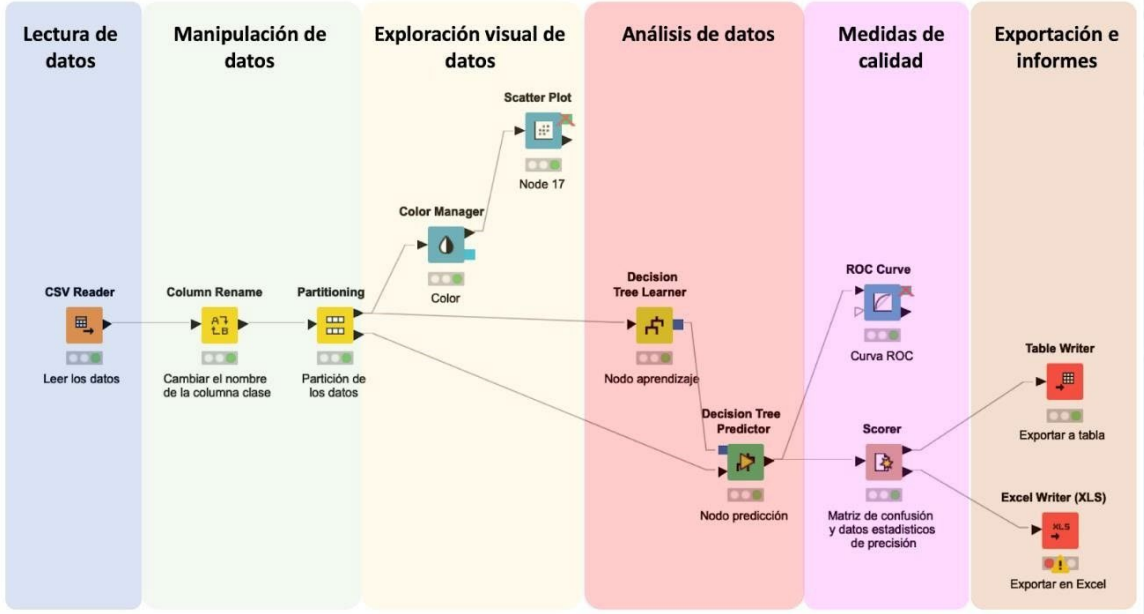


Figura 1 Flujo de datos con el conjunto Iris

2. DISEÑANDO EL FLUJO DE DATOS

Para comenzar es necesario crear un nuevo espacio de trabajo. Esto se puede hacer de dos formas (Figura 2):

- Haciendo clic en "Nuevo" en el panel de la barra de herramientas en la parte superior de KNIME
- Haciendo clic derecho en una carpeta de su espacio de trabajo local en el Explorador KNIME.

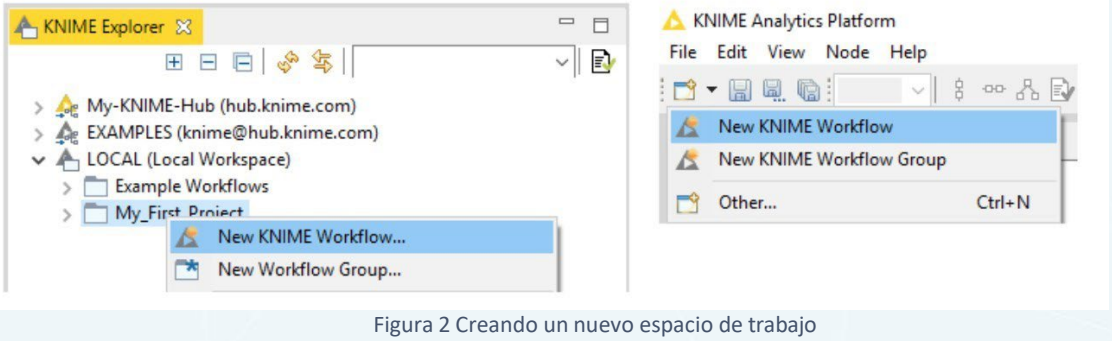


Figura 2 Creando un nuevo espacio de trabajo

Como se ha comentado anteriormente, se va a utilizar el conjunto de datos “iris” para ilustrar el flujo de datos. Iris consiste en 50 muestras de cada una de tres especies de Lirio (setosa, virginica y versicolor). Para cada muestra se midieron cuatro características: la longitud y la anchura de los sépalos y pétalos, en centímetros.

El conjunto está representado en formato *.csv*, por lo que el primer nodo que se necesita es un nodo que permita leer este tipo de archivo. Para ello, nos dirigimos al repositorio de nodos en la sección de IO → Leer (Read). También podemos escribir directamente el nodo a través del buscador disponible en la parte superior.

Como se puede observar en la Figura 3, KNIME ofrece diversos nodos para la lectura de ficheros de diferentes tipos: ARFF, Table, PMML, Excel, etc. En nuestro caso, seleccionamos el nodo “CSV Reader” que permite leer ficheros .csv. Para usar el nodo en nuestro flujo de trabajo se puede arrastrar directamente desde el repositorio y soltarlo en el espacio de trabajo o hacer doble clic sobre el nodo en el repositorio y aparecerá automáticamente en el editor de flujo de trabajo.

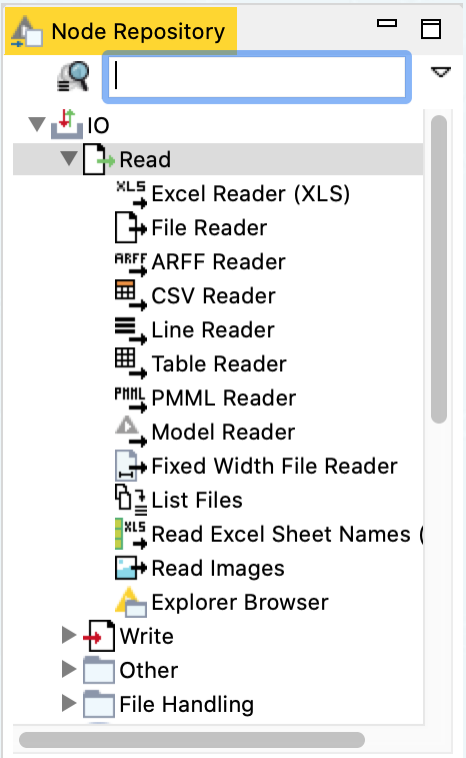


Figura 3 Nodos de lectura

A continuación, es necesario configurar este nodo y para ello tenemos tres opciones: pulsamos la tecla F6, haremos doble clic en el nodo o hacemos clic con el botón derecho y seleccionando "Configurar ..." como se muestra en la Figura 4. Como se puede ver en la imagen, en este menú además de configurar el nodo también se pueden realizar otras tareas: ejecutarlo, ver las salidas, editar el nodo, mostrar los datos de los puertos, etc.

