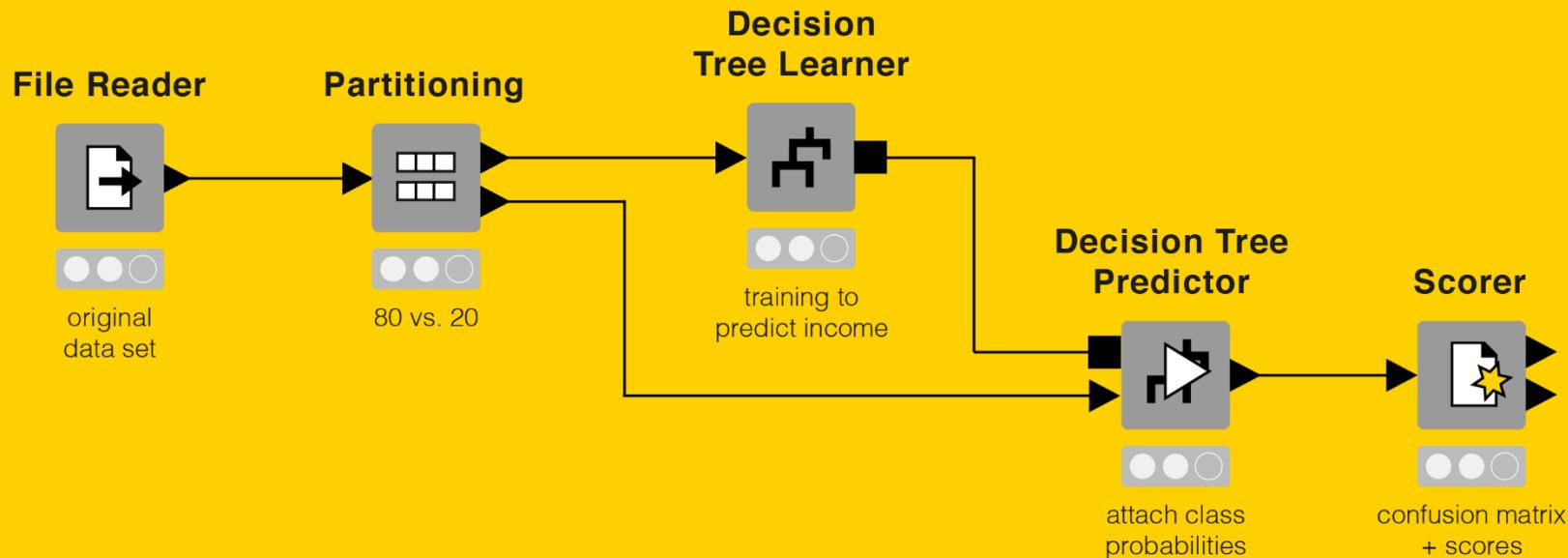


# KNIME® BEGINNER'S LUCK



A Guide to KNIME Analytics Platform for Beginners

Authors: Rafael Borneo, Satoru Hayasaka and Rosaria Silipo

Copyright® 2022 by KNIME Press

Reservados todos los derechos. Esta publicación está protegida por derechos de autor, y se debe obtener permiso del editor antes de cualquier reproducción prohibida, almacenamiento en un sistema de recuperación o transmisión en cualquier forma o por cualquier medio, electrónico, mecánico, fotocopiado, grabación o similar.

Este libro ha sido actualizado para KNIME 4.4.

Para obtener información sobre permisos y ventas, escriba a:

KNIME Press  
Talacker 50  
8001 Zurich  
Switzerland

[knimepress@knime.com](mailto:knimepress@knime.com)

# Table of Contents

Prólogo.....	12
Agradecimientos (versión inglés) .....	13
Nota sobre la traducción al español .....	14
Capítulo 1. Introducción .....	16
1.1. Propósito y estructura del libro.....	16
1.2. Comunidad KNIME.....	17
Páginas web útiles.....	17
Cursos, Eventos y Videos.....	18
Libros.....	19
KNIME Hub.....	19
1.3. Descarga e Instalación de KNIME Analytics Platform .....	23
1.4. Workspace (Ambiente de Trabajo).....	24
La ventana “Workspace Launcher” .....	25
1.5. KNIME Workflow (fllujo de trabajo) .....	26
¿Qué es un workflow? .....	26
¿Que es un nodo? .....	27
1.6. Archivos .knwf and .knar.....	27
1.7. KNIME workbench (banco de trabajo) .....	28
El KNIME Workbench.....	30
Menu Principal .....	32
Hotkeys (teclas importantes) .....	37
Repositorio de Nodos (Node Repository) .....	37
Cuadro de búsqueda .....	38

KNIME Explorer (Explorador de KNIME) .....	38
Mi central KNIME (My-KNIME-Hub) .....	39
El “EXAMPLES Server” .....	40
Montaje de servidores en KNIME Explorer .....	41
El Editor de Workflow .....	42
Personalización del editor de flujo de trabajo .....	43
Anotaciones del Workflow .....	44
Otras personalizaciones del Workbench .....	44
Vista “Node Monitor” .....	45
1.8. Descargar las Extensiones de KNIME .....	45
Instación de KNIME Extensions .....	46
1.9. Datos y workflows para éste libro .....	47
1.10. Ejercicios .....	48
Ejercicio 1 .....	48
Ejercicio 2 .....	49
Ejercicio 3 .....	50
Capítulo 2. Mi primer workflow .....	53
2.1. Operaciones en el Workflow .....	53
Crear un Nuevo Workflow Group .....	54
Crear un nuevo workflow .....	55
Guardar un workflow .....	56
Borrar un workflow .....	56
2.2. Operaciones con Nodos (Node operations) .....	57
Crear un nuevo nodo .....	57
Texto del Nodo .....	58

Configurar un nodo.....	58
Ejecutar un nodo.....	59
Descripción del Nodo .....	60
Ver los Datos Procesados por el Nodo.....	61
2.3. Leer datos desde un archivo.....	61
Crear un nodo lector de CSV “CSV Reader” .....	62
Configure the “CSV Reader node.....	63
Personalizar las propiedades de las columnas.....	64
El protocolo <i>knime://</i> .....	67
2.4. Estructura de datos y tipos de datos en KNIME.....	68
Estructura de datos en KNIME .....	70
2.5. Filtrado de columna de Datos.....	71
Crear un “Column Filter” node .....	71
Configure un nodo “Column Filter” .....	74
2.6. Filtrado de “Data Rows”.....	75
Creando un nodo “Row Filter” .....	76
Configurando el nodo “Row Filter” .....	76
Criterios de filtrado (Row filter) .....	78
2.7. Guardar en un archivo (Write Data to a File).....	80
Creando un nodo “CSV Writer” .....	81
Configurar el nodo “CSV Writer” .....	82
2.8. Ejercicios .....	84
Ejercicio 1 .....	84
Ejercicio 2.....	87
Capítulo 3. Mi primera exploración de datos .....	91

3.1.	Introducción .....	91
3.2.	Reemplazar valores en las columnas.....	92
	Column Rename (Renombrar una columna).....	93
	Nodo “Rule Engine ” .....	96
3.3.	Separación de cadenas de texto (String Splitting ) .....	98
	Cell Splitter por Posición.....	99
	Cell Splitter [by Delimiter].....	100
	RegEx Split (= Cell Splitter by RegEx).....	101
3.4.	Manipulación de cadenas (String Manipulation).....	102
	Nodo “String Manipulation” .....	103
	Case Converter .....	105
	String Replacer.....	106
	Column Combiner.....	107
	Nodo “Column Resorter” .....	109
3.5.	Conversiones de tipos de datos (Type Conversions) .....	110
	Number To String (número a cadena) .....	111
	String To Number (cadena a número) .....	112
	Double To Int.....	113
3.6.	Operaciones de Base de Datos (Database Operations).....	114
	SQLite Connector.....	116
	MySQL Connector.....	117
	Workflow Credentials (credenciales).....	118
	DB Writer .....	120
	Importar un controlador de base de datos JDBC (JDBC Database Driver).....	121
	DB Table Selector ( .....	124

Lector de DB (DB Reader) .....	125
3.7. Agregaciones y agrupaciones (Aggregations and Binning) .....	125
Contenedores numéricos (Numeric Binner).....	127
Agrupando (GroupBy: "Groups" tab) .....	128
GroupBy: Aggregation tabs.....	129
Pivoting (Pivotando).....	130
3.8. Nodos para Visualización de Datos (Nodes for Data Visualization) .....	132
3.9. Gráfico de dispersión (Scatter Plot ) .....	133
Gráfico de dispersión: vista interactiva.....	135
3.10. Propiedades gráficas (Graphical Properties).....	136
Gestor de color "Color Manager" .....	137
3.11. Gráficos de Líneas (Line Plots) y Coordenadas paralelas (Parallel Coordinates) .....	139
Line Plot.....	140
Coordenadas paralelas (Parallel Coordinates) .....	142
3.12. Gráficos de Barras e Histogramas (Bar Charts and Histograms) .....	143
Bar Chart .....	144
Nodo "Table View".....	147
3.13. Ejercicios .....	149
Ejercicio 1 .....	149
Ejercicio 2.....	151
Ejercicio 3.....	151
Chapter 4. Mi Primer Modelo .....	155
4.1. Introducción .....	155
4.2. Dividir y combinar conjuntos de datos .....	156
Muestreo por filas (Row Sampling).....	157

Partitioning.....	158
Shuffle .....	159
Concaternar (Concatenate) .....	160
<b>4.3. Transformando columnas (Transform Columns) .....</b>	<b>162</b>
PMMl .....	163
Valores faltantes (Missing Value) .....	164
Normalizer (normalizador).....	166
Metodos de normalización (Normalization Methods).....	167
Nodo Normalizer (Apply) .....	167
<b>4.4. Modelos de Aprendizaje Automático (Machine Learning Models).....</b>	<b>169</b>
Modelo Naïve Bayes.....	170
Scorer (Javascript).....	173
Arboles de Decisión (Decision Tree).....	178
Curva ROC .....	187
Red neuronal artificial (Artificial Neural Network) .....	189
Write/Read Models to/from file.....	192
Nodo “Statistics”.....	195
Regresión .....	197
Clustering .....	199
Nodo “Cluster Assigner”.....	201
Test de hipótesis .....	202
<b>4.5. Ejercicios .....</b>	<b>203</b>
Ejercicio 1 .....	203
Ejercicio 2.....	204
Ejercicio 3.....	205

Capítulo 5. Preparación de los datos para la elaboración de informes .....	207
5.1.    Introducción .....	207
5.2.    Transformación de filas (Transform Rows) .....	208
RowID .....	211
Nodo "Sorter" .....	214
5.3.    Uniendo Columnas.....	214
Joiner.....	216
Nodo Joiner: pestaña "Joiner Settings" .....	217
Nodo Joiner: pestaña "Column Selection".....	218
5.4.    Nodos Miscelaneos (Misc Nodes) .....	220
Java Snippet (simple).....	221
Java Snippet .....	222
Nodo Math Formula .....	223
Math Formula (Multi Column) .....	225
5.5.    Limpieza final (Cleaning Up the Final Workflow).....	227
Colapsar nodos preexistentes en un Meta-nodo .....	227
Crear un Meta-nodo desde cero .....	228
Expandir y reconfigurar un metanodo.....	230
5.6.    Próximo paso: Crear un Reporte.....	231
5.7.    Ejercicios .....	232
Ejercicio 1.....	232
Ejercicio 2.....	233
Ejercicio 3.....	234
Capítulo 6. Tableros de mando con vistas compuestas.....	237
6.1.    El tablero de mando (Dashboard).....	237

6.2.	Los Nodos .....	238
	Text Output Widget.....	239
6.3.	El Componente.....	240
6.4.	Agregando Colores .....	243
6.5.	La Vista compuesta (Composite View).....	246
6.6.	En el WebPortal.....	250
6.7.	Ejercicios .....	251
	Ejercicio 1 .....	251
	Capítulo 7. Elaboración de informes con BIRT .....	254
7.1.	Informes con BIRT .....	254
7.2.	Instalación de la extensión del diseñador de informes (BIRT).....	255
7.3.	Marcado de datos en el flujo de trabajo.....	256
	Nodo “Data to Report” .....	257
7.4.	De KNIME a BIRT y viceversa.....	257
7.5.	El entorno BIRT.....	259
7.6.	La Plantilla (The Layout).....	263
7.7.	Las Tablas (The Tables).....	266
	Toggle Breadcrumb.....	270
7.8.	Salto de página .....	275
7.9.	Los Gráficos (Charts) .....	275
	Seleccionar tipo de gráfico .....	276
	Selecciona “Data” .....	277
	Formatear el Gráfico .....	279
7.10.	Hojas de estilo (Style Sheets) .....	287
	Crear un Nuevo estilo .....	288

Aplicando un Style Sheet .....	289
7.11.    Generar el documento final.....	290
7.12.    Texto Dinámico .....	290
7.13.    Informes con otras herramientas .....	293
7.14.    Ejercicios .....	295
Ejercicio 1.....	295
Referencias.....	297
Índice de temas.....	298

# Prólogo

Este es el primer libro que escribí en 2010 para [KNIME Press](#) sobre cómo utilizar KNIME Analytics Platform. Dado que nos acercamos al décimo aniversario de este libro, nosotros (el equipo de prensa de KNIME y yo) pensamos que podría necesitar un nuevo texto de Prólogo. En realidad, no necesita una actualización general, ya que desde su nacimiento se ha actualizado dos veces al año cada año, después de cada nueva versión de KNIME Analytics Platform; no inmediatamente después, pero lo suficientemente cerca.

¡Eso es correcto! KNIME Beginner's Luck, como todos los demás libros electrónicos de KNIME Press, es un libro electrónico en vivo que cambia constantemente para adaptarse a la versión más reciente del software. Esta vivacidad del libro electrónico es también la razón por la que rara vez se ha impreso. ¡Actualizar páginas impresas es, sin duda, más difícil que actualizar un archivo pdf!

Como este es el primer libro, inevitablemente se trata de los conceptos básicos: los conceptos básicos de KNIME Analytics Platform, por supuesto, y también los conceptos básicos de un proyecto de ciencia de datos. Este libro lo guía a través de las funciones de acceso más importantes, las operaciones de transformación de datos y, por supuesto, los nodos de aprendizaje automático disponibles en KNIME Analytics Platform. Complementado con muchos ejemplos de flujos de trabajo, ejercicios y capturas de pantalla, lo familiarizará rápidamente con las funciones básicas del software. Si está buscando temas más avanzados, no los encontrará aquí. En cambio....

Si desea obtener más información sobre algoritmos avanzados de aprendizaje automático, variables de flujo o bucles, consulte la secuela de este libro: "[KNIME Advanced Luck](#)". Si desea obtener más información sobre el procesamiento de texto, consulte el libro "[From Words To Wisdom](#)". Si viene de esa escuela de pensamientos donde la lectura de manuales o instrucciones está sobrevalorada, puede comenzar directamente leyendo sobre soluciones a estudios de casos en varios campos de aplicación en nuestra colección "[Practicing Data Science](#)". Si su trabajo consiste más en integrar y combinar diferentes fuentes de datos y tipos de datos, entonces el libro para usted es "[Will they blend?](#)" collection. Si está realizando la transición de SAS, Excel o Alteryx a KNIME, hay más libros disponibles en la página de [KNIME Press](#).

Todo esto es para decir que el equipo de KNIME Press y yo hemos estado trabajando arduamente para proporcionarles el material de aprendizaje, los libros y los tutoriales, para que sean cada vez más productivos con el software KNIME y los conceptos de ciencia de datos.

Rosaria Silipo (Autora de varios libros de KNIME Press, PhD)

# Agradecimientos (versión inglés)

En primer lugar, me gustaría agradecer a todo el equipo KNIME por su paciencia al tratar conmigo y con mis infinitas preguntas.

Entre todos los demás miembros del equipo de KNIME, me gustaría agradecer específicamente a Peter Ohl por haber revisado este libro para encontrar posibles aspectos que no fueran compatibles con las mejores prácticas de KNIME.

También me gustaría agradecer a Casiana Rimbu por ayudarme a proporcionar las capturas de pantalla más bellas, claras y artísticas que jamás pude imaginar y a Meta Brown por animarme en los primeros pasos para desarrollar la idea embrionaria de escribir este libro.

