Módulo 8

8.1 ¿Cómo utilizar KNIME? – Flujos de trabajo

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 A KNIME

KNIME es una herramienta de software libre[1] multiplataforma[2] con interfaz gráfica que permite realizar el ciclo completo de Ciencia de Datos en un entorno visual, sencillo e intuitivo a través de flujos de datos definidos con nodos interconectados entre sí. En este contexto, KNIME dispone de herramientas para:

- Visualización de datos
- Pre-procesado de datos
- Extracción de modelos mediante algoritmos
- Comparación de modelos
- Análisis de resultados

Además, KNIME integra otras plataformas como: BIRT (para creación de informes), WEKA (para minería de datos), Python, R (para análisis estadísticos y visualización) además de otras extensiones para lectura de datos, ETL (Extracción, Transformación y Carga de datos), generación de informes, visualización y análisis.

1.1. INSTALACIÓN DE KNIME

KNIME puede ser descargado y utilizado gratuitamente a través del siguiente enlace: https://www.knime.com/downloads

[1] Software Libre se refiere a un software que se puede utilizar sin coste económico.

[2] Multiplataforma se refiere a que se puede utilizar en distintos sistemas operativos, como Windows, Linux o Mac.

Al ser multiplataforma, se puede instalar la versión adecuada a cada plataforma: Windows, Linux o Mac.

En https://www.knime.com/installation se puede consultar una guía rápida de instalación de la aplicación, así como del funcionamiento básico de cada elemento del sistema.

1.2. EL ENTORNO KNIME

Una vez instalado KNIME, al ejecutarlo, se requiere que se indique el espacio de trabajo. Ese espacio consiste en una carpeta en un directorio de usuario de nuestro ordenador donde se almacenarán todos los flujos de datos creados por el usuario. Se recomienda utilizar el espacio de trabajo asignado por defecto.

Una vez asignado el espacio de trabajo, se lanza el entorno de KNIME. Éste tiene el aspecto que se muestra en la Figura 1, el cual se divide en las distintas zonas, etiquetadas en la propia figura, y que introducen a continuación:

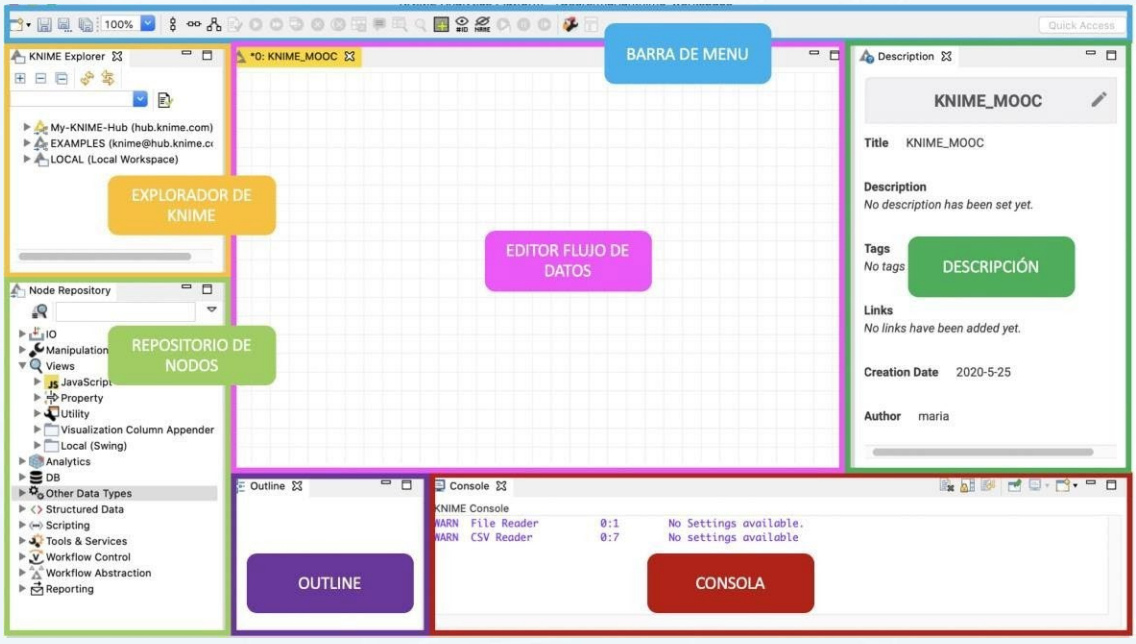


Figura 1: Pantalla de inicio

1. **Barra de menú**

   La barra de menú se encuentra en la parte superior. Consta de accesos directos a varias opciones como guardar flujo de datos, guardarlo como otro flujo de trabajo, alinear y configurar nodos, visualizar los resultados, etc.