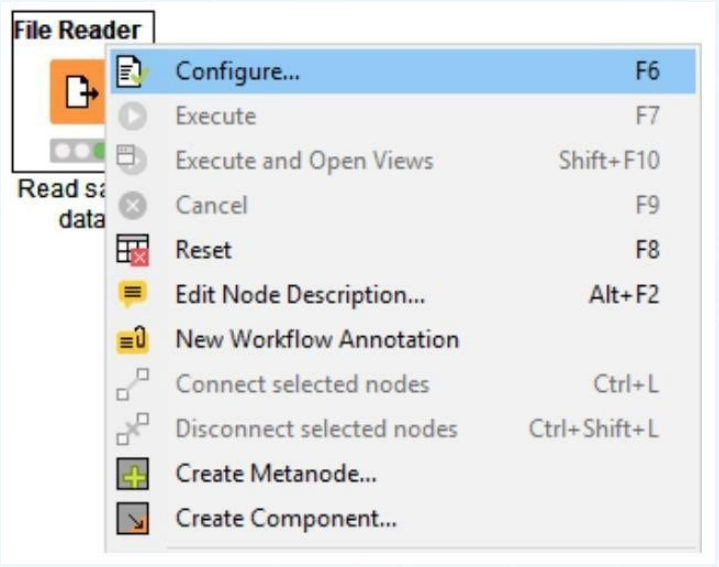


Figura 4 Configurar un nodo

Una vez seleccionada la opción de configurar, aparece un cuadro de diálogo de configuración (Figura 5) donde definiremos la ruta del archivo donde se encuentran los datos haciendo clic en el botón "Examinar" (Browse). En esta ventana también es posible configurar otras opciones disponibles y obtener una vista previa de los datos. Por ejemplo, en nuestro caso, deseleccionamos la opción de “Has column header” y “Has row header” (tiene encabezado de columna y fila) porque nuestro conjunto de datos no lo tiene. Una vez configurado, se hace clic en el botón de “ok” y el nodo aparecerá de color amarillo (configurado, pero no ejecutado). Por tanto, se puede ejecutar bien con la tecla F7, o pinchando en el botón derecho y seleccionar la opción de ejecutar o directamente desde el botón correspondiente en la barra de menú de la parte superior.

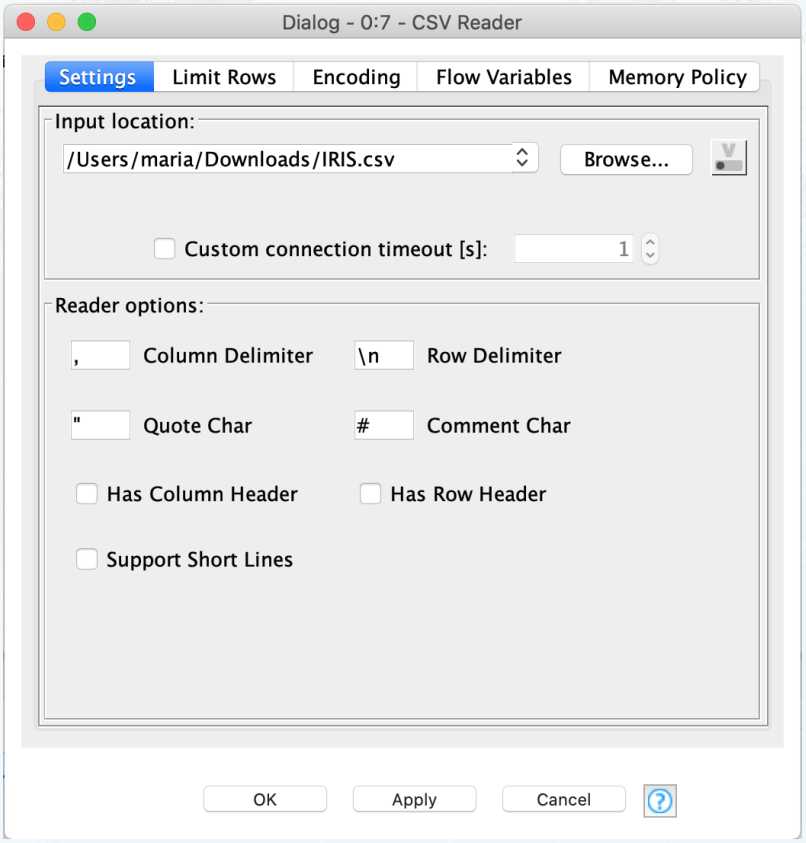


Figura 5 Configuración del nodo de entrada

Una vez ejecutado correctamente (semáforo del nodo estará en verde), podemos visualizar los datos que este nodo ha leído. Los datos se encuentran en modo de tabla, por lo que, para leerlos, hacemos clic con el botón derecho y seleccionamos la opción "File Table" o seleccionamos el botón en la barra superior de menú. En la Figura 6 se puede observar la tabla con los datos del conjunto de datos "iris". En la parte superior se puede ver que el número de filas del conjunto es de 150 (correspondiente a 150 muestras) y que hay 5 columnas (correspondiente a las 4 variables y la clase):

- RowID: identificación de la fila
- Col0: Largo de sépalo
- Col1: Ancho de sépalo
- Col2: Largo de pétalo
- Col3: Ancho de pétalo
- Col4: Especie

Como se puede ver en la Figura 6, en el nombre de cada columna se define el tipo de dato que está contenido en cada una de ellas. Por ejemplo, Las columnas 0, 1, 2, y 3 contienen valores con decimales (D=double) mientras que la columna 4 contiene cadena de texto (S=string).