Muchas gracias finalmente a Heather Fyson por revisar el inglés del libro.

# Nota sobre la traducción al español

La plataforma KNIME es una excelente herramienta para el análisis de datos. Se basa en el uso de programación gráfica en donde el usuario no necesita saber programación en ningún lenguaje computacional para poder utilizarla. Se utilizan nodos que secuencialmente realizan tareas con el fin de realizar el análisis final de los resultados.

La plataforma KNIME se encuentra solamente en el idioma inglés. Todos los nodos, ventanas, menús, asistentes, etc..., están en inglés (el idioma universal del data science). Por tanto, si bien en éste libro se ha hecho en muchos casos la traducción al español de lo que aparece en la plataforma KNIME, cuando se use dicha plataforma todo seguirá apareciendo en el idioma inglés original. Es por esto que todas las figuras aparecen en inglés.

Por ejemplo: "workflow" (base del uso de KNIME) se puede traducir como flujo de trabajo. Sin embargo en la plataforma Ud. seguirá viendo la palabra "workflow" en inglés. NO VERRÁ flujo de trabajo. Otro ejemplo "workspace" se traduce como "espacio de trabajo" pero Ud. seguirá viendo "workspace" en la plataforma.

Lo que se ha hecho es: dejar una mezcla de español e inglés. Así pues, Ud. verá algo así como:

Si ha iniciado KNIME por primera vez, su panel "KNIME Explorer" en la esquina superior izquierda del KNIME workbench (banco de trabajo) contiene solo un grupo de flujo de trabajo (carpeta) llamado "**Example Workflows**". Esta carpeta "Example Workflows" contiene varias subcarpetas, cada una con flujos de trabajo básicos para casos de uso muy comunes:

- **Basic Examples.** Los flujos de trabajo de la subcarpeta "Ejemplos básicos" muestran operaciones generales básicas, como importar datos, combinar datos, ETL, entrenar y evaluar un modelo y, finalmente, mostrar los resultados en un informe simple.
- **Customer Intelligence.** Los flujos de trabajo básicos para la predicción de abandono, la calificación crediticia y la segmentación de clientes están disponibles dentro de la subcarpeta "*Customer Intelligence*".
- **Retail.** Un motor de recomendaciones está integrado en la subcarpeta "*Retail*".
- **Social Media.** Un ejemplo de análisis de redes sociales está disponible en "*Social Media*".

A medida que transcurre la traducción puede que se use más el español que el inglés siempre con la advertencia de que en la plataforma KNIME todo seguirá apareciendo en el idioma inglés. Por supuesto que las explicaciones, instrucciones, comentarios y sugerencias de cómo utilizar la plataforma han sido traducidos al español.

# Capítulo 1. Introducción

## 1.1. Propósito y estructura del libro

¡Vivimos en la era de los datos! Cada compra que hacemos se registra debidamente; cada transacción de dinero se registra cuidadosamente; cada clic web termina en un archivo de clics web. Hoy en día todo lleva un chip RFID y puede registrar datos. Tenemos datos disponibles como nunca antes. ¿Qué podemos hacer con todos estos datos? ¿Podemos darle algún sentido? ¿Podemos usarlo para aprender algo útil y rentable? Necesitamos una herramienta, un bisturí quirúrgico que nos permita cortar cada vez más nuestros datos, mirarlos desde muchas perspectivas diferentes, representar su estructura subyacente.

Supongamos entonces que tenemos esta enorme cantidad de datos ya disponibles, esperando ser analizados. ¿Cuáles son las opciones para que un profesional ingrese al mundo de Business Intelligence (BI) y Data Science (DS)? Las opciones disponibles son, por supuesto, múltiples y están creciendo rápidamente. Si nuestro profesional no controla un presupuesto excesivo, podría recurrir al mundo del software de código abierto. El software de código abierto, sin embargo, es más que una elección basada en el dinero. En muchos casos, representa una filosofía de software para compartir y controlar los recursos que muchos profesionales apoyan.

Dentro del mundo del software de código abierto, podemos encontrar algunas herramientas de ciencia de datos y BI. [KNIME Analytics Platform](#) representa una elección fácil para el profesional novato. No requiere el aprendizaje de un script específico y ofrece una interfaz gráfica de usuario (GUI) para implementar y documentar los procedimientos de análisis. Además, y esto no es una ventaja secundaria, KNIME Analytics Platform puede funcionar como una plataforma de integración a la que se pueden conectar muchas otras herramientas de BI y ciencia de datos. Entonces, no solo es posible, sino incluso fácil, analizar datos con KNIME Analytics Platform y luego construir cuadros de mando sobre los mismos datos procesados con una herramienta de BI diferente.

Aunque KNIME Analytics Platform es muy simple e intuitiva en su uso, cualquier principiante se beneficiaría de una orientación acelerada a través de todos los nodos, categorías y configuraciones. Este libro representa la suerte del principiante, porque tiene como objetivo ayudar a cualquier principiante a preparar su proceso de aprendizaje. Este libro no pretende ser una guía exhaustiva de todo el software KNIME. No cubre implementaciones bajo KNIME Server, que no es de código abierto, o temas que se consideran avanzados. Las variables de flujo, por ejemplo, y las implementaciones de consultas de bases de datos SQL se analizan en el libro siguiente. “[KNIME Advanced Luck](#)”.

Este libro está dividido en siete capítulos. El primer capítulo cubre los conceptos básicos de KNIME Analytics Platform, mientras que el capítulo dos lleva al lector de la mano a la implementación de la primera aplicación KNIME. Desde el tercer capítulo, comenzamos la

exploración de conceptos de ciencia de datos de una manera más profunda. De hecho, el tercer capítulo explica cómo realizar una exploración y visualización de datos básica, en términos de nodos y flujo de procesamiento. El capítulo cuatro está dedicado al modelado de datos. Cubre algunos enfoques demostrativos del aprendizaje automático, Naïve Bayes, árboles de decisión y redes neuronales artificiales. Por último, los capítulos cinco y seis están dedicados a la presentación de informes. Por lo general, los resultados de una investigación basada en la visualización de datos o, en una fase posterior, en el modelado de datos, deben mostrarse en algún momento a los colegas, la gerencia, los directores, los clientes o los trabajadores externos. Por lo tanto, la presentación de informes es una fase muy importante al final del proceso de análisis de datos. El capítulo cinco muestra cómo preparar los datos para exportarlos a un informe, mientras que el capítulo seis muestra cómo crear el informe en sí.

Cada capítulo guía al lector a través de un [ETL](#) o un proceso de aprendizaje automático (ML) paso a paso. Cada paso se explica en detalle y ofrece algunas explicaciones sobre los empleos alternativos de los nodos actuales. Al final de cada capítulo se proponen al lector varios ejercicios para poner a prueba y perfeccionar lo aprendido hasta el momento.

Los ejemplos y ejercicios de este libro se han implementado utilizando KNIME 4.3. También deberían funcionar con versiones posteriores de KNIME, aunque puede haber ligeras diferencias en su apariencia.

## 1.2. Comunidad KNIME

Al ser un software de código abierto, KNIME Analytics Platform se beneficia de una serie de foros y grupos de usuarios de KNIME en todo el mundo. Esta es una buena red de seguridad para consejos, sugerencias y material de aprendizaje. Informamos a continuación los sitios y grupos más populares.

### Páginas web útiles

<http://www.knime.com>

La página raíz del sitio web de KNIME.

<https://www.knime.com/software-overview>

El primer lugar para buscar una descripción general de todos los productos KNIME. La plataforma de análisis KNIME de código abierto se puede descargar aquí.

<https://www.knime.com/knime-self-paced-courses>

Esta es la página raíz de los cursos de autoaprendizaje de KNIME (e-learning) donde puede aprender más sobre las funcionalidades específicas de KNIME. Cubre todo el ciclo de creación de la ciencia de datos, desde el acceso y la exploración de datos hasta el aprendizaje automático y las estructuras de control.

<a href="https://hub.knime.com/">https://hub.knime.com/</a>	KNIME Hub es un repositorio público de material KNIME: flujos de trabajo, ejemplos, extensiones, plantillas de componentes y nodos. Si desea comenzar con ejemplos prácticos, aquí puede encontrar muchos según el término de búsqueda que ingrese.
<a href="https://forum.knime.com/">https://forum.knime.com/</a>	Lo que encuentro particularmente útil es el Foro KNIME. Aquí puede hacer preguntas sobre cómo utilizar KNIME Analytics Platform o sobre cómo ampliarla con nuevos nodos. Alguien de la comunidad KNIME responde siempre y rápidamente.

## Cursos, Eventos y Videos

<b>Cursos de KNIME Analytics Platform</b>	KNIME ofrece periódicamente cursos presenciales y en línea para el software KNIME. Esto incluye elementos básicos y avanzados. Para verificar la próxima fecha disponible y registrarse, simplemente vaya a la página web de KNIME Events <a href="https://www.knime.com/learning/events">https://www.knime.com/learning/events</a> y seleccione la carpeta Online Courses.
<b>KNIME Webinars</b>	Con frecuencia se encuentran disponibles varios seminarios web sobre temas específicos, como nodos de química, minería de texto, integración con otras herramientas de análisis, aprendizaje automático automatizado, aprendizaje profundo, análisis de series de tiempo, mejores prácticas, etc. Para conocer los próximos seminarios web programados, consulte la página web de KNIME Events en <a href="https://www.knime.com/learning/events">https://www.knime.com/learning/events</a> y seleccione el tab de Webinars.
<b>KNIME Meetups y KNIME Summits</b>	KNIME Meetups y KNIME Data Talks se llevan a cabo periódicamente en todo el mundo. Estas son buenas oportunidades para aprender más sobre el software KNIME, inspirarse en nuevos proyectos de ciencia de datos y conocer a otras personas de la comunidad KNIME. ( <a href="https://www.knime.com/learning/events">https://www.knime.com/learning/events</a> )
<b>Canal KNIME TV en YouTube</b>	KNIME tiene su propio canal de videos en YouTube, llamado KNIMETV. Allí, una serie de videos están disponibles para aprender más sobre muchos temas diferentes y especialmente para actualizarse sobre las nuevas funciones en los nuevos lanzamientos de KNIME. ( <a href="http://www.youtube.com/user/KNIMETV">http://www.youtube.com/user/KNIMETV</a> )

## Libros

### Funciones avanzadas en la plataforma de análisis KNIME

Para uso avanzado:

Rosaria Silipo, Jeanette Prinz, "KNIME Advanced Luck"  
(<https://www.knime.com/knimepress/advanced-luck>)

### Reporting Suite

KNIME Reporting Suite se basa en BIRT, otra herramienta de código abierto para informes. Aquí hay una guía básica sobre cómo usar BIRT:  
*D. Peh, N. Hague, J. Tatchell, "BIRT. A field Guide to Reporting.", Addison-Wesley, 2008*

### Data Science y KNIME

Para obtener una descripción general de la ciencia de datos, la minería de datos y el análisis de datos, consulte:

*Berthold M.R., Borgelt C., Höppner F., Klawonn F., Silipo R., "[Guide to intelligent data science](#)", Springer 2020.*

## KNIME Hub

Existe un lugar privilegiado donde encontrar información sobre los nodos KNIME y flujos de trabajo, así como ejemplos para sus próximos proyectos: el KNIME Hub (<https://hub.knime.com/>).

KNIME Hub es un repositorio de aplicaciones, componentes y nodos para reciclar, reutilizar y ensamblar en KNIME Analytics Platform. O como dice en la página de inicio: KNIME Hub es "el lugar para buscar y colaborar en flujos de trabajo y nodos de KNIME. Aquí puede encontrar soluciones para sus preguntas sobre ciencia de datos".

Cuando accede a KNIME Hub por primera vez, se encontrará con figura 1.1. Esta página ofrece algunos enlaces a la documentación de la guía de inicio, el Foro KNIME y el blog KNIME. Lo más importante es que en la parte superior ofrece un cuadro de búsqueda para encontrar aplicaciones, nodos y componentes cargados por los usuarios de KNIME en este lugar.

### 1.1. KNIME Hub home page <https://hub.knime.com/>

The screenshot shows the KNIME Hub homepage with a search bar at the top containing the text "Customer Intelligence". Below the search bar, there are four large numerical counts: 3 991 Nodes, 446 Components, 3 876 Workflows, and 209 Extensions. Three cards are displayed below these counts: "KNIME Examples" (Workflow examples), "KNIME Space Blueprints" (Blueprints for finance, accounting, and auditing), and "Verified Components" (Easily reuse bundled functionalities verified by KNIME Experts). The search results for "Customer Intelligence" are not visible in this specific screenshot.

Si escribimos “Customer Intelligence” en el cuadro de búsqueda, encontraremos una lista de nodos y flujos de trabajo relacionados con Customer Intelligence. Seleccionemos solo “Flujos de trabajo” en el menú superior. Luego, a continuación, en la figura 1.2, puede ver la lista de aplicaciones (flujos de trabajo) que implementan algunos aspectos de la inteligencia del cliente, y que están debidamente etiquetadas, tal como las cargan los usuarios de la comunidad KNIME. De hecho, puede cargar sus propias aplicaciones desarrolladas en KNIME Hub. Todo lo que necesita es una cuenta con el [KNIME Forum](#).

## 1.2. Lista de aplicaciones (workflows) relacionados (y etiquetados) con Customer Intelligence disponibles en el KNIME Hub

The screenshot shows the KNIME Hub interface with a search bar containing 'Customer Intelligence'. Below the search bar, it displays '405 results'. The results are categorized by type: All, Nodes, Components, Workflows, and Extensions. Under the 'Workflows' category, four workflows are listed:

- B2B Customer Intelligence Use Case**: Showcases tools and methods available for the Citizen Data Scientist to improve and predict B2B customer behaviour. It includes a green circular badge with the letter 'A'.
- Training a Churn Predictor**: An example of how to build a basic PMML model for a churn prediction using a Decision Tree algorithm. It includes a profile picture of a woman.
- Basic Customer Segmentation**: Implements a basic customer segmentation through a clustering procedure. It includes a profile picture of a man.
- Customer Segmentation**: Performs customer segmentation by means of clustering k-Means node. It includes a profile picture of a man.

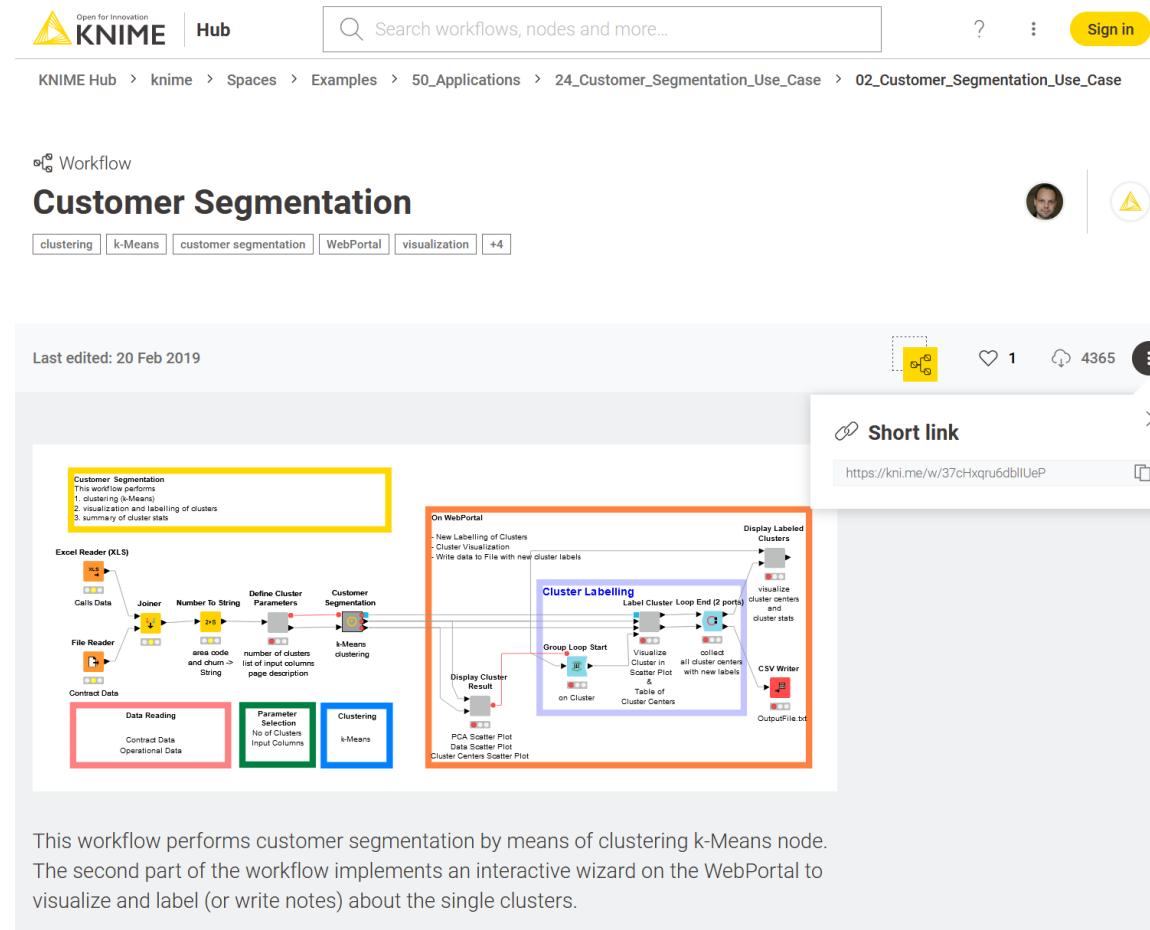
Al hacer clic en una de las aplicaciones de la lista, se abre la página web correspondiente (Fig. 1.3), con una imagen explicativa del flujo de trabajo implementado.