2. **Explorador de KNIME** (*KNIME Explorer*)

   El explorador de KNIME se encuentra en la parte superior izquierda. En esta zona se puede encontrar la descripción general de los flujos de trabajo, así como menús desplegables con diversos flujos de datos disponibles en los espacios de trabajo activos de KNIME, es decir, su espacio de trabajo local, así como los servidores KNIME. En esto últimos se pueden descargar gran variedad de ejemplos concretos de diversa índole, como por ejemplo el análisis de la cesta de la compra con reglas de asociación o ejemplos de filtrado de nodos.

3. **Repositorio de nodos** (*Node repository*)

   El repositorio de nodos se encuentra debajo del explorador de KNIME, en la esquina inferior izquierda. En esta sección se enumeran todos los nodos disponibles en la plataforma principal de KNIME y los de las extensiones que se han instalado. Los nodos están organizados por categorías, pero también puede usar el cuadro de búsqueda en la parte superior del repositorio de nodos para buscar nodos. Como veremos más adelante, el uso de KNIME se basa en el diseño de flujos de datos que están compuestos de nodos (iconos) que se conectan entre sí.

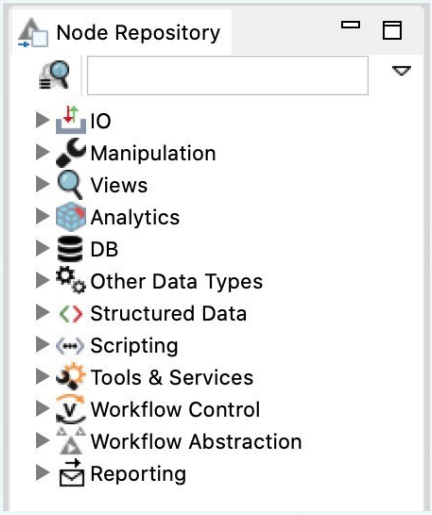


Figura 2: Repositorio de nodos

A continuación, se detallan las principales categorías de nodos:

a) Entrada de datos [IO > Read] y salida de datos [IO > Write].
b) Preprocesamiento [Data Manipulation], para filtrar, discretizar, normalizar, filtrar, seleccionar variables, entre otros.

c) Visualización de datos [Views] para mostrar resultados en pantalla, ya sea de forma textual o gráfica.
d) Minería de datos [Mining], para construir modelos como reglas de asociación, clustering, clasificación, PCA, etc.
e) Otros tipos de nodos específicos, cuyo uso queda lejos del contenido de este curso.

1. **Outline (Salida)**
   La descripción y visión general del flujo de datos que está actualmente activo se puede visualizar en esta sección que se ubica a la derecha del repositorio de nodos.

2. **Consola**
   Justo a la derecha del outline se encuentra la consola. Ésta muestra mensajes de ejecución que indican al usuario lo que sucede al ejecutar el flujo de trabajo.

3. **Editor del flujo de datos (Workflow editor)**
   En la parte central de la pantalla se encuentra la zona para crear y diseñar el flujo de datos.

4. **Descripción**
   En la parte superior derecha se encuentra la descripción de los nodos seleccionados tanto en el repositorio de nodos como los nodos del flujo de trabajo.