File Table - 0:7 - CSV Reader

File Hilite Navigation View

Table "IRIS.csv" – Rows: 150 | Spec – Columns: 5 | Properties | Flow Variables

Row ID	D Col0	D Col1	D Col2	D Col3	S Col4
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row1	4.9	3	1.4	0.2	Iris-setosa
Row2	4.7	3.2	1.3	0.2	Iris-setosa
Row3	4.6	3.1	1.5	0.2	Iris-setosa
Row4	5	3.6	1.4	0.2	Iris-setosa
Row5	5.4	3.9	1.7	0.4	Iris-setosa
Row6	4.6	3.4	1.4	0.3	Iris-setosa
Row7	5	3.4	1.5	0.2	Iris-setosa
Row8	4.4	2.9	1.4	0.2	Iris-setosa
Row9	4.9	3.1	1.5	0.1	Iris-setosa
Row10	5.4	3.7	1.5	0.2	Iris-setosa
Row11	4.8	3.4	1.6	0.2	Iris-setosa
Row12	4.8	3	1.4	0.1	Iris-setosa
Row13	4.3	3	1.1	0.1	Iris-setosa
Row14	5.8	4	1.2	0.2	Iris-setosa
Row15	5.7	4.4	1.5	0.4	Iris-setosa
Row16	5.4	3.9	1.3	0.4	Iris-setosa
Row17	5.1	3.5	1.4	0.3	Iris-setosa
Row18	5.7	3.8	1.7	0.3	Iris-setosa
Row19	5.1	3.8	1.5	0.3	Iris-setosa
Row20	5.4	3.4	1.7	0.2	Iris-setosa
Row21	5.1	3.7	1.5	0.4	Iris-setosa
Row22	4.6	3.6	1	0.2	Iris-setosa
Row23	5.1	3.3	1.7	0.5	Iris-setosa
Row24	4.8	3.4	1.9	0.2	Iris-setosa

Figura 6: Tabla de datos leída con el nodo CSV reader

El nombre para identificar estas columnas lo asigna KNIME por defecto, pero es posible cambiarlo con el nodo "column rename" que se encuentra en el repositorio de nodos correspondientes a manipulación. A modo de ejemplo, para mostrar la potencialidad y versatilidad de KNIME en cuanto a manipulación de datos, lo agregamos a nuestro flujo de datos y unimos el puerto de salida del nodo de lectura con el puerto de entrada del nodo para renombrar la columna (Figura 7). Como se puede observar, el nodo aparece con la luz amarilla que nos indica que el nodo debe ser configurado antes de ejecutarlo. Para ello, como se ha mencionado anteriormente, se hace doble clic sobre el nodo o se selecciona la opción en el menú que se abre tras hacer clic con el botón derecho. En la Figura 8 se puede ver el menú de configuración de este nodo, donde en la parte izquierda se pueden seleccionar las columnas que se quieren renombrar y en la parte derecha se indica el nuevo nombre y el tipo de datos que contiene esa columna.

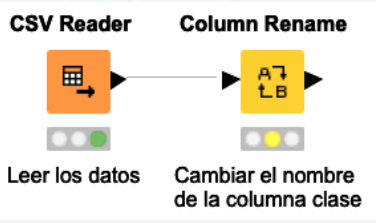


Figura 7: Enlace de los dos primeros nodos del flujo de datos de ejemplo

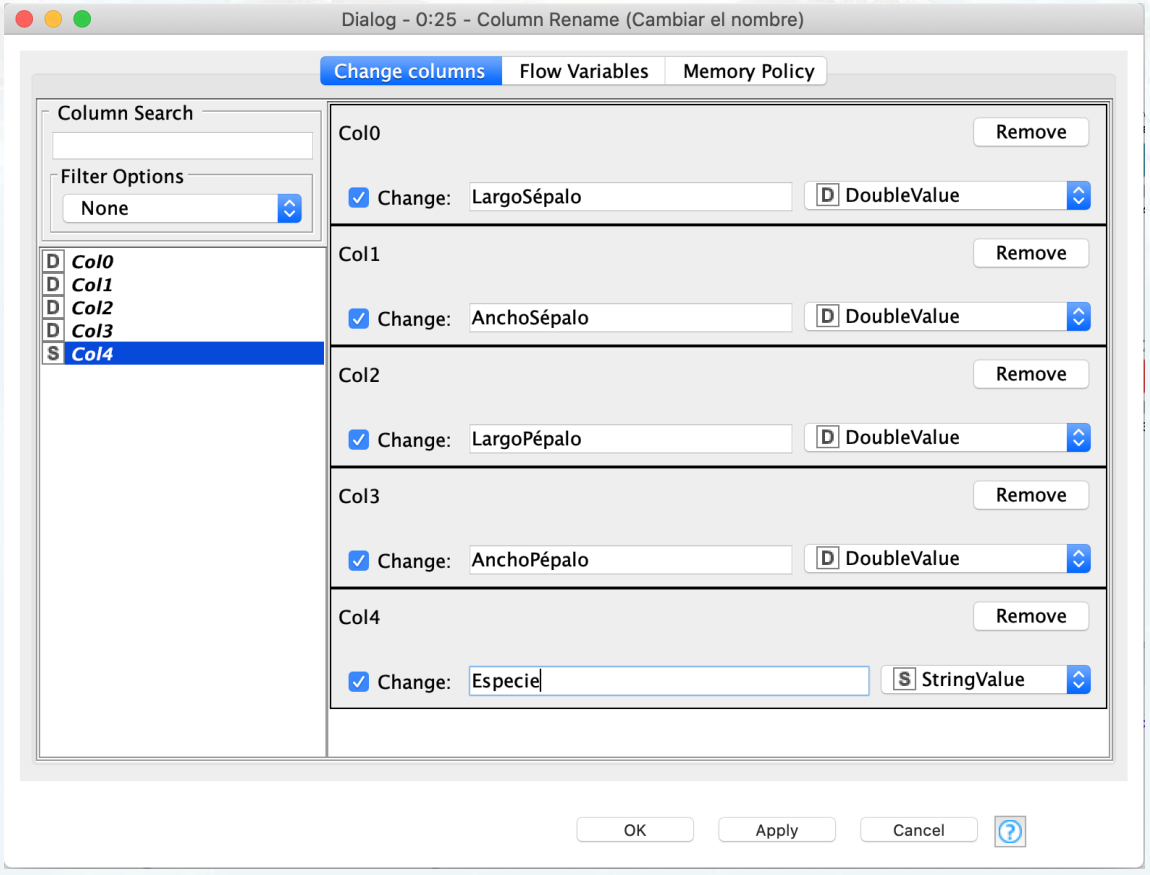
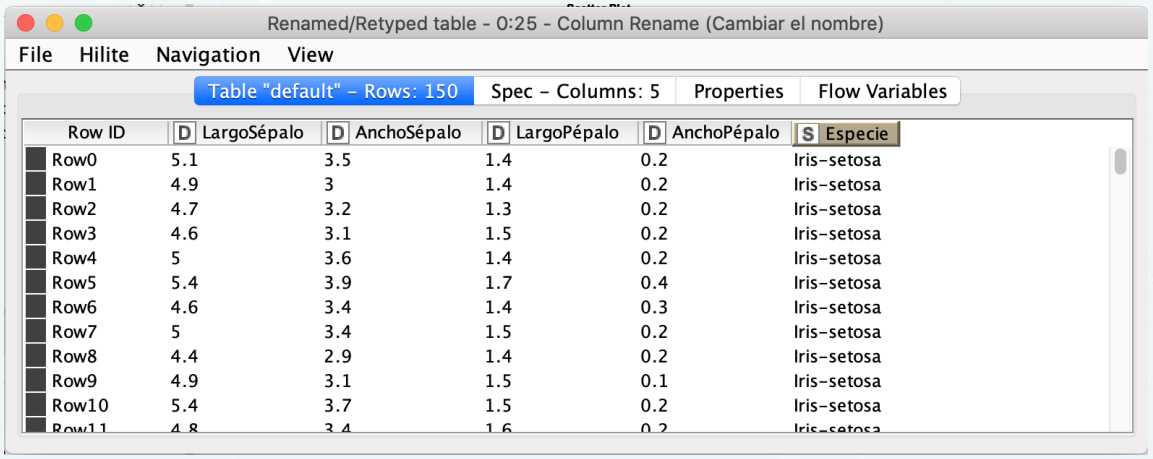