En la esquina superior derecha, puede ver el botón para iniciar sesión con la cuenta de KNIME Forum. Estar conectado le permite cargar, descargar, comentar, dar me gusta y actualizar los espacios para los que tiene permisos. Debajo de eso, puede encontrar la imagen del autor y debajo una serie de botones de utilidad: para descargar el flujo de trabajo, darle Me gusta, arrastrarlo y soltarlo en su instalación de KNIME Analytics Platform, y el enlace permanente corto a este flujo de trabajo para compartir.

Si pasa el cursor sobre la imagen del autor y tiene permisos de edición para este espacio Hub, aparecerá un bolígrafo. Al hacer clic en él, podrá permitir que otros usuarios de KNIME carguen y cambien este espacio.

El KNIME Hub es un repositorio de workflows, pero también de nodos, componentes y extensiones de KNIME.

### 1.3. La página dedicada a la aplicación denominada "Segmentación de clientes" en el KNIME Hub, con enlace corto <https://kni.me/w/37chxqr6dbllUeP>



## 1.3. Descarga e Instalación de KNIME Analytics Platform

Hay dos productos KNIME disponibles:

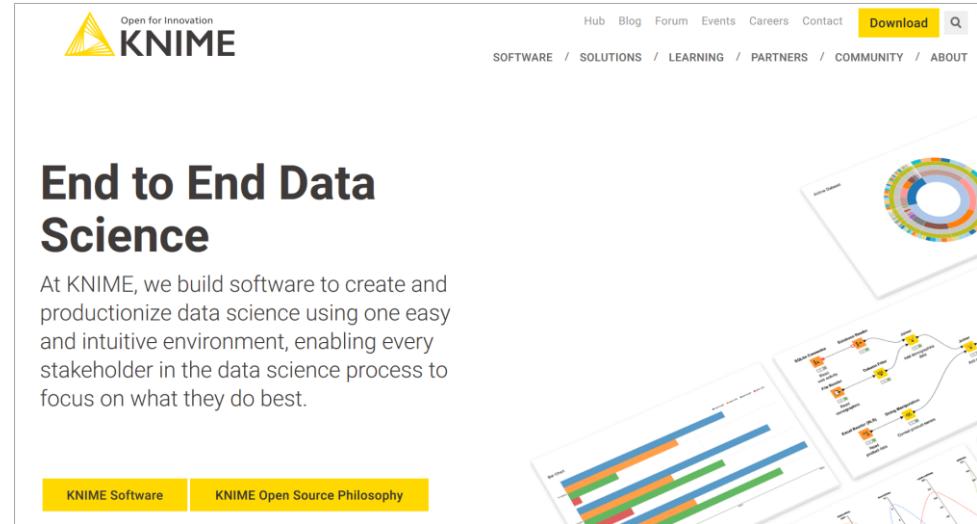
- la [KNIME Analytics Platform \(código abierto\)](https://www.knime.com/software-overview), que puede ser descargada de manera gratuita en: <https://www.knime.com/software-overview> bajo licencia GLP version 3.
- el [KNIME server](https://www.knime.com/knime-server), que se describe en: <https://www.knime.com/knime-server>

Desde el punto de vista analítico, las funcionalidades de los dos productos son las mismas. El servidor KNIME además incluye una serie de funciones TI útiles para la colaboración en equipo, la implementación y la gestión del flujo de trabajo empresarial, el almacenamiento de datos, la integración y la escalabilidad para el laboratorio de ciencia de datos. En este libro, sin embargo, trabajaremos con KNIME Analytics Platform (código abierto). Para comenzar a jugar con KNIME Analytics Platform, primero, debe descargarlo en su máquina.

### Como descargar KNIME Analytics Platform

- Vaya a [www.knime.com](https://www.knime.com)
- En la esquina superior derecha de la página principal, haga clic en "Descargar".
- Proporcione un poco de información sobre usted (que se agradece), luego continúe con el paso 2 "Descargar KNIME"
- Elija la versión que se adapte a su entorno (Windows / Mac / Linux, 32 bits / 64 bits, con o sin instalador para Windows), incluyendo opcionalmente todas las extensiones gratuitas
- Acepta los términos y condiciones.
- Empiece a descargar. Terminará con un archivo comprimido (\*.zip), un archivo de almacenamiento autoextraíble (\*.exe) o una aplicación de instalación

### 1.4. La página web de KNIME



- Para archivos .zip y .exe, simplemente descomprimalos en la carpeta de destino. Si seleccionó la versión del instalador, simplemente ejecútela y siga las instrucciones del instalador.

## 1.4. Workspace (Ambiente de Trabajo)

Para iniciar KNIME Analytics Platform, abra la carpeta donde se instaló KNIME y ejecute knime.exe (o knime en una máquina Linux / Mac). Si ha instalado KNIME utilizando el instalador, puede simplemente hacer clic en el ícono en su escritorio o en el menú principal de Windows.

Si está iniciando KNIME Analytics Platform por primera vez, se le preguntará si desea compartir sus estadísticas de uso con KNIME. Estos números se utilizarán para impulsar el motor de recomendaciones de mejores prácticas proporcionado dentro del KNIME Analytics Platform workbench: Workflow Coach. Ninguna información personal llegará a KNIME y sus estadísticas anónimas nunca se compartirán.

Después de la pantalla de presentación, la ventana “Workspace Launcher” requiere que ingrese la ruta del workspace

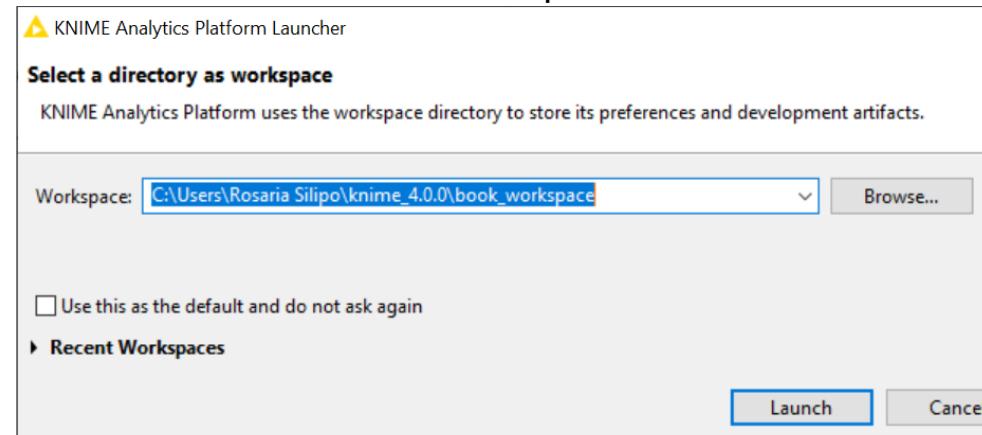
# La ventana “Workspace Launcher”

El **workspace** es una carpeta donde todas las preferencias y aplicaciones (flujos de trabajo), tanto desarrolladas como actualmente en desarrollo, se guardan para la próxima sesión de KNIME.

La carpeta del espacio de trabajo se puede ubicar en cualquier lugar del disco duro.

De forma predeterminada, la carpeta del espacio de trabajo es ".. \ knime-workspace". Sin embargo, puede cambiarlo fácilmente, cambiando la ruta propuesta en la ventana “KNIME Analytics Platform Launcher”, antes de comenzar la sesión de trabajo de KNIME.

## 1.5. La ventana “Workspace Launcher”



Una vez que la KNIME Analytics Platform se haya abierto, desde dentro del workbench KNIME puede cambiar a otra carpeta del espacio de trabajo, seleccionando “File” en el menú superior y luego “Switch Workspace”. Después de seleccionar el nuevo espacio de trabajo, KNIME Analytics Platform se reinicia y muestra la lista de flujo de trabajo del espacio de trabajo recién seleccionado. Tenga en cuenta que si la carpeta del espacio de trabajo no existe, se creará automáticamente.

Si se tiene una gran cantidad de clientes, por ejemplo, puedo usar un workspace diferente para cada uno de ellos. Esto mantiene mi espacio de trabajo limpio y ordenado y me protege de mezclar información por error. Para este proyecto utilicé el espacio de trabajo “KNIME\_4.3.1 \ book\_workspace”.

## 1.5. KNIME Workflow (flujo de trabajo)

KNIME Analytics Platform no funciona con scripts, funciona con flujos de trabajo gráficos.

Pequeños íconos, llamados nodos, están dedicados cada uno a implementar y ejecutar una tarea determinada. Una secuencia de nodos crea un flujo de trabajo para procesar los datos para alcanzar el resultado deseado.

### ¿Qué es un workflow?

Un workflow es un **flujo de análisis**, es decir, una **secuencia de pasos de análisis** necesarios para alcanzar un resultado dado. Es la tubería del proceso de análisis, algo como:

- Paso 1. Leer datos
- Paso 2. Limpiar datos
- Paso 3. Filtrar datos
- Paso 4. Entrena un modelo

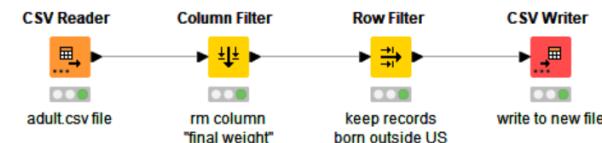
KNIME Analytics Platform implementa sus workflows de forma gráfica. Cada paso del análisis de datos se implementa y ejecuta a través de una pequeña caja, llamada nodo. Una secuencia de nodos crea un flujo de trabajo

En el KNIME whitepaper [1] un workflow is definido como: *"Los flujos de trabajo en KNIME son gráficos que conectan nodos, o más formalmente, gráficos acíclicos directos (DAG)." ([http://www.kdd2006.com/docs/KDD06\\_Demo\\_13\\_Knime.pdf](http://www.kdd2006.com/docs/KDD06_Demo_13_Knime.pdf))*

Below is an example of a KNIME workflow, with:

- a node to read data from a file
- a node to exclude some data columns
- a node to filter out some data rows
- a node to write the processed data into a file

#### 1.6. Ejemplo de un KNIME "workflow"



**Nota.** Un flujo de trabajo es una secuencia de análisis de datos, que en un lenguaje de programación tradicional se implementaría mediante una serie de instrucciones y llamadas a funciones. KNIME Analytics Platform lo implementa gráficamente. Esta representación gráfica es más intuitiva de usar, le permite mantener una descripción general del proceso de análisis y también sirve para la documentación.

## ¿Que es un nodo?

Un nodo (node) en una unidad simple de procesamiento de un workflow

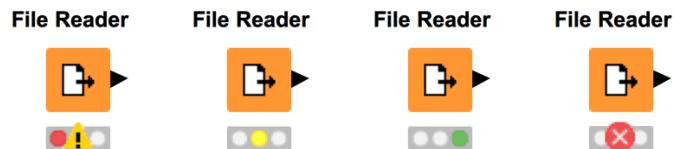
Un nodo toma un conjunto de datos como entrada, lo procesa y lo pone a disposición en su puerto de salida. La acción de "procesamiento" de un nodo varía desde el modelado, como un nodo de aprendizaje de red neuronal artificial, hasta la manipulación de datos, como la transposición de la matriz de datos de entrada, desde herramientas gráficas, como un diagrama de dispersión, hasta operaciones de lectura / escritura. Cada nodo en KNIME puede tener 4 estados:

- Inactivo y no configurado → **red** light
- Configurado pero no ejecutado → **yellow** light
- Ejecutado exitosamente → **green** light
- Ejecutado con errores → **red with cross** light

Nodos que contienen otros nodos se llaman **metanodes** or **components**.

A continuación se muestran cuatro ejemplos del mismo nodo (un nodo Lector de archivos) en cada uno de los cuatro estados.

1.7. Nodo lector "File Reader" con diferentes estados



## 1.6. Archivos .knwf and .knar

Los flujos de trabajo de KNIME se pueden empaquetar y exportar en archivos .knwf o .knar. Un archivo .knwf contiene solo un flujo de trabajo, mientras que un archivo .knar contiene un grupo de flujos de trabajo. Dichas extensiones están asociadas con KNIME Analytics Platform. Un doble clic abre el flujo de trabajo dentro de KNIME Analytics Platform.

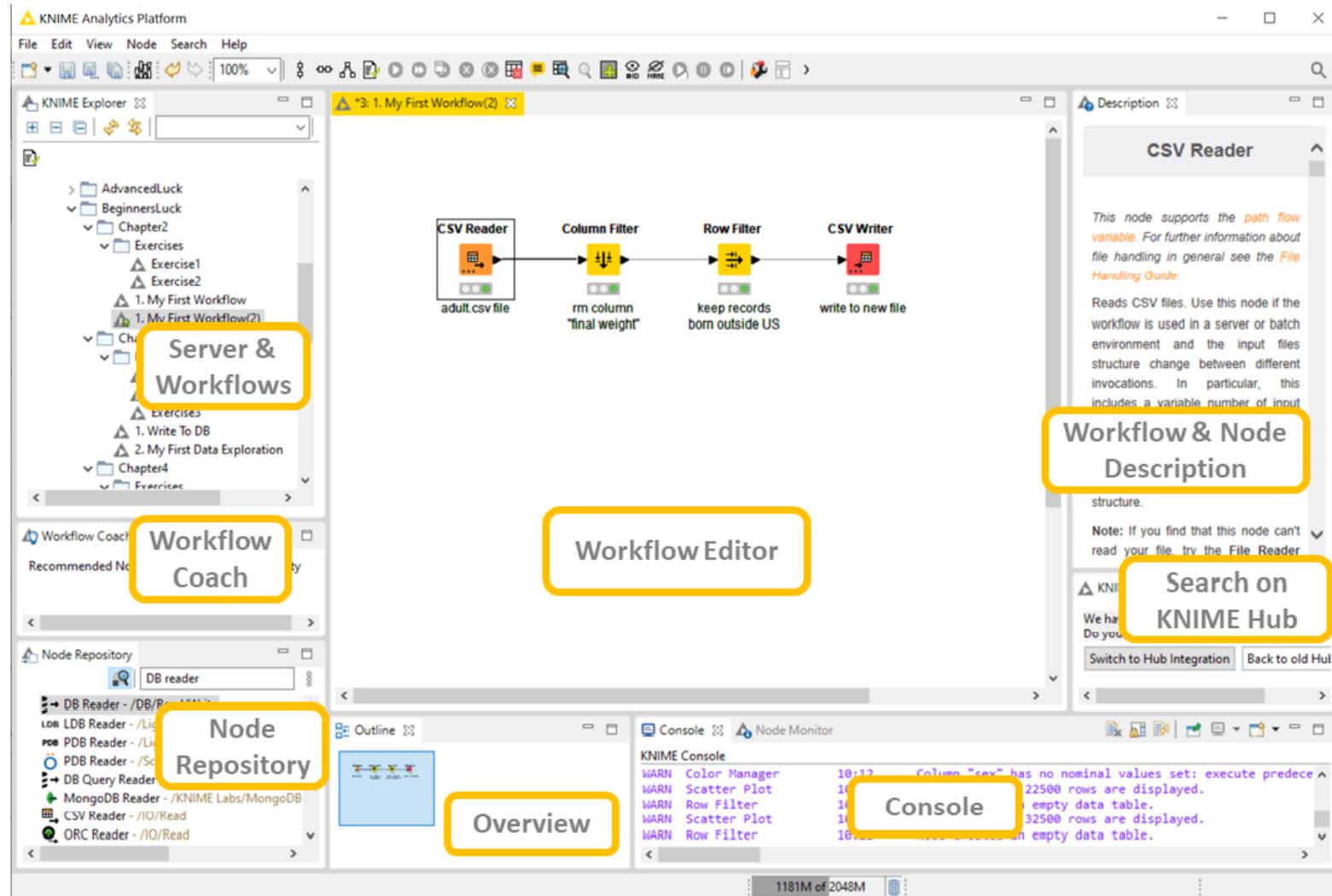
## 1.8. Los archivos .knwf and .kna files estan asociados a la plataforma KNIME Analytics