1.3. NODOS Y FLUJOS DE DATOS

Como se ha comentado antes, el uso de KNIME se basa en el diseño de flujos de datos, los cuales representan las distintas etapas de un proyecto de extracción de conocimiento. Estos flujos de datos están compuestos de nodos (iconos) que se conectan entre sí. Los nodos disponen de una serie de conexiones, llamados puertos, de entradas y salidas que se interconectan entre sí para definir el flujo de trabajo. En este sentido, las salidas de unos nodosse utilizan como entrada de otros nodos donde transportan datos a través de los puertos. Cadanodo implementa varios procedimientos, procesos, etc., siendo básicamente unidades de procesamiento de un flujo de datos.

1.3.1 Los nodos

Los nodos pueden realizar todo tipo de tareas, incluida la lectura / escritura de archivos, la transformación de datos, la generación de modelos a partir de los conjuntos de datos, la creación de visualizaciones, etc. Un nodo se representa visualmente en KNIME como un pequeño icono compuesto de puertos de entrada y salida, así como del estado actual del nodo. El puerto de entrada se encuentra situado a la izquierda del nodo, el puerto de salida a la derecha del nodo y el estado en la parte inferior del nodo. Por los puertos circulan datos.

1.3.2 Los puertos

Las entradas son los datos que procesa el nodo, y las salidas son los conjuntos de datos resultantes. KNIME contempla distintos tipos de puertos de entrada o salida como mecanismos de transmisión de datos. Éstos suelen representar distintos tipos de conexiones de datos, entre los que se encuentran principalmente: datos, bases de datos y modelos. Generalmente, la información se transfiere en forma de tablas como las de Excel, con una cabecera donde se indica el nombre de la variable y de que tipo de dato es (Cadena de texto, decimal, entero, etc.)

¡IMPORTANTE! *Sólo se pueden conectar puertos de la misma tipología, por ejemplo, podemos conectar dos nodos a través de un puerto de datos, pero no a través de un puerto de datos y un puerto de modelos*.

En la Figura 3 se pueden observar tres nodos con cada uno de los puertos mencionados.

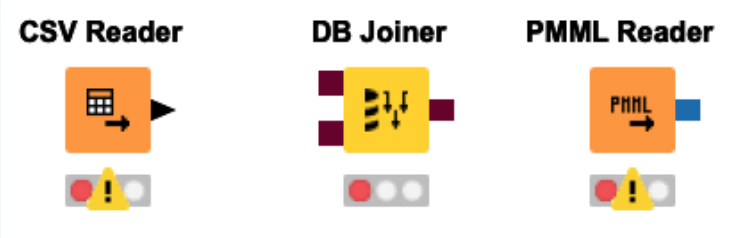


Figura 3 Tres nodos con puertos diferentes

Como ejemplo, el nodo "*CSV Reader*" tiene un puerto de salida de datos. Este tipo de datos es una tabla con datos, donde cada fila es una instancia del problema y cada columna es una de las variables. La variable de salida suele colocarse al final de la tabla. Se representa con un triángulo negro.

El nodo "*DB Joiner*" tiene puertos de entrada y salida de bases de datos y se representa con un cuadrado marrón.

El nodo "PMML Reader" dispone de un puerto de salida que se representa con un cuadrado azul y proporciona el modelo que se ha generado. Este puerto se va a utilizar mucho en este MOOC para conectar nodos "*Learner*" con nodos "*Predictor*" para aplicar el modelo aprendido sobre otro conjunto de datos. La Figura 4 muestra un ejemplo de un nodo "*Learner*" y cómo el modelo aprendido se transfiere a otro nodo a través de un puerto de salida de modelo (cuadrado azul) y cómo éste se utiliza como modelo de entrada a un nodo "*Predictor*" para realizar la predicción, en concreto con un árbol de decisión.

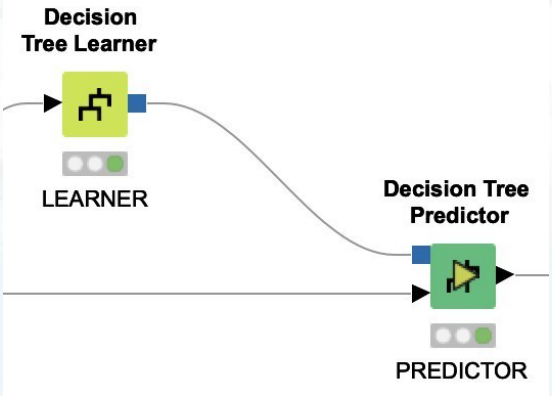


Figura 4: Ejemplo de puerto de modelo

1.3.3 El estado de un nodo

El estado del nodo está representado por un semáforo con colores. Cada color representa un estado: rojo representa no configurado, amarillo representa configurado, pero no ejecutado, verde representa ejecutado y un círculo rojo con una cruz, error en la configuración del nodo.