Figura 8: Configuración del nodo para renombrar las columnas

El resultado que se obtiene al ejecutar este nodo se ilustra en la Figura 9, donde se puede observar cómo se ha cambiado el nombre de las columnas.



Row ID	LargoSépalo	AnchoSépalo	LargoPépalo	AnchoPépalo	Especie
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row1	4.9	3	1.4	0.2	Iris-setosa
Row2	4.7	3.2	1.3	0.2	Iris-setosa
Row3	4.6	3.1	1.5	0.2	Iris-setosa
Row4	5	3.6	1.4	0.2	Iris-setosa
Row5	5.4	3.9	1.7	0.4	Iris-setosa
Row6	4.6	3.4	1.4	0.3	Iris-setosa
Row7	5	3.4	1.5	0.2	Iris-setosa
Row8	4.4	2.9	1.4	0.2	Iris-setosa
Row9	4.9	3.1	1.5	0.1	Iris-setosa
Row10	5.4	3.7	1.5	0.2	Iris-setosa
Row11	4.8	3.4	1.6	0.2	Iris-setosa

Figura 9: Salida del nodo que permite renombrar las columnas

Por último, para acabar con el preprocesamiento de los datos, se va a dividir el conjunto de datos obtenido en dos conjuntos diferentes. Así tendremos un conjunto de datos del cual se puede aprender un modelo (training set) y otro con el que validar el modelo obtenido (test set). Recordad que el objetivo del ejemplo que se está desarrollando es aprender un modelo que nos permita obtener la clase de una flor en función de 4 características de ésta.

Para la partición del conjunto de datos vamos a utilizar la metodología que se ha detallado en el módulo 3 (cápsula 2). Para ello, utilizamos el nodo "*Partitioning*" que divide la tabla de entrada en dos particiones que se pueden consultar a través de los dos puertos de salida. En la Figura 10 se puede observar el menú de configuración de este nodo. En primer lugar, hay que especificar si la partición se va a realizar de manera absoluta (número absoluto de filas en la partición) o relativa (porcentaje del número de filas en la tabla de entrada que están en la primera partición). En el ejemplo, realizamos una partición relativa del 80%. Además, en la parte inferior, se puede definir como se ordenan los datos en las particiones según las diversas metodologías explicadas en el módulo 3. Por ejemplo, la opción "take from top" ordena las filas superiores en la primera tabla de salida y el resto en la segunda tabla y "Stratified sampling" realiza un muestreo estratificado, es decir, la distribución de valores en la columna seleccionada se mantiene (aproximadamente) en las tablas de salida. Recordad que en la zona de descripción de nodos de la pantalla inicial de KNIME se encuentran detalladas las diferentes opciones de los nodos.

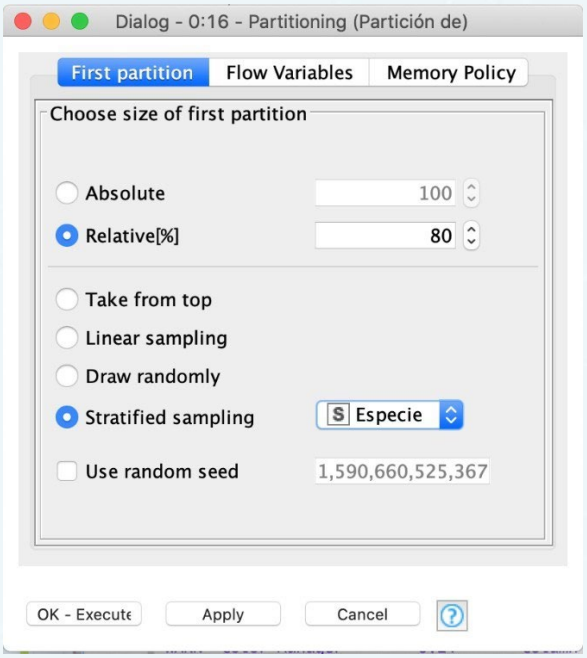


Figura 10: Configuración del nodo para hacer las particiones del conjunto de datos

A continuación, vamos a utilizar un par de nodos del repositorio de "views" que nos permiten explorar los datos de manera visual, un aspecto muy interesante en el análisis de datos. En primer lugar, vamos a utilizar el nodo "color manager" para distinguir visualmente las especies (clases). Seleccionamos la opción de configurar el nodo para establecer la columna donde se ubican las clases (Figura 11) en la parte superior. Una vez configurado se ejecuta y se puede ver que se ha asignado un color para cada una de las clases (Figura 12). En concreto, para la clase *Iris setosa* ha asignado el color rojo y para la clase *iris versicolor* le ha asignado el color morado.