▲ 01_From.Strings_to.Documents.knwf	10/4/2017 9:45 AM	KNIME Workflow ...	18,619 KB
▲ 04_Interaction.Graph.knwf	9/29/2017 8:20 AM	KNIME Workflow ...	9,465 KB
▲ 06_REST_Examples_Google_Geocode.knwf	7/29/2017 7:09 PM	KNIME Workflow ...	62 KB
▲ 06_Semantic.Web_updated.knar	11/3/2016 2:24 PM	KNIME Archive File	178 KB
▲ AzureDemoWorkflowArchive.knar	5/5/2017 11:24 AM	KNIME Archive File	24,104 KB
▲ Building.a.Simple.Classifier_.knwf	2/18/2017 5:46 PM	KNIME Workflow ...	43 KB
▲ Cookbook_Ch5.knar	11/24/2017 10:03 ...	KNIME Archive File	477 KB
▲ Cookbook_Ch6.knar	11/24/2017 10:26 ...	KNIME Archive File	155 KB
▲ Corsair.knwf	7/10/2017 4:20 PM	KNIME Workflow ...	106 KB

## 1.7. KNIME workbench (banco de trabajo)

Después de aceptar la ruta del espacio de trabajo, el KNIME workbench se abre en una página de "Bienvenido a KNIME". Esta página proporciona algunos enlaces para comenzar, como por ejemplo al KNIME Hub, a cierta documentación básica, a los cursos y eventos actuales, a las actualizaciones disponibles, etc. El "KNIME Workbench" consta de un menú superior, una barra de herramientas y algunos paneles. Los paneles se pueden cerrar, volver a abrir y mover.

## 1.9. EL KNIME workbench



# El KNIME Workbench

**Top Menu:** File, Edit, View, Node, Help

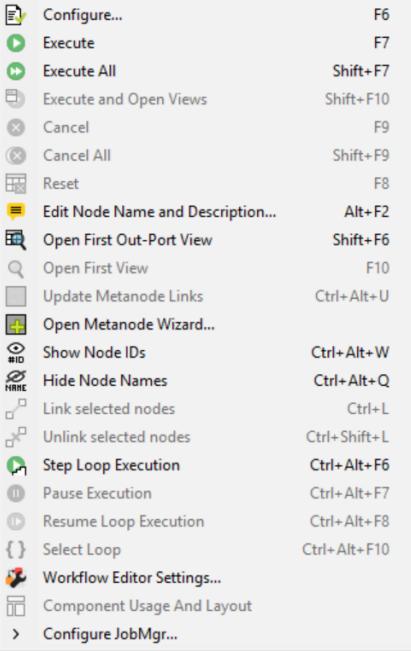
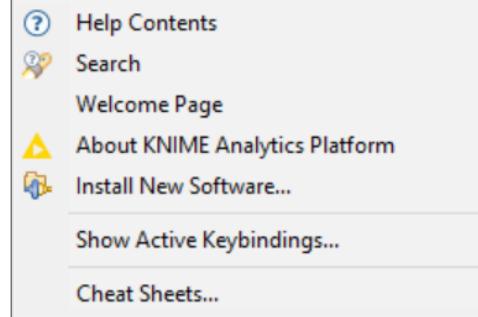
**Tool Bar:** New, Save (Save As, Save All), Undo/Redo, Open Report (if reporting was installed), zoom (in %), Align selected nodes vertically/vertically, Auto layout, Configure, Execute options, Cancel execution options, Reset, Edit node name and description, Open node's first output port table, Open node's first view, Open the "Add Meta node" Wizard, Append IDs to node names, Hide all node names, Loop execution options, Change Workflow Editor Settings, Edit Layout in Components, configure job manager.

KNIME Explorer	Workflow Editor	Node Description
Este panel muestra la lista de proyectos de workflows disponibles en el espacio de trabajo seleccionado (LOCAL), en el servidor de EJEMPLOS, en My-KNIME-Hub (su propio espacio en KNIME Hub) o en otros servidores KNIME conectados.	<p>El área central consta del propio "Workflow Editor".</p> <p>Se puede seleccionar un nodo del panel "Node Repository" "Repositorio de nodos" y arrastrarlo y soltarlo aquí, en el panel "Workflow Editor"</p> <p>Los nodos se pueden conectar haciendo clic en el puerto de salida de un nodo y soltando el mouse en el puerto de entrada del siguiente nodo o en el siguiente nodo.</p>	Si se selecciona un nodo o un flujo de trabajo, este panel muestra una descripción resumida de las funcionalidades del nodo o la meta información del flujo de trabajo.
Workflow Coach		<b>Search box for KNIME Hub</b> Para buscar material en KNIME Hub

<b>Node Repository</b>	<b>Outline</b>	<b>Console</b>
<p>Este panel contiene todos los nodos que están disponibles en su instalación KNIME. Es algo similar a una paleta de herramientas cuando se trabaja en un informe o con un software de diseño web. Allí usamos herramientas gráficas, mientras que en KNIME usamos herramientas de análisis de datos.</p>	<p>El panel "Outline" contiene una pequeña descripción general del contenido del "Workflow Editor". Es posible que el panel "Outline" no sea de gran interés para flujos de trabajo pequeños. Sin embargo, tan pronto como los flujos de trabajo alcancen un tamaño considerable, es posible que todos los nodos del flujo de trabajo ya no estén visibles en el "Workflow Editor" sin desplazarse. El panel "Outline", por ejemplo, puede ayudarlo a ubicar los nodos recién creados.</p>	<p>El panel "Console" muestra errores y mensajes de advertencia (warnings) al usuario.</p> <p>Este panel también muestra la ubicación del archivo de registro, que puede ser de interés cuando la consola no muestra todos los mensajes.</p> <p>También hay un botón en la barra de herramientas para mostrar el archivo de registro asociado con esta instancia de KNIME.</p>

# Menu Principal

File (archivo)	Edit (editar)	View (Ver)
 New... &nbsp; Ctrl+N  Save &nbsp; Ctrl+S  Save As...  Save All &nbsp; Ctrl+Shift+S  Close All &nbsp; Ctrl+Shift+W Recent Workflows >  Print... &nbsp; Ctrl+P  Import KNIME Workflow...  Export KNIME Workflow...  Export to SVG... Switch Workspace > Preferences  Export Preferences...  Import Preferences... Install KNIME Extensions... Update KNIME... Restart Exit	 Undo &nbsp; Ctrl+Z  Redo &nbsp; Ctrl+Y  Cut &nbsp; Ctrl+X  Copy &nbsp; Ctrl+C  Paste &nbsp; Ctrl+V  Delete &nbsp; Delete Select All &nbsp; Ctrl+A	 Console &nbsp; Alt+Shift+Q, C  Description  Error Log &nbsp; Alt+Shift+Q, L  KNIME Explorer  KNIME Hub Search  Node Monitor  Node Repository  Outline &nbsp; Alt+Shift+Q, O  Workflow Coach Other... &nbsp; Alt+Shift+Q, Q Reset Perspective...  Quick Node Insertion... &nbsp; Ctrl+Space  Open KNIME log
<b>File</b> incluye los comandos de archivo tradicionales, como "Nuevo" y "Guardar", además de algunos comandos específicos de KNIME, como: <ul style="list-style-type: none"> <li>- Importar / Exportar flujo de trabajo KNIME ...</li> <li>- Exportar a SVG</li> <li>- Cambiar espacio de trabajo</li> <li>- Preferencias con preferencias de exportación / importación</li> <li>- Instalar extensiones KNIME</li> </ul>	<b>Edit</b> contiene comandos de edición.  Deshacer y Rehacer se refieren a las últimas acciones realizadas.  Cortar, Copiar, Pegar y Eliminar hacen referencia a los nodos seleccionados en el flujo de trabajo.	<b>View</b> contiene la lista de todos los paneles que se pueden abrir en el banco de trabajo KNIME.  Aquí se puede volver a abrir un panel cerrado.  Además, cuando se estropea la disposición del panel, la opción "Restablecer perspectiva" recrea el diseño del panel original cuando se inició el banco de trabajo por primera vez.

- Actualizar KNIME	Seleccionar todo selecciona todos los nodos del flujo de trabajo en el editor de flujo de trabajo.	La opción "Otro" abre vistas adicionales útiles para personalizar el banco de trabajo.
	<b>Node</b>	<b>Help</b>
		
<b>Node</b> se refiere a todas las operaciones posibles que se pueden realizar en un nodo. Un nodo puede estar:	<p><b>Help Contents</b> proporciona ayuda general sobre Workbench, BIRT y KNIME.</p> <p><b>Search</b> abre un panel a la derecha del panel "Descripción de nodo" para buscar temas de ayuda o nodos específicos.</p> <p><b>Welcome Page</b> (re abre) la Welcome Page</p> <p><b>Install New Software</b> es la puerta para instalar KNIME Extensions desde KNIME Update sites.</p>	
<ul style="list-style-type: none"> <li>- Configured (Configurado)</li> <li>- Executed (Ejecutado)</li> <li>- Cancelled (Cancelado -detenido durante la ejecución)</li> <li>- Reset (restablece los resultados de la última operación "Ejecutar")</li> <li>- Given a name and description</li> <li>- Set to show its View (if any)</li> </ul>		

- Reset (restablece los resultados de la última operación "Ejecutar")
- Dado un nombre y una descripción
- Establecer para mostrar su vista (si corresponde)

Las opciones solo están activas si son posibles. Por ejemplo, un nodo ya ejecutado con éxito no se puede volver a ejecutar a menos que se restablezca primero o se haya cambiado su configuración. Entonces las opciones "Cancel" and "Execute" están inactivas.

La opción "Open Meta Node Wizard" inicia el asistente para crear un nuevo metanodo en el editor de flujo de trabajo.

**Show Active Keybindings** resume todos los comandos del teclado para el editor de flujo de trabajo.

**Cheat Sheets** ofrece tutoriales sobre temas específicos: la herramienta de informes, csv, Plug-ins.

Veamos ahora los elementos que se utilizan con más frecuencia en Top Menu.

**“File” → “Import KNIME workflow”** reads and copies workflows into the current workspace.

Opción “Select root directory”, copia el workflow directamente del folder el workspace actual (LOCAL).

Opción “Select archive file”, lee un workflow desde una archivo .knwf or .knar en el Los archivos .knwf ./knaar pueden ser creados con la opcion “File”→ “Export KNIME workflow”.

**“File” → “Export KNIME workflow”** exporta el workflow seleccionado a un archivo .knaar

Opción “Reset Workflow(s) before export” exporta los workflows sin la data producida por la ejecucion de los nodos. Esto genera archivos exportables significativamente más pequeños.

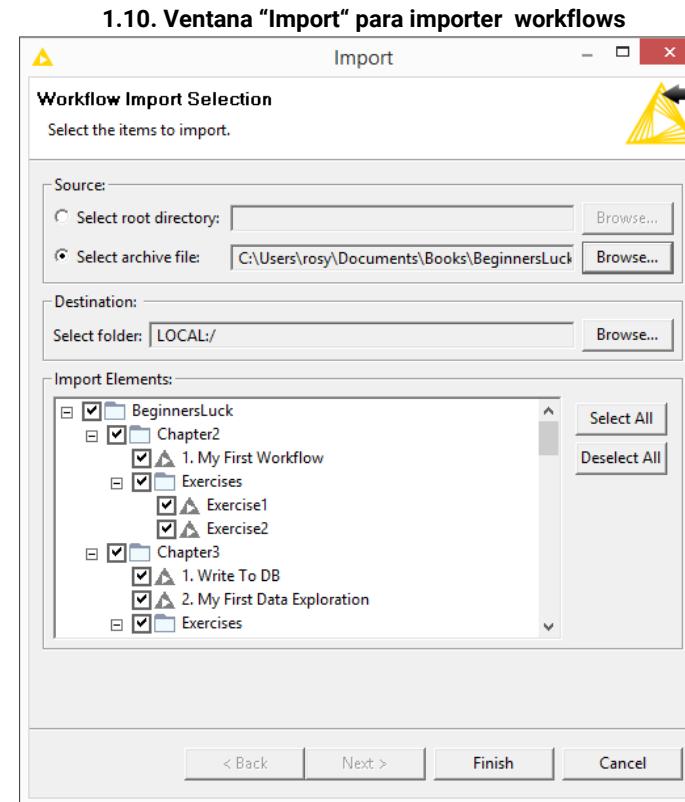
Copiar workflows de una carpeta a otra puede crear ciertos problemas relacionados con las actualizaciones internas de KNIME. Copiar workflows usando la opción “Import KNIME workflow” o mediante double-click es definitivamente más seguro.

**“File” → “Install KNIME Extensions”** y **“Help” → “Install New Software”** ambos abre un asistente para la instalacion de extensiones de KNIME (ver proximas secciones)

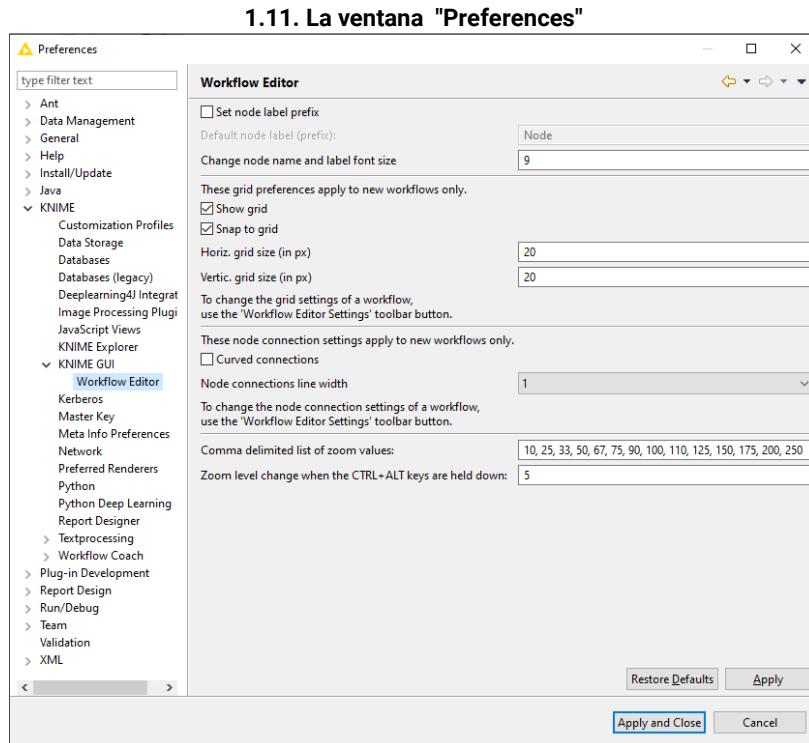
**“File” → “Switch Workspace”** cambia el workspace actual por uno nuevo.

**“File” → “Preferences”**. Abre una ventana donde se pueden personalizar todos los ajustes de KNIME. Se pueden encontrar en el elemento “KNIME”. Vamos a comprobarlos.

- **Chemistry** tiene configuraciones relacionadas con los renderizadores KNIME en los mòdulos de química.