La Figura 5 muestra un ejemplo de un nodo, así como los distintos estados que representa.

Figura 5: Estado de los Nodos en KNIME

- *Estado rojo:* Nodo no configurado, es un nodo que no está listo para su ejecución. Éste debe configurarse.
- *Estado amarillo:* Nodo configurado, listo para ejecución. Si aparece un icono con forma de triángulo de exclamación significa que el nodo está listo para ejecución, pero con los parámetros por defecto.
- *Estado verde:* Nodo ejecutado. Los datos se encuentran disponibles en los puertos de salida del nodo.
- *Estado círculo rojo con cruz:* Nodo que no se puede ejecutar. Error de configuración.

Las alertas se muestran con un icono amarillo con forma de triángulo y un signo de exclamación.

1.3.4 Metanodos

Los metanodos es un nodo que contiene subflujos de datos, es decir, en el flujo de datos principal se ven como un solo nodo, aunque pueden contener muchos nodos e incluso más metanodos en el interior. El uso de los metanodos es de un nivel avanzado a los contenidos de este MOOC, pero se presentan en esta cápsula de manera introductoria. En la cápsula 3 se expone un ejemplo de metanodo para realizar la validación cruzada.

Para incluir un nodo en el flujo de datos se puede utilizar el asistente de metanodos seleccionando "Nodo / Agregar metanodo" desde el menú o haciendo clic en el botón con el icono del metanodo en la barra de herramientas (el editor de flujo de trabajo debe estar activo). Se puede crear un metanodo predefinido seleccionando uno y haciendo clic en "Finalizar" y se agrega al flujo de datos Figura 6.

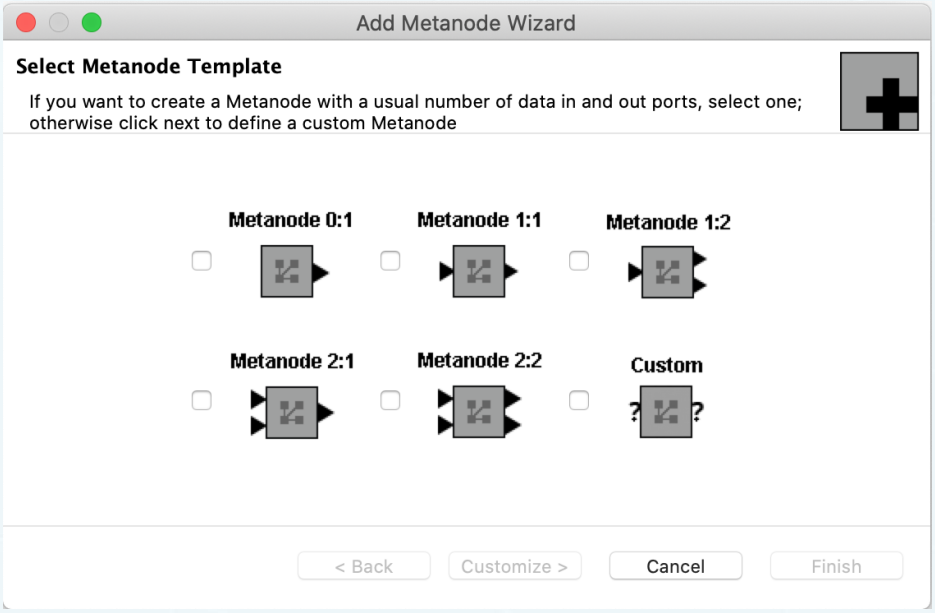


Figura 6: Menú para crear nodos

Si necesita personalizar el metanodo, por ejemplo, con un número diferente de puertos de entrada o salida o desea tener diferentes tipos de puertos, puede seleccionar uno de los metanodos predefinidos como plantilla y luego hacer clic en "Personalizar" para acceder a la siguiente página del asistente.
Para abrir un metanodo, puede hacer doble clic en él o elegir "Abrir editor de subflujo de datos" en su menú contextual. Una vez dentro del metanodo puede incluir nodos igual que en el flujo de datos principal.