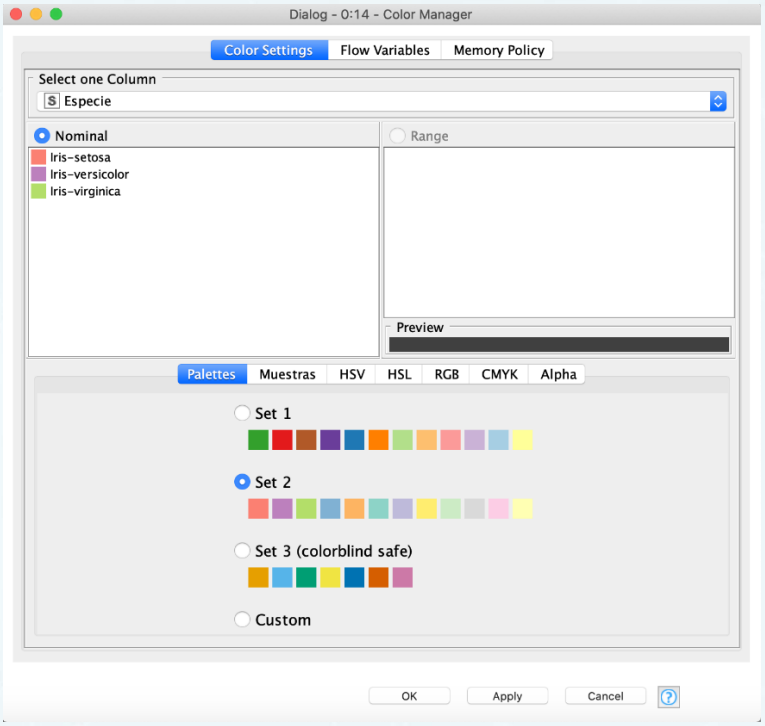


Figura 11: Ventana de configuración del nodo "color manager"

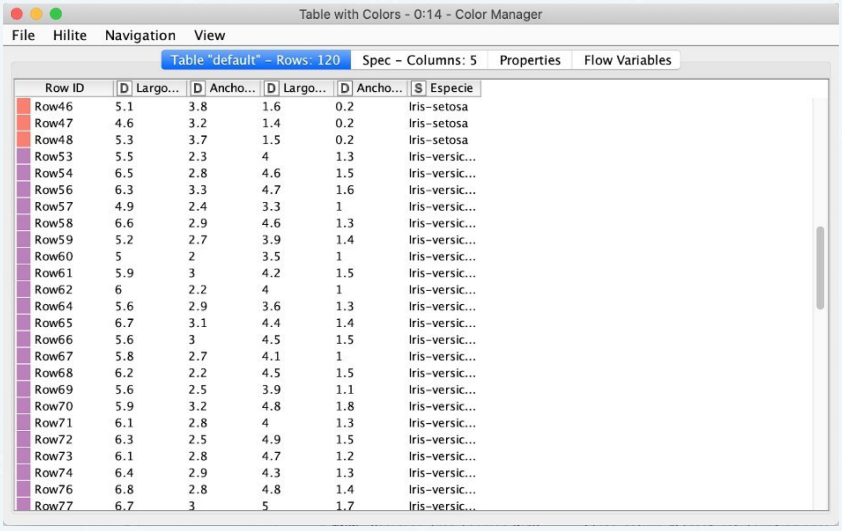


Figura 12: Salida del nodo "color manager"

Otro nodo que es común es el nodo "*Scatter Plot*" que presenta un diagrama de dispersión en el que ofrece la capacidad de elegir diferentes columnas para x e y. En el menú de configuración se seleccionan las dos variables a analizar y tras ejecutar el nodo se obtiene una imagen como la que se muestra en la Figura 13. En esta figura aparecen los distintos casos según los colores que se han asignado en el nodo anterior para cada clase.

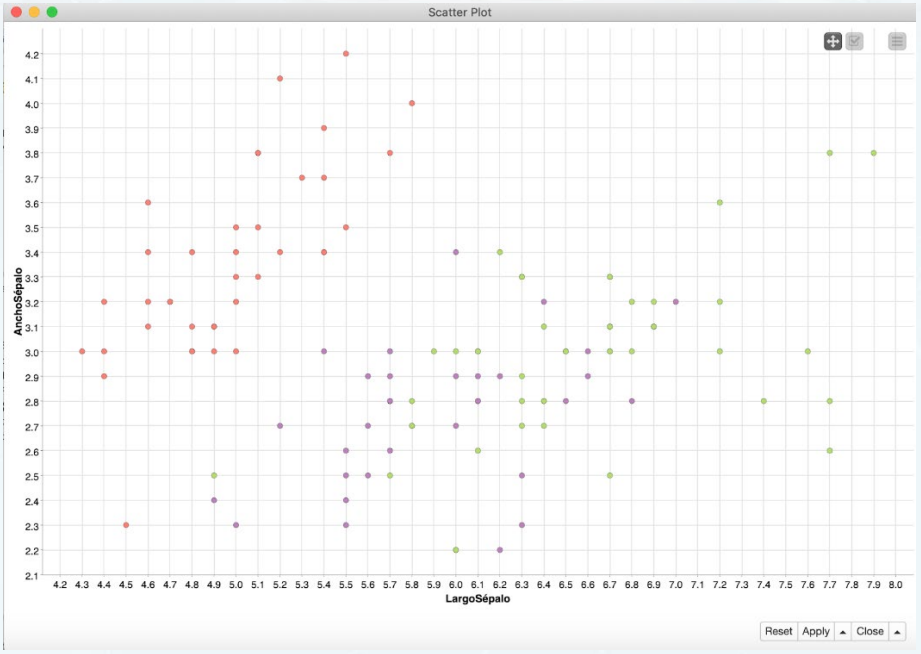


Figura 13: Salida del nodo Scatter plot

A continuación, vamos a comenzar con los nodos que permiten desarrollar el modelo (fase de análisis de datos). En este ejemplo se va a utilizar el árbol de decisión, que se explicó en el Módulo 5 (cápsula 2), como técnica de representación para la clasificación de las especies, aunque como hemos comentado, se podría utilizar cualquier otro algoritmo de aprendizaje automático. El primer paso consiste en buscar en el repositorio los dos nodos que se necesitan, el de aprendizaje y el de predicción (*Decision Tree Learn*er y *Decision Tree Predictor*) y arrastrarlos hasta el flujo de trabajo. A continuación, se deben conectar los puertos con el nodo anterior que, en este ejemplo, es el que utilizamos para hacer la partición del conjunto de datos inicial. Recordad que la salida de este nodo eran dos conjuntos de datos que se utilizarían para el aprendizaje y test de nuestro modelo. Por tanto, el puerto de salida superior del nodo de "*Partitioning*" habrá que conectarlo con el puerto de entrada del nodo de aprendizaje "*Decision Tree Learner*" y el puerto de salida inferior con el nodo de predicción "*Decision Tree Predictor*". Estos dos nodos hay que conectarlos entre si a través del puerto de color azul que permite transportar el modelo que ha generado el nodo *Decision Tree Learner*.