- **Databases** especifica la ubicación de controladores de bases de datos específicos, que aún no están disponibles en KNIME. De hecho, los controladores de bases de datos más comunes y más recientes ya están disponibles en el menú de controladores de los nodos de la base de datos. Sin embargo, si necesita algún archivo de controlador específico, puede establecer su ruta aquí.



- **KNIME Explorer** contiene la lista de los repositorios compartidos a través de KNIME Server.
- **KNIME GUI** permite la personalización de las opciones y el diseño del banco de trabajo KNIME a través de una serie de configuraciones.
- **Master Key** contiene la clave maestra que se utilizará en los nodos con una opción de cifrado, como los nodos de conexión de la base de datos. Dado que las contraseñas de la base de datos KNIME version 2.3 se pasan a través de las variables de flujo de trabajo "Credentials" y la preferencia de clave maestra ha quedado obsoleta. Aún puede encontrarlo en el menú Preferencias para compatibilidad con versiones anteriores.
- En **Meta Info Preferences** puede cargar una plantilla de metainformación para nodos y flujos de trabajo.
- Aquí puede encontrar la configuración de preferencias para los modulos externos, como: H2O, R, Report Designer, Perl, Perl, Open Street Map y otros si los tiene instalados. En particular, para los scripts externos, esta página ofrece la opción de establecer la ruta a la instalación del script de referencia.
- Finalmente, **Workflow Coach** contiene el conjunto de datos que se utilizará para el motor de recomendación de nodos: la comunidad, un espacio de trabajo del servidor o su propio espacio de trabajo local.

Para todos los amantes del teclado, la mayoría de los comandos de KNIME también se pueden ejecutar a través de teclas de acceso rápido. Todas las teclas de acceso rápido se enumeran en los menús de KNIME al lado de los comandos correspondientes o en los mensajes de información sobre herramientas de los iconos de la barra de herramientas en el menú superior. Estas son las teclas de acceso rápido que se utilizan con más frecuencia.

## Hotkeys (teclas importantes)

### Configuración de nodos

- **F6** abre la ventana de configuración del nodo seleccionado

### Ejecución de nodos

- **F7** ejecuta los nodos configurados seleccionados
- **Shift + F7** ejecuta todos los nodos configurados
- **Shift + F10** ejecuta todos los nodos configurados y abre todas las vistas

### Para detener la ejecución de nodos

- **F9** cancela los nodos en ejecución seleccionados
- **Shift + F9** cancela todos los nodos en ejecución seleccionados

### Para mover nodos

- **Ctrl + Shift + Arrow** mueve el nodo seleccionado en la dirección de la flecha

### Reseteo de Nodos

- **F8** restablece los nodos seleccionados

### Grabar Workflows

- **Ctrl + S** graba el workflow
- **Ctrl + Shift + S** graba los workflows abiertos
- **Ctrl + Shift + W** cierra todas las ventanas de workflows

### Meta-Node

- **Shift + F12** abre el asistente de metanodo

### To move Annotations

- **Ctrl + Shift + PgUp/PgDown** mueve la anotación seleccionada al frente o al final de todas las anotaciones superpuestas

## Repositorio de Nodos (Node Repository)

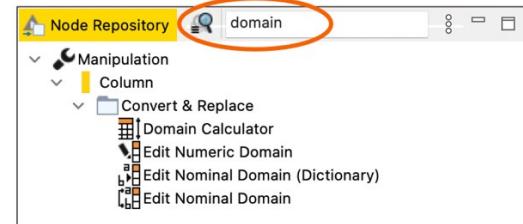
En la esquina inferior izquierda encontramos el Repositorio de nodos, que contiene todos los nodos instalados organizados en categorías y subcategorías. KNIME Analytics Platform ha acumulado hasta ahora más de 1500 nodos. Se ha vuelto difícil recordar la ubicación de cada nodo en el repositorio de nodos. Para resolver este problema, hay dos opciones de búsqueda disponibles: por coincidencia exacta y por coincidencia aproximada, ambas en el cuadro de búsqueda ubicado en la parte superior del panel del repositorio de nodos.

## Cuadro de búsqueda

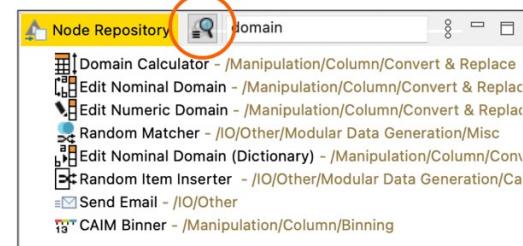
En la parte superior del panel "Repositorio de nodo" hay un cuadro de búsqueda. Si escribe una palabra clave en el cuadro de búsqueda y presiona "Enter", obtendrá la lista de nodos que contienen una coincidencia exacta de esa palabra clave. Presione la tecla "Esc" para ver todos los nodos nuevamente.

Al hacer clic en la lente a la izquierda del cuadro de búsqueda, se ejecuta un algoritmo de búsqueda difusa (fuzzy) que conduce a una lista de resultados coincidentes más amplia que la que se encuentra en la figura anterior.

1.13. Búsqueda de palabras en el panel del repositorio de nodos: modo de coincidencia exacta



1.14. Búsqueda de palabras en el panel del repositorio de nodos: modo de búsqueda difusa (fuzzy)



## KNIME Explorer (Explorador de KNIME)

En la esquina superior izquierda del banco de trabajo KNIME, encontramos el panel KNIME Explorer. Este panel contiene:

- En LOCAL los workflows que se han desarrollado en el espacio de trabajo seleccionado
- Los puntos de montaje a varios servidores KNIME
- Los workflows contenidos en el espacio de trabajo de referencia de dichos servidores.
- El acceso a My-KNIME-Hub, es decir, a su espacio en KNIME Hub. Recuerda que necesitas una cuenta en KNIME Forum para acceder a este espacio

Al principio, el panel KNIME Explorer solo contiene LOCAL, My-KNIME-Hub y EXAMPLES. Como ya dijimos, LOCAL muestra el contenido del espacio de trabajo seleccionado. EXAMPLES apunta a un servidor público de solo lectura, accesible mediante inicio de sesión anónimo.

Este servidor aloja una serie de flujos de trabajo de ejemplo que puede utilizar para iniciar un nuevo proyecto. My-KNIME-Hub permite acceder a su espacio en KNIME Hub.

Cuando abra KNIME Analytics Platform por primera vez, encontrará una carpeta denominada "Example Workflows" que contiene las soluciones para algunos casos de uso de ciencia de datos comunes.

Las carpetas en "KNIME Explorer", que contienen workflows , también se denominan "Workflow Groups".

**Nota.** El panel KNIME Explorer también puede albergar datos. Simplemente cree una carpeta debajo de la carpeta del área de trabajo, llénela con archivos de datos en la máquina y seleccione "Refresh" en el menú contextual (clic derecho) del panel "KNIME Explorer".

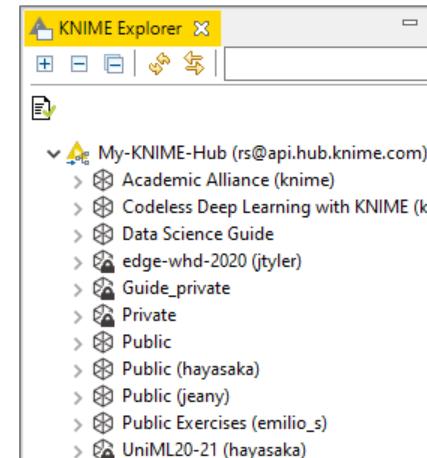
## Mi central KNIME (My-KNIME-Hub)

Desde el panel KNIME Explorer, puede acceder a sus espacios en KNIME Hub y cargar y actualizar contenido nuevo o existente allí.

De forma predeterminada, un usuario KNIME autenticado tiene un espacio público, para que el material se comparta públicamente, y un espacio privado para estacionar su propio material para su uso posterior. Sin embargo, se pueden crear nuevos espacios públicos o privados con un clic derecho en My-KNIME-Hub en el panel KNIME Explorer y luego una selección de la opción "Crear nuevo espacio..." .

De forma predeterminada, eres el único propietario de tus propios espacios. Sin embargo, al acceder a este espacio desde un navegador web, después de colocar el cursor sobre su imagen en la esquina superior derecha, aparece un bolígrafo. Esto permitirá agregar colegas y compañeros de equipo como contribuyentes al espacio.

**1.15. Panel KNIME Explorer.** En la parte superior, el contenido del servidor de EXAMPLES; debajo del contenido del espacio de trabajo LOCAL



# El "EXAMPLES Server"

Un enlace al servidor público KNIME (EXAMPLES) está disponible en el panel "KNIME Explorer". Este es un servidor proporcionado por KNIME a todos los usuarios para tutoriales y demostraciones. Allí puede encontrar una serie de ejemplos útiles sobre cómo implementar tareas específicas con KNIME. Para conectarse al servidor de EJEMPLOS:

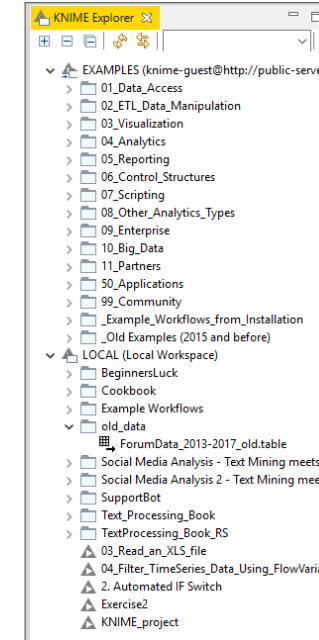
- haga doble clic en "EJEMPLOS" en el panel "KNIME Explorer"
- haga doble clic en "Haga doble clic para conectarse ..."

Debería iniciar sesión automáticamente como invitado.

Para transferir ejemplos de flujos de trabajo del servidor de EXAMPLES a su espacio de trabajo LOCAL, simplemente arrastre y suelte o copie y pegue (Ctrl-C, Ctrl-V en Windows) de "EXAMPLES" a "LOCAL".

También puede abrir los workflows de EXAMPLES en el editor de flujo de trabajo, pero solo temporalmente y en modo de solo lectura. Un cuadro de advertencia amarillo en la parte superior advierte que esta copia del flujo de trabajo no se guardará.

**1.16. Panel KNIME Explorer.** En la parte superior, el contenido del servidor de EXAMPLES; debajo del contenido del espacio de trabajo LOCAL



El panel KNIME Explorer puede albergar más de un servidor KNIME. Es suficiente agregar puntos de montaje del servidor a la lista de servidores KNIME disponibles.

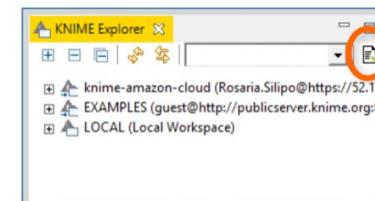
# Montaje de servidores en KNIME Explorer

Para agregar el KNIME servers al panel "KNIME Explorer" :

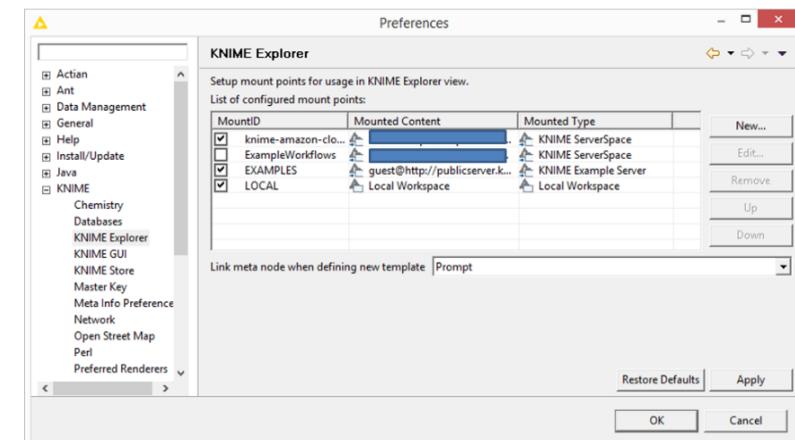
- Seleccione el panel "KNIME Explorer"
- Haga Click en "Configure Explorer View"
- La ventana The "Preferences (Filtered)" se abre en la página "KNIME Explorer" y enumera todos los espacios KNIME ya montados en esta instancia KNIME. Los tres espacios KNIME disponibles de forma predeterminada en cada instancia KNIME son el espacio de trabajo local "LOCAL", el servidor público KNIME "EXAMPLES" y el My-KNIME-Hub ubicado en el servidor KNIME Hub (hub.knime.com).
- Utilice el botón "New" y "Remove" para agregar / eliminar conexiones a servidores remotos.
- Después de hacer clic en el botón "New", complete la información requerida sobre el servidor en la ventana "Select New Content"
- Utilice el botón "Test Connection" para recuperar automáticamente el punto de montaje predeterminado para el servidor seleccionado.

La pagina KNIME Explorer "Preferences" puede ser accedida mediante "File" → "Preferences" → "KNIME Explorer".

1.17. El botón "Configure KNIME Explorer"



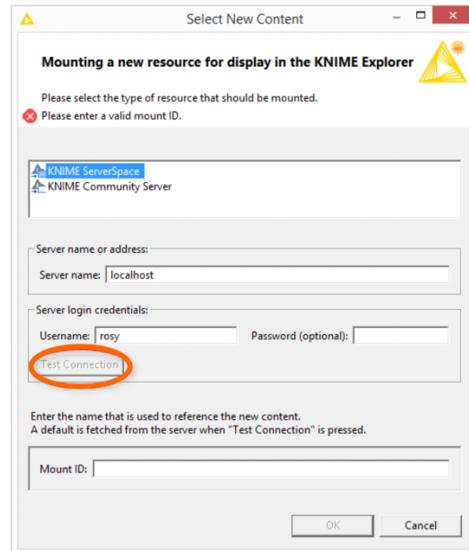
1.18. La ventana "Preferences (Filtered)"



Para iniciar sesión en cualquiera de los servidores disponibles en el panel "KNIME Explorer":

- haga clic con el botón derecho o haga doble clic en el nombre del servidor
- proporcionar las credenciales

### 1.19. La ventana “Select New Content”



## El Editor de Workflow

La pieza central del banco de trabajo KNIME consiste en el propio editor de flujo de trabajo. Este es el lugar donde se construye un flujo de trabajo agregando un nodo tras otro. Los nodos se insertan en el editor de flujo de trabajo arrastrando y soltando o haciendo doble clic desde el repositorio de nodos o el entrenador de flujo de trabajo. El proceso de construcción del flujo de trabajo se describirá ampliamente en las próximas secciones de este libro. Aquí, describiremos cómo personalizar y probablemente mejorar la función del canvas del espacio del editor de flujo de trabajo.

Describiremos dos opciones:

- cambiar la apariencia del canvas con cuadrículas y diferentes visualizaciones para las conexiones;
- introducción de anotaciones para comentar el trabajo.

#### *Agregar una cuadricula al canvas y conexiones curvas a los workflows*

Casi hacia el final, a la derecha de la barra de herramientas, puede ver el botón “Change Workflow Editor Settings”. Si hace clic en él, se abre la ventana “Workflow Editor Settings”.

1.19. Botón "Change Workflow Editor Settings" en la barra de herramientas

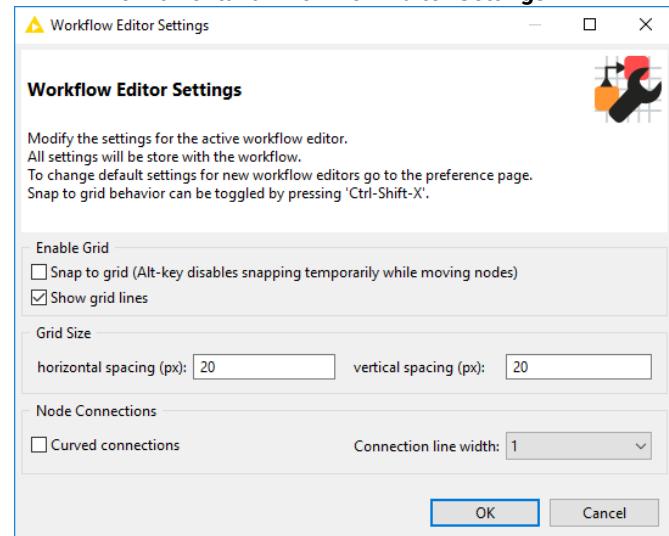


## Personalización del editor de flujo de trabajo

La función de cuadrícula contiene algunas opciones:

1. "Show grid lines" Esto muestra líneas de cuadrícula en el editor de flujo de trabajo y permite alinear mejor los nodos y anotaciones manualmente.
2. "Snap to grid". Esta opción adjunta nodos y anotaciones a la esquina disponible más cercana de la cuadrícula. Le brinda menos libertad manual, pero el resultado es más limpio y ordenado en menos tiempo.
3. "Curved Connections". Aquí puede permitir que las conexiones de nodo sigan una curva en lugar de una línea recta. Esto podría dar lugar a gráficos de flujo de trabajo más atractivos.