Al igual que los nodos, los metanodos tienen estados (Figura 7):

- Inactivo / configurado: Si hay al menos un nodo dentro del metanodo que no se ejecuta ni se ejecuta.
- Ejecutando: si al menos un nodo se está ejecutando
- Ejecutado: si se ejecutan todos los nodos contenidos

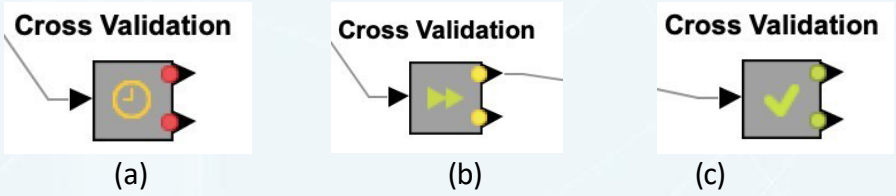


Figura 7: Estados de los metanodos: (a) Inactivo, (b) Ejecutando y (c) Ejecutado

Un ejemplo de metanodo puede ser el *Cross Validation* (validación cruzada), que se ilustra en la Figura 8. Como se puede observar, dentro de este metanodo se realiza la partición del conjunto de datos (*X-Partitioner*), se ejecuta el algoritmo de clasificación (*Decision Tree Learner* y *Decision Tree Predictor*) y se agregan los resultados de la validación cruzada (*X-Aggregator*).

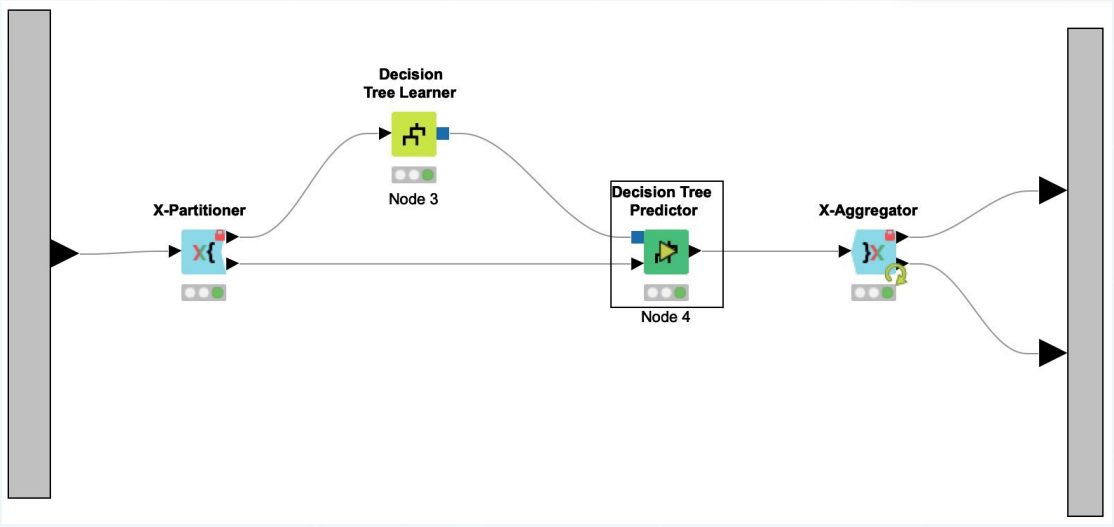


Figura 8: Metanodo de Cross Validation

1.4 CREANDO UN FLUJO DE DATOS.

Un flujo de datos está compuesto de una colección de nodos interconectados. Generalmente representa una parte, o la totalidad, de un ciclo de ciencia de datos.
Para crear un flujo de datos, se seleccionan los correspondientes nodos en el Repositorio de nodos y se arrastran a la zona del editor de flujo de datos. Como se ha mencionado anteriormente, las salidas de unos nodos se utilizan como entradas de otros y se conectan arrastrando el puerto de salida de un nodo al puerto de entrada de otro nodo.