A continuación, antes de ejecutarlos, es necesario configurar dichos nodos con los valores y/o parámetros que se consideren adecuados para el conjunto de datos. Como se puede ver en la *Figura 14*, en el menú del nodo de aprendizaje lo primero que hay que indicar es en que columna del conjunto de datos está contenida la especie o clase a clasificar y, además, debajo se pueden configurar: medida de calidad, método de poda, número de registros mínimo por nodo, etc.

En el menú de configuración del nodo que predecirá la especie indicamos el nombre de la nueva columna con el resultado de la clasificación de la especie (Figura 15). A continuación, se ejecutan los dos nodos.

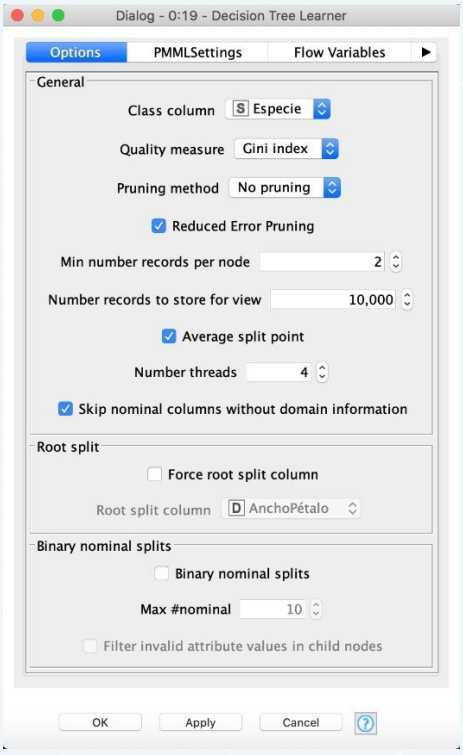


Figura 14: Configuración del nodo "Decision Tree Learner"

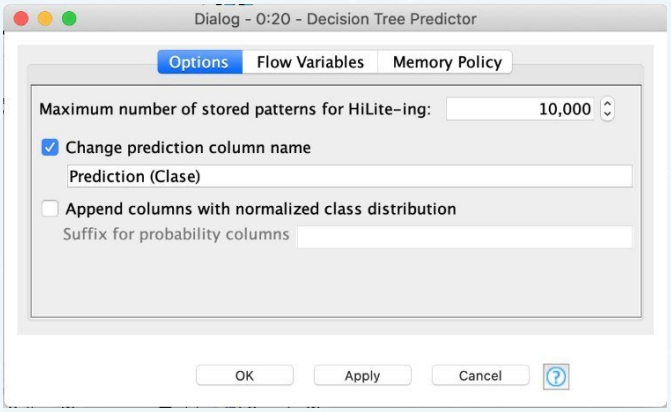
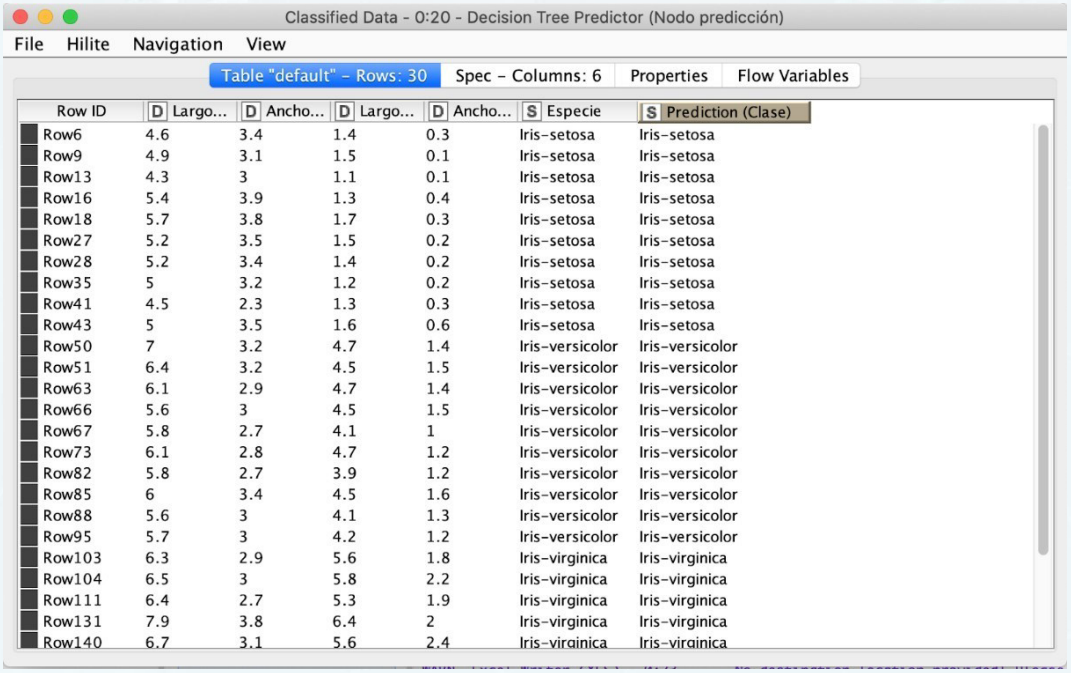


Figura 15: Configuración del nodo "Decision Tree Predictor"

El resultado de este nodo se puede observar en la Figura 16. Para acceder a esta tabla se puede seleccionar la opción dentro del menú desplegable haciendo clic con el botón derecho o seleccionando la opción en la barra de menú. Un aspecto importante en el ciclo de vida de ciencia de datos es la evaluación de la calidad de los modelos aprendidos para medir cómo pueden predecir nuevos ejemplos. Para ello, vamos a incluir dos nodos que nos permitan analizar en detalle la calidad del modelo aprendido ("*Scorer*" y "*ROC curve*"), es decir, cómo está clasificando el modelo:


Classified Data - 0:20 - Decision Tree Predictor (Nodo predicción)