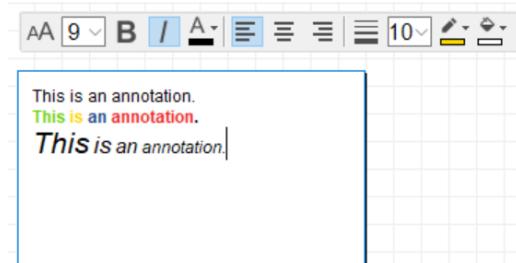
1.20. La ventana "Workflow Editor Settings"



### Agregar anotaciones al canvas

También es posible incluir anotaciones en el editor de flujo de trabajo. Las anotaciones pueden ayudar a explicar la tarea workflow y la función de cada nodo o grupo de nodos. El resultado es una descripción general mejorada, similar a la documentación, de la tarea general del workflow y de las subtareas individuales.

### 1.21. El editor Annotations



## Anotaciones del Workflow

Para insertar una nueva anotación:

- haga clic con el botón derecho en cualquier lugar del editor de workflow y "New Workflow Annotation"
- aparece un pequeño marco de color amarillo pálido; este es el marco de anotaciones predeterminado
- haga doble clic en el marco para editar su contenido
- Observe la barra de herramientas que aparece en la parte superior para editar el estilo del texto, el color de la fuente, el color de fondo, la alineación del texto y las propiedades del borde (color, grosor).
- Para volver a abrir una anotación, simplemente haga doble clic en la esquina superior izquierda, donde está el ícono de lápiz.

## Otras personalizaciones del Workbench

Otra posibilidad de personalización consiste en agregar vistas. Las vistas disponibles se encuentran en el elemento "View" del menú superior.

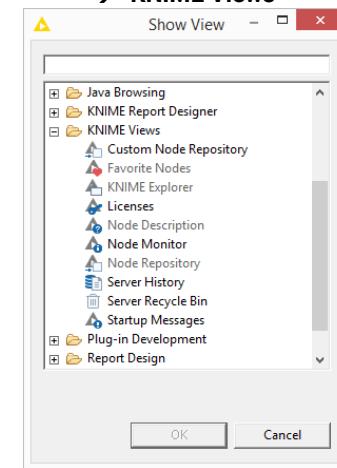
Algunas vistas populares son las vistas "Node Monitor", "Custom Node Repository", y "Licenses" y "Server" si ud esta conectado al servidor. Estas puede invocarse siguiendo el esquema: Top menu under "View" → "Other" → "KNIME Views".

"Node Monitor" ayuda, especialmente durante la fase de desarrollo, a monitorear y depurar la ejecución del flujo de trabajo.

"Custom Node Repository" permite una personalizacion del "Node Repository" con solo un conjunto de nodos.

"Licenses" permite monitorear la situacion de sus licencias .

### 1.22. Vistas adicionales "View" → "Other" → "KNIME Views"

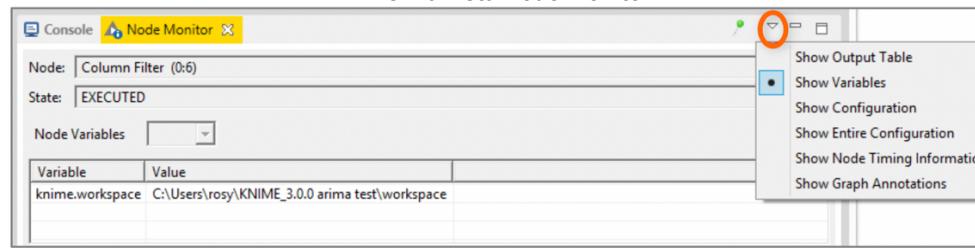


## Vista “Node Monitor”

Para insertar la vista “Node Monitor” en el workbench:

- Selecctcione “View”→ “Other...” en el menu principal
- Expanda el elemento “KNIME Views” y haga doble clic en “Node Monitor” un panel, llamado “Node Monitor ”, aparece en el lateral del panel “Console”; el panel muestra los valores de las variables de flujo de salida, los datos de salida o los valores de configuración del nodo seleccionado en el editor de flujo de trabajo.
- Aquí puede decidir qué mostrar (datos, configuración, variables), a través del menú en la esquina superior derecha.

1.23. La vista Node Monitor



## 1.8. Descargar las Extensiones de KNIME

KNIME Analytics Platform es un producto de código abierto. Como todo producto de código abierto, se beneficia de los comentarios y las funcionalidades que desarrolla la comunidad. Hay varias extensiones disponibles para KNIME Analytics Platform. Si ha descargado e instalado KNIME Analytics Platform, incluidas todas sus extensiones gratuitas, verá las categorías correspondientes en el panel Node Repository, como KNIME Labs, Text Processing, R Integration y muchas otras. Sin embargo, si en el momento de la instalación ha elegido instalar la plataforma KNIME Analytics sin las extensiones gratuitas, es posible que deba instalarlas por separado en algún momento de una instancia en ejecución.

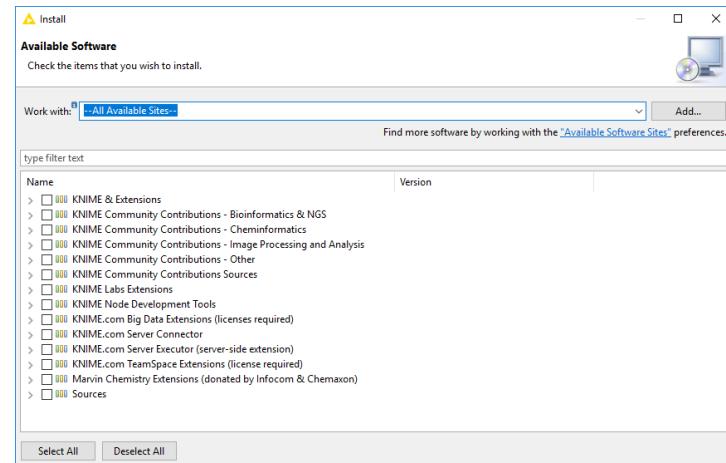
# Instación de KNIME Extensions

Para instalar una nueva extensión KNIME desde KNIME Analytics Platform, hay tres opciones.

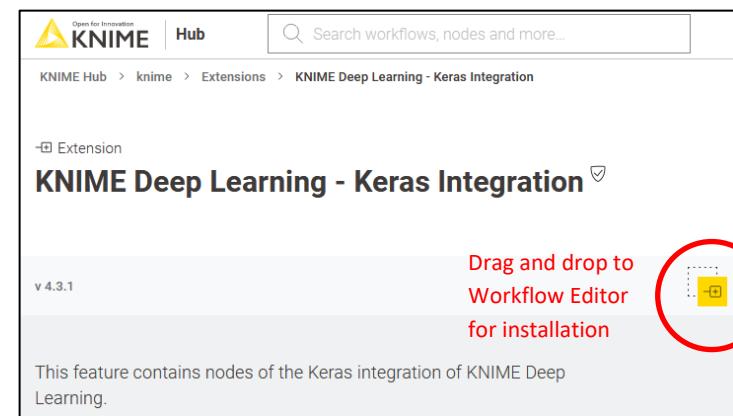
1. Desde el Top Menu, seleccione “File” → “Install KNIME Extensions”, seleccione la extensión deseada y haga click en el botón “Next”. Siga las instrucciones del asistente
2. Desde el Top Menu, seleccione “Help” → “Install New Software”. En la ventana “Available Software”, en el cuadro de texto “Work with”, seleccione el URL de actualización de KNIME (usualmente llamado “KNIME Analytics Platform 4.x Update Site” - <http://update.knime.com/analytics-platform/4.x>). Entonces seleccione la extensión, haga click the “Next” button y siga las instrucciones del asistente
3. Busque en el KNIME Hub, en un navegador web o desde el panel KNIME Hub. Cuando encuentre la extensión deseada, arrastre y suelte el ícono de la extensión desde el navegador al Workflow Editor.

Una vez que se hayan instalado las extensiones KNIME seleccionadas y se haya reiniciado KNIME, debería ver la nueva categoría, correspondiente a la extensión instalada, en el “Node Repository”.

1.24. La ventana “Available Software”



1.25. Una extensión del KNIME Hub



En la ventana “Software disponible” puede encontrar algunos grupos de extensiones: KNIME & Extensions, KNIME Labs Extensions, KNIME Node Development Tools, Sources y más. “KNIME & Extensions” contiene todas las extensiones proporcionadas para la versión actual; “KNIME Labs Extensions” contiene una serie de extensiones listas para usar, pero que aún no tienen la calidad de lanzamiento x.1; “KNIME Node Development Tools” contiene paquetes con algunas herramientas útiles para que los programadores de Java desarrollen nodos; “Sources” contiene el código fuente de KNIME. Los paquetes específicos donados por terceros o entidades comunitarias también pueden estar disponibles en la lista de extensiones. Por lo general, se agrupan en categorías de “Community”. Mi consejo es instalar todas las

extensiones, incluso las de cheminformatics. Muchos de ellos contienen varios nodos útiles que no están necesariamente restringidos a un dominio en particular.

## 1.9. Datos y workflows para éste libro

Este libro construye algunos ejemplos y proporciona las soluciones a los ejercicios. Están contenidos en carpeta "BeginnersLuck" descargable desde el [espacio KNIME Hub de uno de los autores](#) de este libro. Para acceder el KNIME Hub, necesitas crear una cuenta con el [KNIME Forum](#). Después de ingresar al KNIME Hub, para descargar los flujos de trabajo, simplemente haga clic en el ícono de la nube.

- Descargue la carpeta completa en su máquina, lo que resultará en un archivo .knar. Luego:
- Haga doble click o importelo al KNIME Explorer haciendo: Select File → Import KNIME Workflow ...

1.26. Workflows y datos para este libro en el KNIME Hub ([hub.knime.com/hayasaka/spaces/Public/latest/KNIMEPress/KNIME\\_Beginners\\_Luck\\_4.4\\_20210802](https://hub.knime.com/hayasaka/spaces/Public/latest/KNIMEPress/KNIME_Beginners_Luck_4.4_20210802))

The screenshot shows a web interface for the KNIME Hub. At the top left, there's a 'Public space' button. Below it, the word 'Public' is displayed in large, bold letters. On the right side, there are two user profile icons. In the center, there's a file listing for 'KNIME\_Beginners\_Luck\_4.4\_20210802'. The file details include 'Last edited: 24 Jun 2020' and a download icon with the number '2'. Below the file listing, there's a breadcrumb navigation: 'Home > KNIMEPress > KNIME\_Beginners\_Luck\_4.4\_20210802'. Underneath this, there's a folder structure with a 'BeginnersLuck' folder highlighted. A red circle is drawn around the download icon next to the folder name.

Al final de la operación de importación, en el panel KNIME Explorer debería encontrar una carpeta BeginnersLuck que contiene las subcarpetas Chapter2, Chapter3, Chapter4, Chapter5 y Chapter6, cada una con flujos de trabajo y ejercicios que se implementarán en los siguientes capítulos. También debe encontrar una carpeta KBLdata que contenga los datos requeridos

Los datos utilizados para los ejercicios y para los flujos de trabajo demostrativos de este libro fueron generados por el autor o descargados del Repositorio de Aprendizaje Automático de la UCI, un repositorio público de datos. (<http://archive.ics.uci.edu/ml/datasets>). Si el conjunto de datos pertenece al repositorio de la UCI, aquí se proporciona un enlace completo para descargar. Los datos generados por el autor, que no son datos públicos, se encuentran en la carpeta KBLData.

Datos del repositorio de aprendizaje automático de la UCI:

- Adult.data: <http://archive.ics.uci.edu/ml/datasets/Adult>
- Iris data: <http://archive.ics.uci.edu/ml/datasets/Iris>
- Yellow-small.data (Balloons) <http://archive.ics.uci.edu/ml/datasets/Balloons>
- Wine data: <http://archive.ics.uci.edu/ml/datasets/Wine>

## 1.10. Ejercicios

### Ejercicio 1

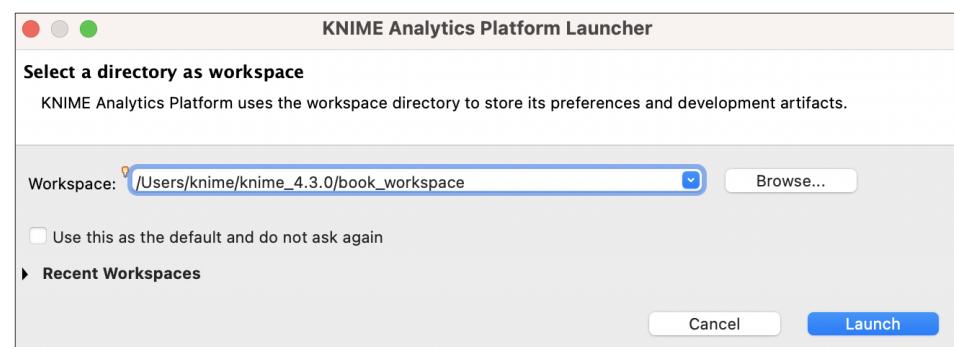
Cree su propio espacio de trabajo y asígnele el nombre "book\_workspace". Utilizará este espacio de trabajo para los próximos ejercicios y flujos de trabajo.

#### Solución al ejercicio 1

- Lanzar KNIME
- En la ventana del Workspace Launcher, haga clic en "Browse"
- Seleccione la ruta para su nuevo workspace
- Haga clic en "OK"

Para mantener esto como su workspace predeterminado, habilite la opción en la esquina inferior izquierda.