A modo ilustrativo de cómo se crea un flujo de datos y cómo muestra los resultados, nos basaremos en el siguiente flujo de datos:

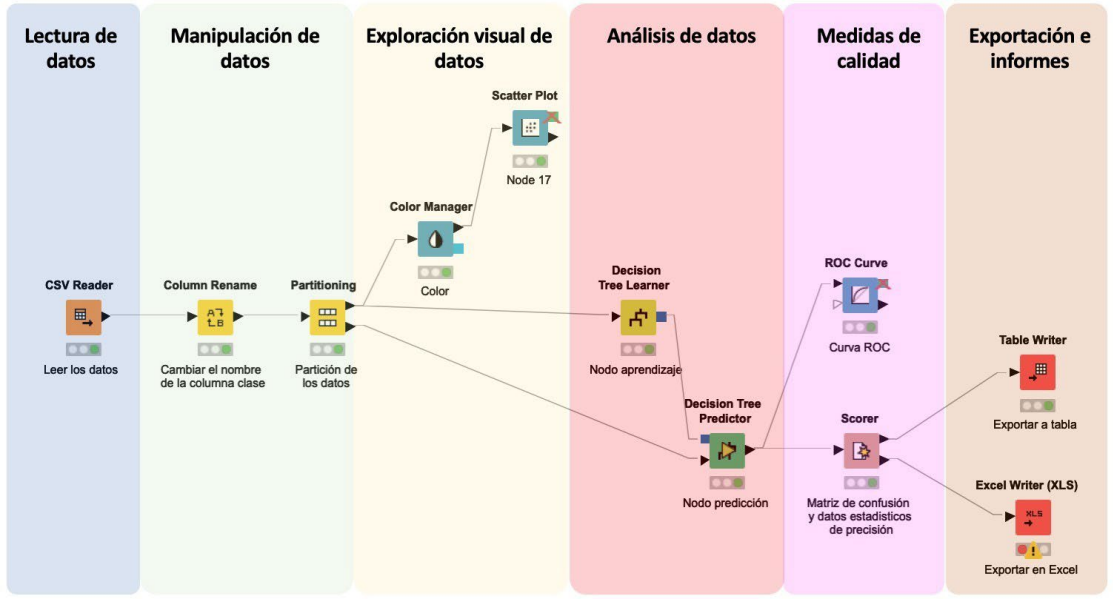


Figura 9 Ejemplo de flujo de datos

El flujo de datos de la Figura 9 muestra, de manera general, las distintas etapas de un proyecto de extracción de conocimiento. Comenzando por una lectura de datos (zona azul inicial), manipulación de datos (zona verde), exploración visual de los datos (zona amarilla), el análisis propiamente dicho para obtener un modelo (zona roja), una evaluación de la calidad del modelo obtenido (zona violeta) y una generación de informes (zona beige).

A continuación, se detallan los principales aspectos a tener en cuenta para crear un flujo de datos, tales como, configurar un nodo, ejecutar el nodo y examinar los datos de salida del nodo.

1.4.1 Configurar un nodo

Para configurar un nodo, basta con seleccionarlo en el editor del flujo de trabajo y pulsar la tecla F6 o botón derecho para desplegar menú y seleccionar opción Configurar (Figura 10).

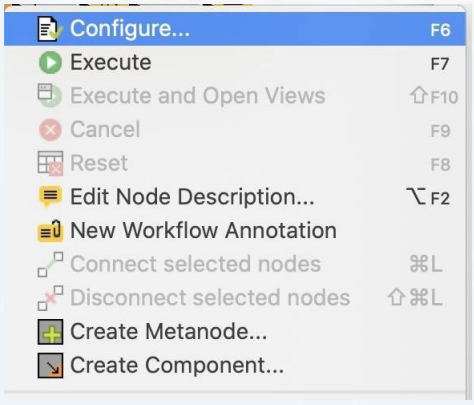


Figura 10: Menú para configurar un nodo

1.4.2 Ejecutar un nodo

Para ejecutar un nodo, basta con seleccionarlo en el editor del flujo de datos y pulsar la tecla F7 o botón derecho para desplegar menú y seleccionar opción Ejecutar (Figura 11).