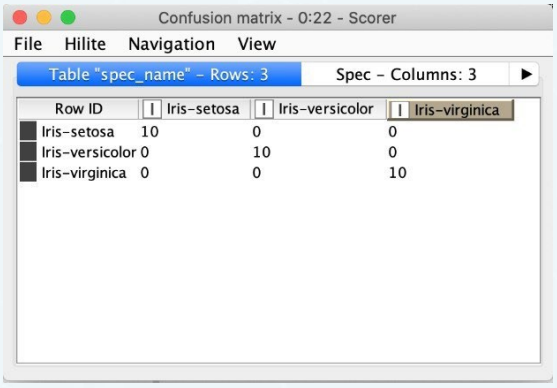
File Hilite Navigation View

Table "default" – Rows: 30 | Spec – Columns: 6 | Properties | Flow Variables

Row ID	Largo...	Ancho...	Largo...	Ancho...	Especie	Prediction (Clase)
Row6	4.6	3.4	1.4	0.3	Iris-setosa	Iris-setosa
Row9	4.9	3.1	1.5	0.1	Iris-setosa	Iris-setosa
Row13	4.3	3	1.1	0.1	Iris-setosa	Iris-setosa
Row16	5.4	3.9	1.3	0.4	Iris-setosa	Iris-setosa
Row18	5.7	3.8	1.7	0.3	Iris-setosa	Iris-setosa
Row27	5.2	3.5	1.5	0.2	Iris-setosa	Iris-setosa
Row28	5.2	3.4	1.4	0.2	Iris-setosa	Iris-setosa
Row35	5	3.2	1.2	0.2	Iris-setosa	Iris-setosa
Row41	4.5	2.3	1.3	0.3	Iris-setosa	Iris-setosa
Row43	5	3.5	1.6	0.6	Iris-setosa	Iris-setosa
Row50	7	3.2	4.7	1.4	Iris-versicolor	Iris-versicolor
Row51	6.4	3.2	4.5	1.5	Iris-versicolor	Iris-versicolor
Row63	6.1	2.9	4.7	1.4	Iris-versicolor	Iris-versicolor
Row66	5.6	3	4.5	1.5	Iris-versicolor	Iris-versicolor
Row67	5.8	2.7	4.1	1	Iris-versicolor	Iris-versicolor
Row73	6.1	2.8	4.7	1.2	Iris-versicolor	Iris-versicolor
Row82	5.8	2.7	3.9	1.2	Iris-versicolor	Iris-versicolor
Row85	6	3.4	4.5	1.6	Iris-versicolor	Iris-versicolor
Row88	5.6	3	4.1	1.3	Iris-versicolor	Iris-versicolor
Row95	5.7	3	4.2	1.2	Iris-versicolor	Iris-versicolor
Row103	6.3	2.9	5.6	1.8	Iris-virginica	Iris-virginica
Row104	6.5	3	5.8	2.2	Iris-virginica	Iris-virginica
Row111	6.4	2.7	5.3	1.9	Iris-virginica	Iris-virginica
Row131	7.9	3.8	6.4	2	Iris-virginica	Iris-virginica
Row140	6.7	3.1	5.6	2.4	Iris-virginica	Iris-virginica

Figura 16: Salida del nodo "Decision Tree Predictor" con los datos clasificados

El nodo "*Scorer*" proporciona, por un lado, la matriz de confusión (Módulo 5.1) que ofrece un conteo de los aciertos y errores de cada una de las especies (setosa, virginica y versicolor) permitiendo comprobar si nuestro modelo está confundiéndose entre clases, y en qué medida (Figura 16). En segundo lugar, proporciona una tabla con datos estadísticos: True positives, True negatives, false positives, false negatives, recall, precision, F-measure, etc. (Figura 17).

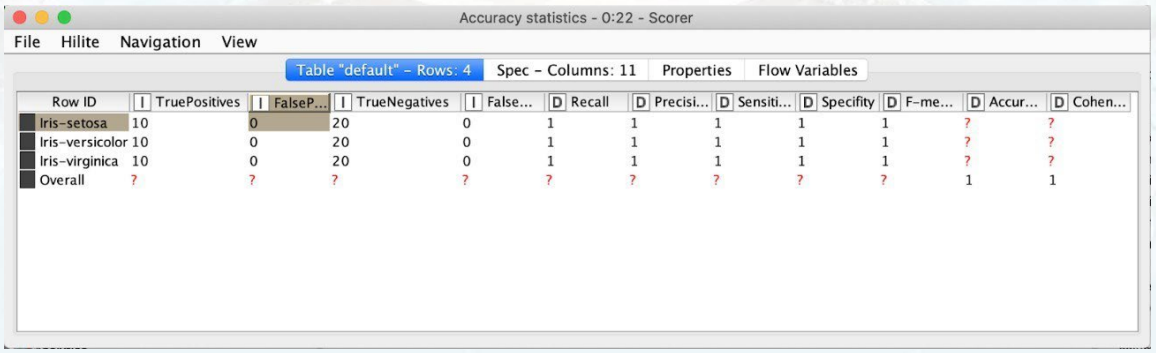

Confusion matrix - 0:22 - Scorer

File Hilite Navigation View

Table "spec_name" – Rows: 3 | Spec – Columns: 3

Row ID	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	10	0	0
Iris-versicolor	0	10	0
Iris-virginica	0	0	10

Figura 17: Matriz de confusión



Row ID	TruePositives	FalseP...	TrueNegatives	False...	Recall	Precisi...	Sensiti...	Specifity	F-me...	Accur...	Cohen...
Iris-setosa	10	0	20	0	1	1	1	1	1	?	?
Iris-versicolor	10	0	20	0	1	1	1	1	1	?	?
Iris-virginica	10	0	20	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

Figura 18: Tabla con los datos estadísticos de precisión

El nodo “ROC curve” permite obtener curvas ROC para problemas de clasificación de dos clases, tal y como se detalló en el módulo 5 (cápsula 1). La tabla de entrada debe contener una columna con los valores de clase reales (incluidos todos los valores de clase como valores posibles) y una segunda columna con las probabilidades de que un elemento se clasifique como perteneciente a la clase seleccionada.

En el ejemplo hemos configurado este nodo con los siguientes datos:

- Class Column: (Predición clase)
- Positive class value: Iris setosa

El resultado se puede ver en la Figura 19.