#### 1.27. Ejercicio 1: Crear el workspace "book\_workspace"



# Ejercicio 2

Instale las siguientes extensiones:

- KNIME Database
- KNIME Javascript Views
- KNIME Report Designer

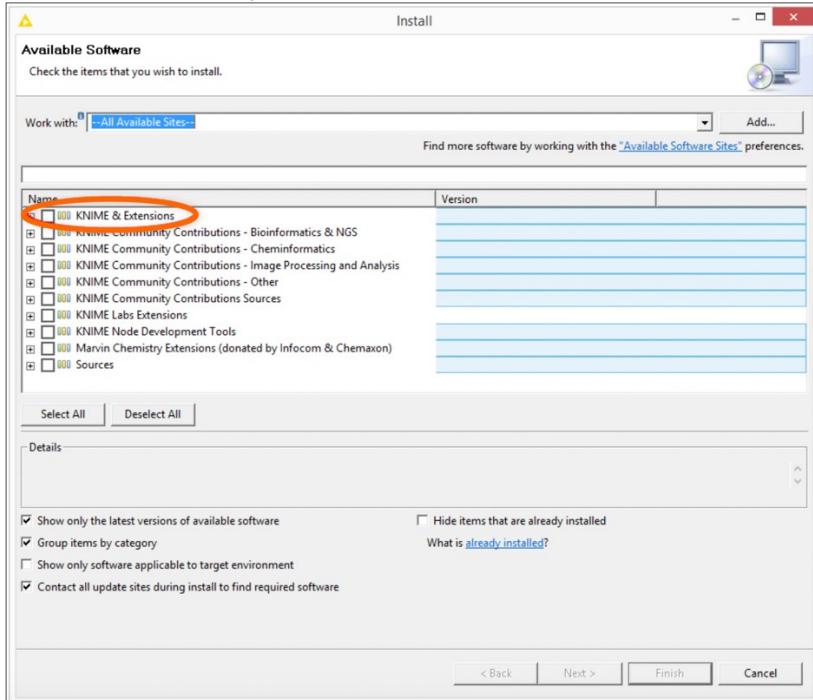
## Solucion al Ejercicio 2

Del Top Menu, seleccione “File” → “Install KNIME Extensions”

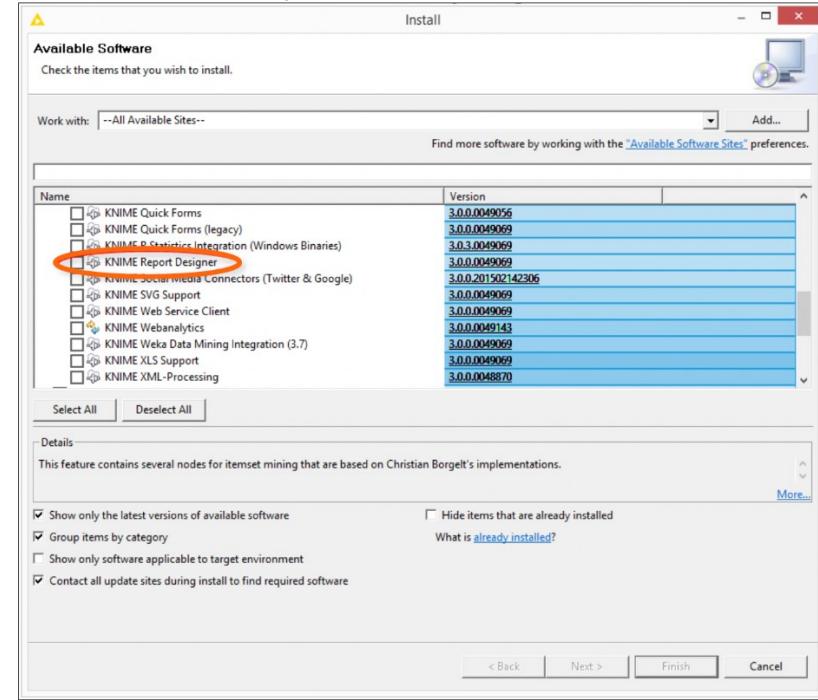
Seleccione las extensiones requeridas (KNIME Database,...)

Haga Click en “Next” y siga instrucciones

1.28. Ejercicio 2: Lista de KNIME Extensions



1.29. Ejercicio 2: Reportando Extension



## Ejercicio 3

Busque todos los “Row Filter” nodes en el Node Repository.

En el panel “Node Description”, ¿puede ud explicar la diferencia entre un “Row Filter”, a “Reference Row Filter”, and a “Nominal Value Row Filter”?

Muestre el efecto de esos nodos en las siguientes tablas

Tabla Original

Position	name	team
1	The Black Rose	4
2	Cynthia	4
3	Tinkerbell	4
4	Mother	4
5	Augusta	3
6	The Seven Seas	3

Tabla de Referencia

Ranking	scores
1	22
3	14
4	10

### Solucion al Ejercicio 3

Row Filter

Este nodo permite el filtrado de filas de acuerdo con ciertos criterios. Puede incluir o excluir: ciertos rangos (por número de fila), filas con un determinado ID de fila y filas con un cierto valor en una columna seleccionable (atributo). En el siguiente ejemplo, usamos el siguiente criterio de filtro: team > 3

*Original*

<b>Position</b>	<b>name</b>	<b>team</b>
1	The Black Rose	4
2	Cynthia	4
3	Tinkerbell	4
4	Mother	4
5	Augusta	3
6	The Seven Seas	3

*Filtrada*

<b>Position</b>	<b>name</b>	<b>team</b>
1	The Black Rose	4
2	Cynthia	4
3	Tinkerbell	4
4	Mother	4

*Reference Row Filter*

Este nodo tiene dos tablas de entrada. La primera tabla de entrada, conectada al puerto inferior, se toma como tabla de referencia; la segunda tabla de entrada, conectada al puerto superior, es la tabla a filtrar. Tienes que elegir la columna de referencia en la tabla de referencia y la columna de filtrado en la segunda tabla. Todas las filas con un valor en la columna de filtrado que también existe en la columna de referencia se mantienen, si se selecciona la opción "incluir"; se eliminan si se selecciona la opción "excluir"

*Tabla de Referencia*

<b>Ranking</b>	<b>scores</b>
1	22
3	14
4	10

*Tabla Filtrada*

<b>Position</b>	<b>name</b>	<b>team</b>
1	The Black Rose	4
2	Cynthia	4
3	Tinkerbell	4
4	Mother	4
5	Augusta	3
6	The Seven Seas	3

*Tabla resultante*

Position	name	team
1	The Black Rose	4
3	Tinkerbell	4
4	Mother	4

En el ejemplo de arriba, estamos usando el "Ranking" como la columna de referencia en la tabla de referencia y la "Position" como la columna de filtrado en la tabla filtrada. Hemos escondido de incluir las filas comunes

#### *Nominal Value Row Filter*

Filtrar las filas según el valor seleccionado de un atributo nominal. Se puede seleccionar una columna nominal y uno o más valores nominales de este atributo como criterio de filtro. Las filas que tienen estos valores nominales en la columna seleccionada se incluyen en los datos de salida. Básicamente es un filtro de fila aplicado a una columna con valores nominales. Las columnas nominales son columnas de cadena y los valores nominales son los valores que contienen.

En el siguiente ejemplo, usamos "name" como columna nominal y "name = Cynthia" como criterio de filtrado.

*Original table*

Position	name	team
1	The Black Rose	4
2	Cynthia	4
3	Tinkerbell	4
4	Mother	4
5	Augusta	3
6	The Seven Seas	3

*Filtered table*

Position	name	team
2	Cynthia	4

# Capítulo 2. Mi primer workflow

## 2.1. Operaciones en el Workflow

Si ha iniciado KNIME por primera vez, su panel "KNIME Explorer" en la esquina superior izquierda del KNIME workbench (banco de trabajo) contiene solo un grupo de flujo de trabajo (carpeta) llamado "**Example Workflows**". Esta carpeta " Example Workflows " contiene varias subcarpetas, cada una con flujos de trabajo básicos para casos de uso muy comunes:

- **Basic Examples.** Los flujos de trabajo de la subcarpeta "Ejemplos básicos" muestran operaciones generales básicas, como importar datos, combinar datos, ETL, entrenar y evaluar un modelo y, finalmente, mostrar los resultados en un informe simple.
- **Customer Intelligence.** Los flujos de trabajo básicos para la predicción de abandono, la calificación crediticia y la segmentación de clientes están disponibles dentro de la subcarpeta "*Customer Intelligence*".
- **Retail.** Un motor de recomendaciones está integrado en la subcarpeta "*Retail*".
- **Social Media.** Un ejemplo de análisis de redes sociales está disponible en "*Social Media*".

Estos ejemplos de flujos de trabajo (workflows) se pueden reutilizar y readaptar para su propia aplicación. Sin embargo, en este capítulo queremos construir nuestro propio primer flujo de trabajo, para realizar las siguientes operaciones básicas:

- Leer datos de un archivo de texto
- Filtrar filas no deseadas
- Filtrar columnas no deseadas
- Escribir los datos resultantes en un archivo CSV

Usaremos este primer flujo de trabajo para explorar estructuras de datos y tipos de datos, comandos de flujo de trabajo y nodos, posibilidades de depuración e inspección de datos, opciones de comentarios, ventanas de configuración y comandos de ejecución, y otras características disponibles dentro del banco de trabajo (workbench) KNIME.

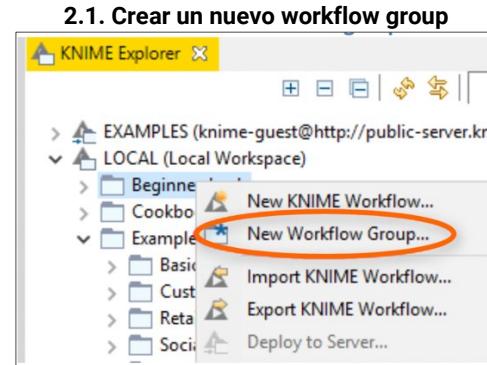
Para mantener limpio nuestro espacio, utilizamos grupos de workflows para organizar los workflows por capítulo o tema. Vamos a crear ahora un nuevo grupo de flujo de trabajo y llamémoslo "Chapter 2". Una vez hecho esto, debemos completar el grupo de flujo de trabajo recién creado con un nuevo flujo de trabajo, llamémoslo " My First Workflow ". Finalmente, en el panel "KNIME Explorer", debería ver el grupo de flujo de trabajo "Chapter 2" con un flujo de trabajo llamado " My First Workflow " en él. Por ahora, " My First Workflow " es un flujo de trabajo vacío. De hecho, si hace doble clic en él, el editor de flujo de trabajo se abre en una página vacía.

Veamos ahora cómo realizar algunas operaciones de flujo de trabajo, incluida la creación, el guardado y la eliminación de un flujo de trabajo.

## Crear un Nuevo Workflow Group

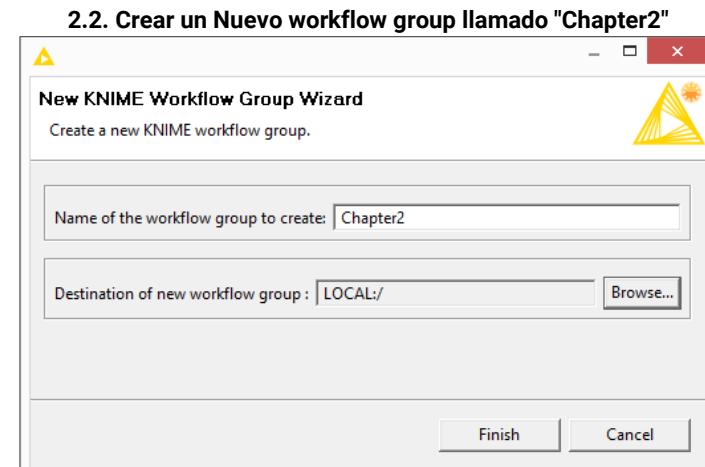
En el panel “KNIME Explorer” :

- Haga clic con el botón derecho en cualquier lugar del espacio de trabajo LOCAL (o en un espacio del servidor)
- Seleccione “New Workflow Group”



En el asistente “New KNIME Workflow Group Wizard” :

- Ingrese el nombre del grupo de flujo de trabajo
- Ingrese el destino dentro del panel KNIME Explorer.  
Para visualizar los posibles destinos del grupo de flujo de trabajo, haga clic en el botón “Browse”
- Haga clic en “Finish”



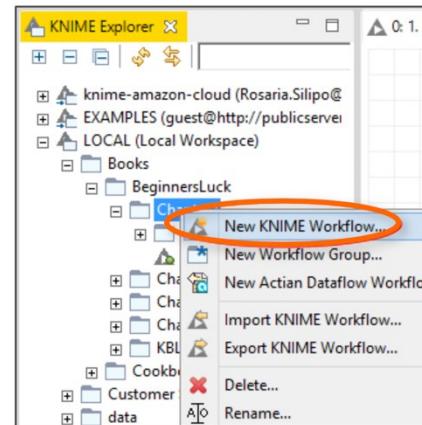
**Nota.** Si selecciona un grupo de flujo de trabajo existente en KNIME Explorer, hace clic con el botón derecho e inicia el “New KNIME Workflow Group Wizard”, el destino predeterminado será el grupo de flujo de trabajo seleccionado.

# Crear un nuevo workflow

In the “KNIME Explorer” panel:

- Haga clic con el botón derecho en cualquier lugar del espacio de trabajo LOCAL (o en un espacio del servidor)
- Seleccione “New KNIME Workflow”

## 2.3. Crear un nuevo workflow



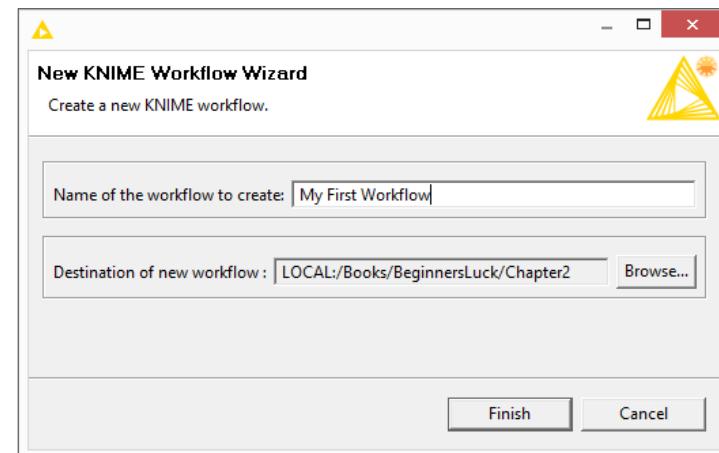
En el “New KNIME Workflow Wizard” :

- Ingrese el nombre del nuevo flujo de trabajo
- Especifique dónde debe ubicarse, por ejemplo, en un grupo de flujo de trabajo existente, utilizando el botón “Browse” si es necesario
- Haga clic en “Finish”

**Nota.** Si selecciona un grupo de flujo de trabajo existente en KNIME Explorer, haga clic con el botón derecho e inicie el Asistente “New KNIME Workflow Wizard”, el destino predeterminado estará en el workflow group seleccionado.

**Para abrir un flujo de trabajo, simplemente haga doble clic en el flujo de trabajo en KNIME Explorer.**

## 2.4. Crear un Nuevo workflow llamado "My First Workflow" en la carpeta "Chapter2"



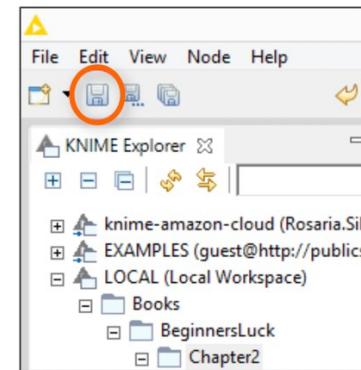
# Guardar un workflow

Para guardar un flujo de trabajo, haga clic en el icono de disco en el Top Menu. Esto solo guarda el workflow seleccionado en el editor de flujo de trabajo (workflow editor). Al guardar el flujo de trabajo, se guarda la arquitectura del flujo de trabajo, la configuración de los nodos y los datos producidos en la salida de cada nodo.

Si desea guardar una copia del flujo de trabajo seleccionado actualmente, debe hacer clic en el icono de disco "Save as ..." a la derecha del icono de disco único "Save".

Si desea guardar TODOS los flujos de trabajo abiertos y no solo el abierto, debe hacer clic en el icono de la pila de discos "Save all" a la derecha del icono de disco "Save as..."

2.5. Opciones "Save", "Save as...", and "Save all"



# Borrar un workflow

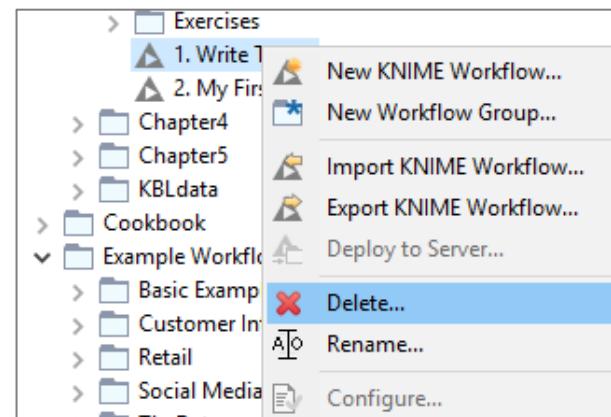
Para borrar un workflow

- Click-derecho en el workflow en el panel "KNIME Explorer"
- Seleccione "Delete"

En la ventana "Confirm Deletion", se le preguntará si realmente quiere eliminar ese workflow.

**Cuidado!** El comando "Delete" remueve el proyecto de workflow físicamente del disco duro. **Una vez que sea borrado no puede ser recuperado!!!!**

2.6. Eliminar un workflow



## 2.2. Operaciones con Nodos (Node operations)

En el Capítulo 1, hemos visto que un nodo es la unidad computacional básica en un flujo de trabajo KNIME. También hemos visto que los nodos están disponibles, organizados por categorías, en el panel "Repositorio de nodos" en la esquina inferior izquierda del banco de trabajo KNIME. Y hemos visto que cada nodo tiene cuatro estados: aún no configurado (rojo), configurado (amarillo), ejecutado exitosamente (verde) y ejecutado con error (rojo con cruz).