IMPORTANTE: *un nodo solo puede ejecutarse si todos sus nodos predecesores en el flujo han terminado su ejecución. Si ejecutamos un nodo, automáticamente todos sus nodos predecesores no ejecutados se ejecutarán (siempre que estén configurados y estén listos para ejecutarse)*

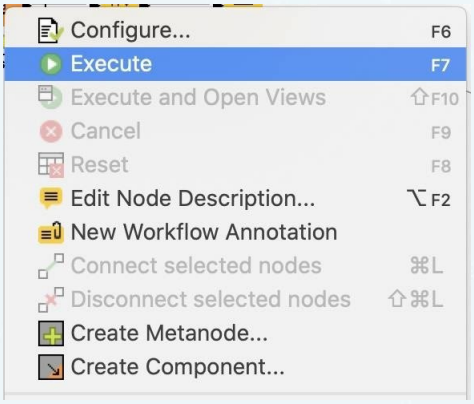
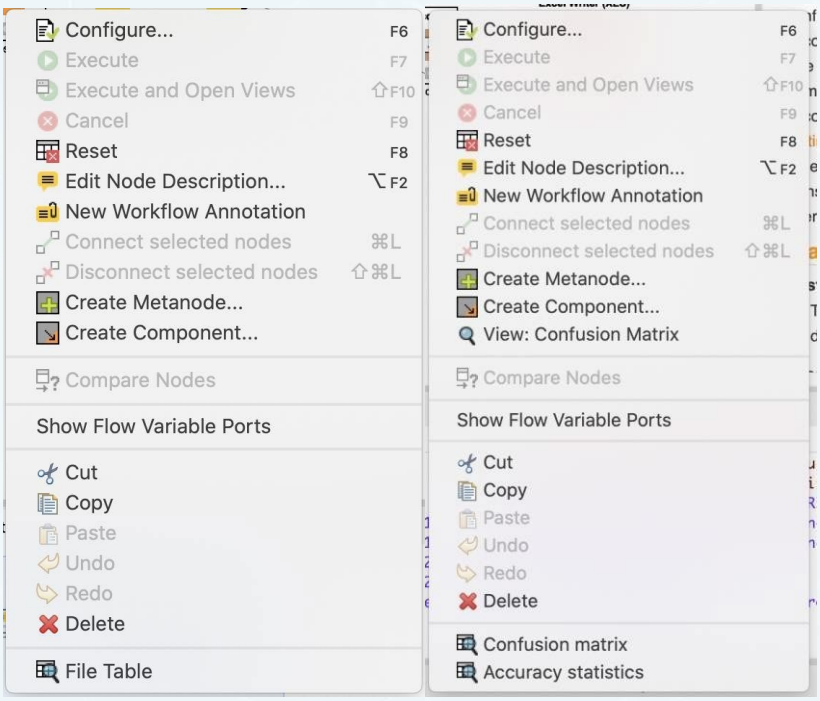


Figura 11:Menu de configuración

1.4.3 Examinar datos de salida de un nodo

Si el nodo se ha ejecutado correctamente (semáforo en color verde), se pueden visualizar los datos que transportan los puertos de salida. Para ello, basta con pulsar el botón derecho, y a través de las últimas opciones del menú se pueden visualizar los datos de los puertos de salida. Por ejemplo, para el nodo "*CSV Reade*r" del flujo de datos de la Figura 12, que sólo tiene un puerto de salida se pueden visualizar los datos en forma de Tabla a través del menú y la última opción: File Table (Figura 12a). Mientras que para el nodo "*Scorer*", que tiene dos puertos de salida, se pueden visualizar los datos correspondientes a la matriz de confusión y las estadísticas a través de la opción: Confusion matrix y Accuracy statistics (Figura 12b)



(a) (b)

Figura 12: (a) Salida del nodo CSV Reader (b) Salida del nodo Scorer

2. 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.