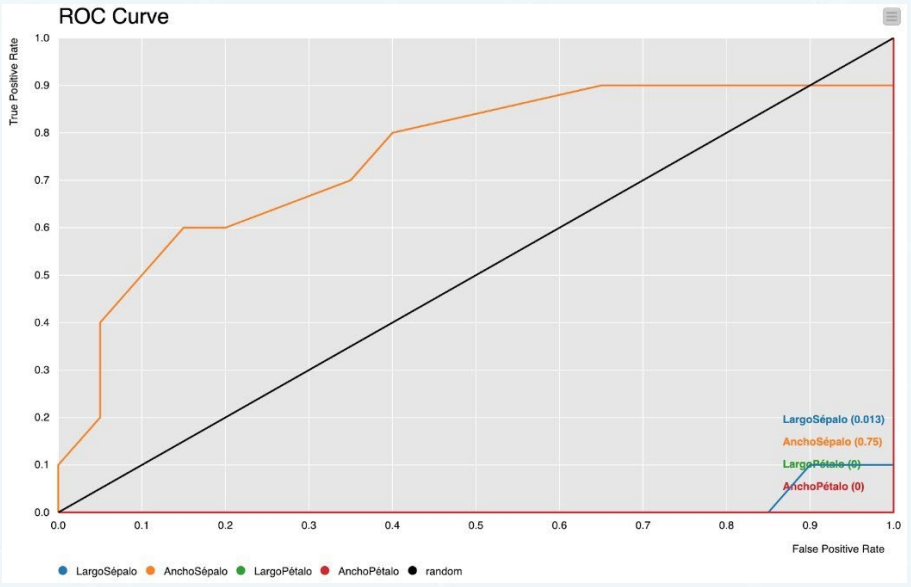


Figura 19: Curva ROC

Por último, los últimos nodos del flujo de trabajo nos permiten escribir los resultados obtenidos. En este ejemplo, se ha incluido el nodo de "*Table Writer*" y "*Excel Writer*" para exportar la tabla de confusión y los datos estadísticos respectivamente. En los nodos de lectura es necesario indicar en el menú de configuración dónde se guardará el fichero, así como los datos que se quieren exportar. Por ejemplo, en la Figura 19 se puede ver el menú del nodo que permite exportar los datos en formato Excel. Como se puede observar, en la parte superior se indica la localización del fichero de salida, en la parte intermedia algunas opciones de formato y en la parte inferior se pueden definir los datos que se quieren exportar.

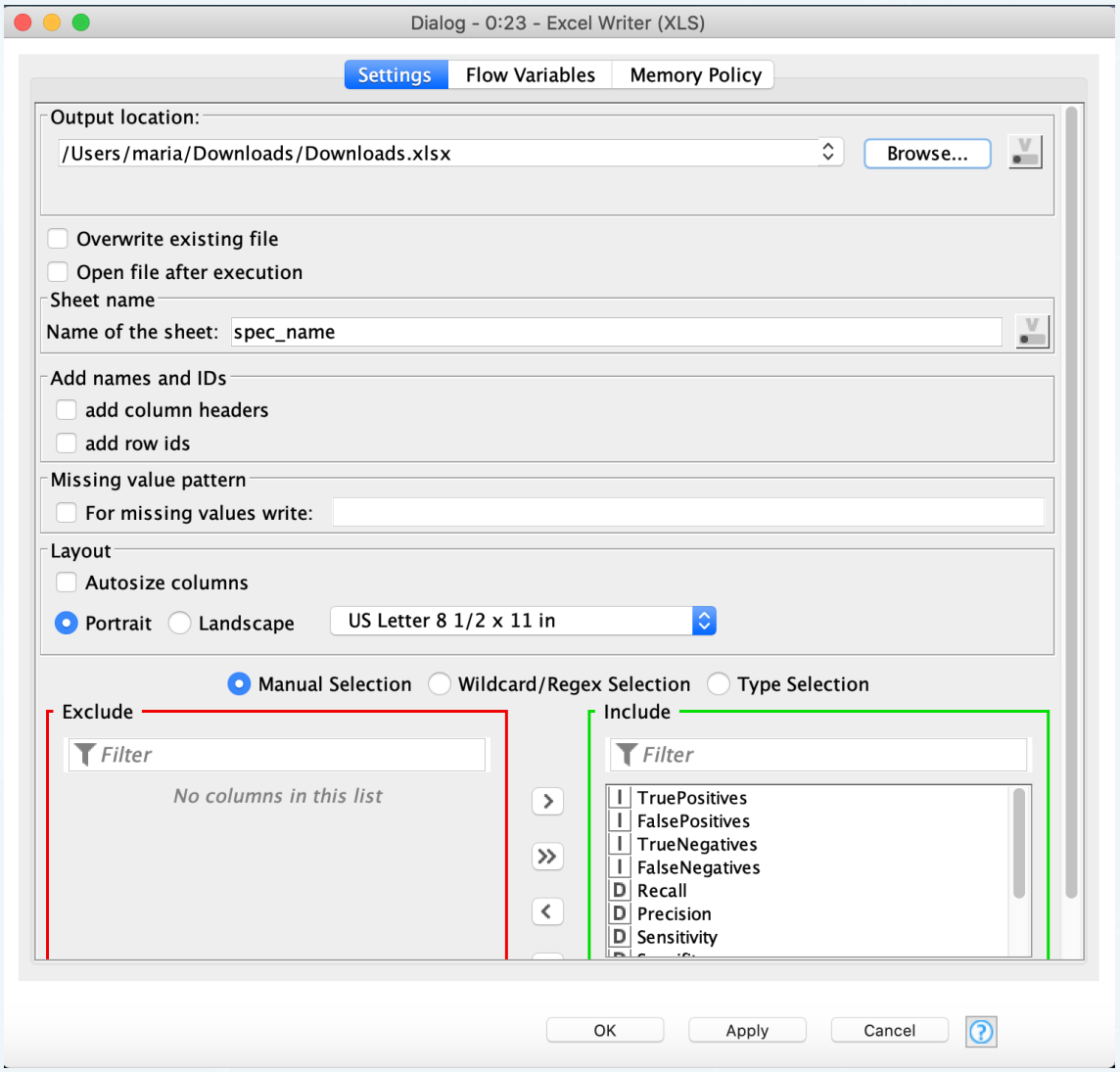


Figura 20: Menú de configuración del nodo "Excel Writer"

3. REFERENCIAS BIBLIOGRÁFICAS

- Bakos, G. (2013). KNIME essentials. Packt Publishing Ltd.
- Blokdyk, G. (2019). KNIME a Complete Guide - 2019 Edition. Emereo Pty Limited, 2019.
- McCormick, K. (2019). Introduction to Machine Learning with KNIME. linkedin.com
- Silipo, R. (2016). Introduction to Data Analytics with KNIME: A Data Science Approach to Analytics. O'Reilly.
- Silipo, R., & Mazanetz, M. P. (2012). The KNIME cookbook. KNIME Press, Zürich, Switzerland.
- Strickland, J. (2016). Data Analytics Using Open-Source Tools. Lulu. com.