En esta sección vamos a explorar: cómo agregar un nuevo nodo a un flujo de trabajo (estado final = inactivo, no configurado; **luz roja**), cómo configurar el nodo (estado final = configurado, no ejecutado; **luz amarilla**) y cómo ejecutar el nodo (estado final = ejecutado con éxito; **luz verde**).

### Crear un nuevo nodo

Para crear un nuevo nodo, tiene dos opciones:

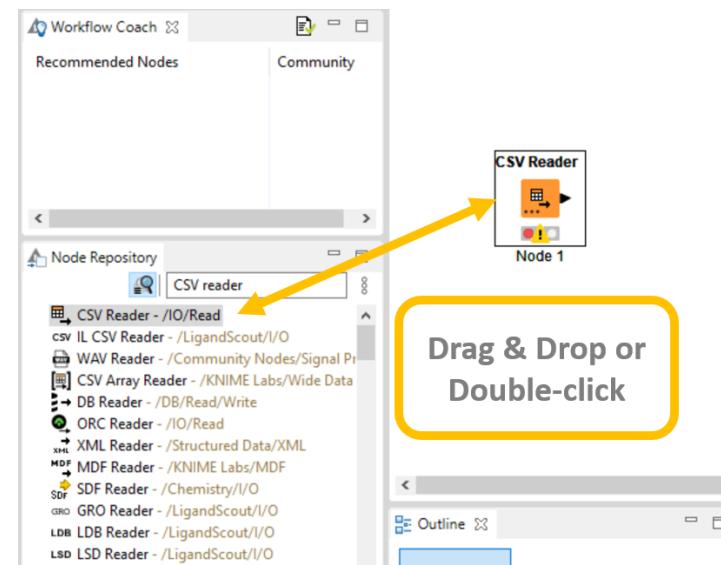
- arrastre y suelte el nodo del panel "Repositorio de nodos" en el editor de flujo de trabajo
- haga doble clic en el nodo en el panel "Repositorio de nodos"

El nodo generalmente se importa con el estado de semáforo en rojo.

Para conectar un nodo con nodos existentes, hay dos opciones:

- haga clic en el puerto de salida del primer nodo y suelte el mouse en el puerto de entrada del segundo nodo
- seleccione un nodo en el flujo de trabajo y haga doble clic en un nodo en el repositorio de nodos: esto crea un nuevo nodo y conecta automáticamente su primer puerto de entrada al primer puerto de salida del nodo existente. Shift + doble clic en el nuevo nodo mueve la conexión al siguiente puerto de entrada.

2.7. Arrastre y libere o haga double-click en el nodo para crear un Nuevo nodo en el editor de workflow



Una vez creado el nodo, debemos configurarlo, es decir, establecer los parámetros necesarios para que ejecute la tarea del nodo.

A continuación, abramos la ventana de configuración del nodo y configuremos el nodo.

Finalmente, necesitamos asociar una descripción a este nodo con fines de documentación, para reconocer fácilmente qué tarea está realizando dentro del flujo de trabajo. Cada nodo se crea con un texto predeterminado debajo como "Node n", donde "n" es un número progresivo. Este texto del nodo se puede personalizar. Esto, junto con las anotaciones del flujo de trabajo descritas en el capítulo 1, mantiene clara la descripción general del flujo de trabajo y cumple el propósito de la documentación del flujo de trabajo.

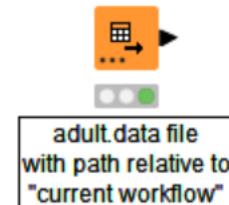
2.8. Doble-click en el nombre del nodo para editarlo

## Texto del Nodo

Para cambiar el texto ubicado debajo del nodo:

- Haga doble clic en el texto del nodo para que se pueda editar
- Escribe el nuevo texto. El texto puede ocupar más líneas, separadas por "Enter"
- Haga clic fuera del nodo para confirmar el cambio de texto

CSV Reader



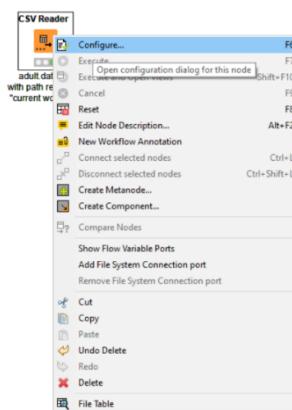
2.9. Haga click-derecho en el nodo y seleccione "Configure" ó Haga Doble-click en el nodo

## Configurar un nodo

Para configurar un nodo existente:

- Haga Double-click en el nodo  
ó
- Haga click-derecho en el nodo y seleccione "Configure"

Si todos los puertos de entrada están conectados, aparece el cuadro de diálogo de configuración para que complete los ajustes de configuración. Cada nodo tiene un diálogo de configuración diferente, ya que cada nodo realiza una tarea diferente.



Después de una configuración exitosa, el nodo cambia su semáforo a amarillo.

## Ejecutar un nodo

El nodo ahora está configurado, lo que significa que sabe qué hacer.

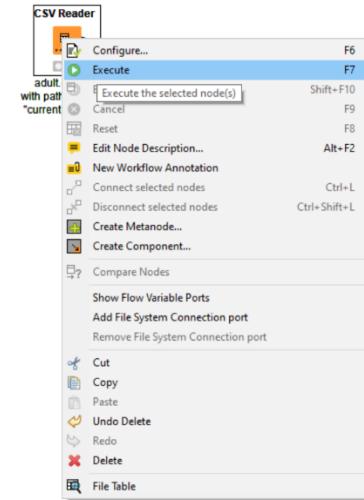
Para que realmente realice su tarea, debemos ejecutarlo.

Para ejecutar un nodo y dejar que ejecute su tarea:

- Haga click-botón derecho en el nodo y seleccione "Execute"
- ó
- Seleccione el nodo y haga click en la flecha verde de la barra de tareas

Si la ejecución ha sido exitosa el semáforo del nodo cambia de Amarillo a verde

**2.10. Hag click-botón derecho y seleccione "execute" para ejecutar el nodo**

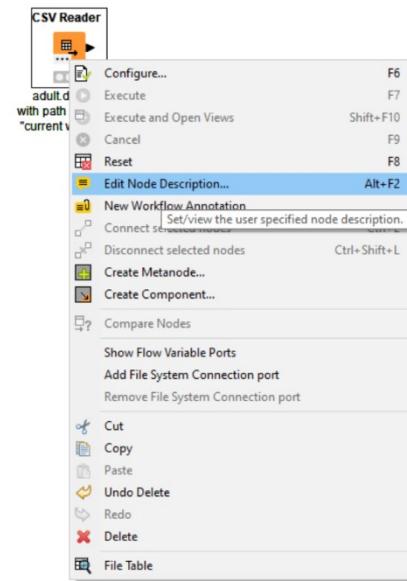


# Descripcion del Nodo

Además del texto del Nodo, Ud puede insertar una descripción oculta de la tarea que hace el nodo:

- Haga clic derecho en el nodo
- Seleccione la opción "Edit Node Description ..."
- En la ventana "Descripción de nodo"
  - En el campo "Custom Description ", escriba la descripción del nodo
  - Haga clic en "OK"

2.11. Opcion "Edit Node Description" para insertar una descripción oculta



## Ver los Datos Procesados por el Nodo

Si la ejecución fue exitosa (**luz verde**), puede inspeccionar los datos procesados.

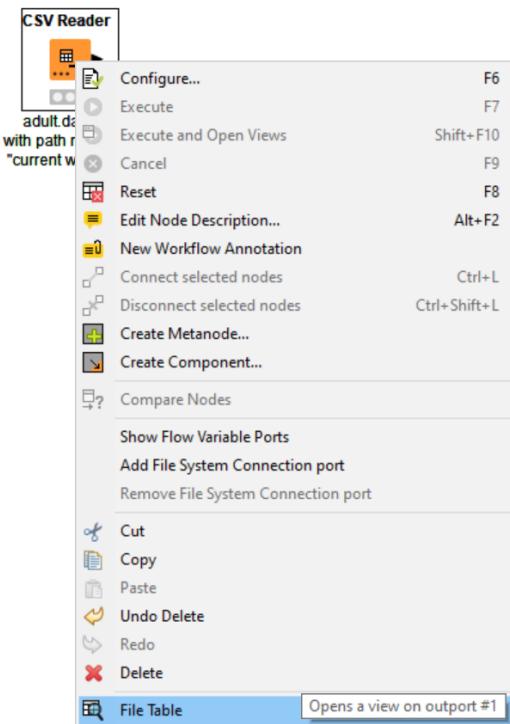
- Haga clic derecho en el nodo
- Seleccione la última opción en el menú contextual
- A continuación, debería aparecer la tabla de datos con los datos procesados.

La opción para ver los datos procesados es el último elemento del menú contextual (menú contextual) para todos los nodos con datos de salida, pero toma diferentes nombres para diferentes nodos dependiendo de su tarea.

Algunos nodos pueden producir más de un conjunto de datos de salida. En este caso, hay más de un elemento de vista en la parte inferior del menú contextual del nodo.

Los nodos presentan tantos puertos de salida (en este caso, un triángulo negro) como se produzcan conjuntos de datos de salida.

2.12. Haga clic derecho en el nodo y seleccione la última opción en el menú contextual para visualizar los datos procesados



## 2.3. Leer datos desde un archivo

El primer paso en todos los proyectos de análisis de datos consiste en leer datos. Los datos locales generalmente se leen de un archivo o de una base de datos. En este capítulo describimos cómo leer y escribir datos desde y hacia un archivo de texto. La lectura y escritura de datos desde y hacia una base de datos se describe en el Capítulo 3 en la sección "Operaciones de la base de datos".

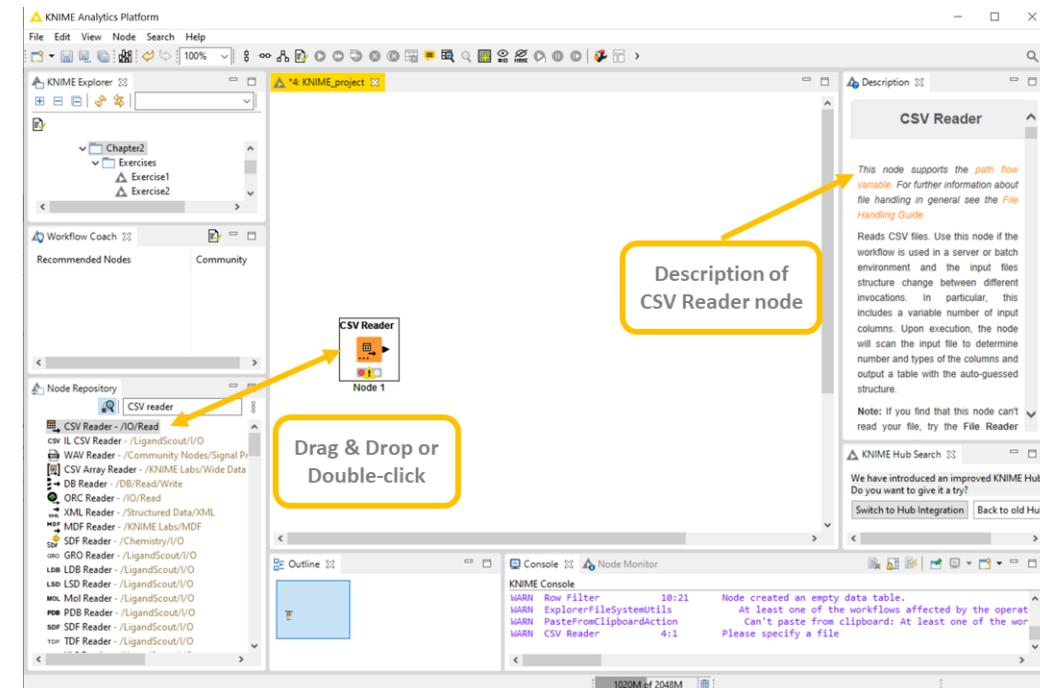
El formato de archivo más común es el formato de valores separados por comas (CSV), que cubre todos los archivos de texto donde los campos están separados por un carácter especial. Para leer este tipo de archivos de datos, el nodo CSV Reader es el nodo más versátil de KNIME Analytics Platform.

## Crear un nodo lector de CSV “CSV Reader”

En el panel “Node Repository” esquina inferior izquierda:

- Expanda la categoría "IO" y luego la subcategoría "Read" ó, alternativamente, escriba "CSV Reader" en el cuadro de búsqueda
- Arrastre y suelte el nodo "CSV Reader" en el editor de flujo de trabajo (o haga doble clic en él)
- Si el panel "Description" de la derecha está habilitado, muestra todo lo que necesita saber sobre el nodo "CSV Reader": tarea, puerto de salida y configuraciones requeridas.
- Para activar el panel "Description", vaya al Menú superior, abra "View" y seleccione "Description".

2.13. Crear un nodo lector de CSV



**Nota.** Debajo del nodo "CSV Reader" recién creado, es posible que observe un pequeño triángulo de advertencia amarillo. Si pasa el mouse sobre él, aparecerá la siguiente información sobre herramientas: "No hay configuraciones disponibles". Esto se debe a que el nodo CSV Reader aún no se ha configurado (¡necesita al menos la ruta del archivo!). Por el momento, el nodo se encuentra en estado de semáforo en rojo: ni siquiera configurado.

# Configure the “CSV Reader node

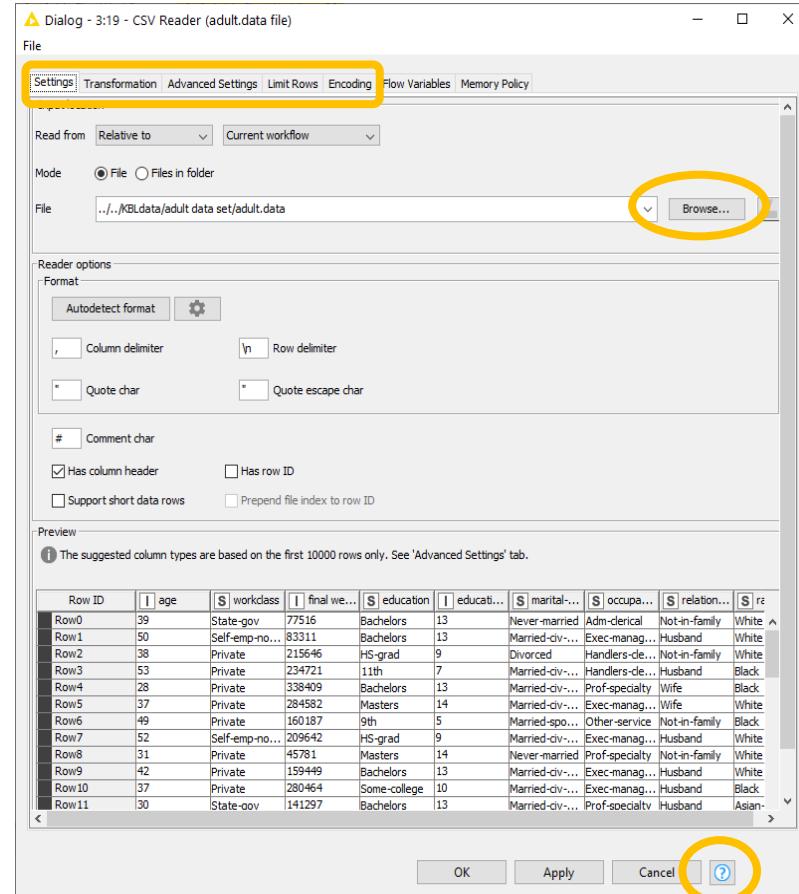
- Haga doble clic en el nodo
- ó
- Haga clic con el botón derecho en el nodo y seleccione "Configure"

En el diálogo de configuración:

- Especifique la ruta del archivo, escribiendo o usando el botón "Examine. Para este ejemplo, usamos el archivo adult.csv, descargable del Repositorio de Aprendizaje Automático de UCI (<http://archive.ics.uci.edu/ml/datasets/Adult>) o disponible en KBLdata / adult data set / adult.csv .
- En la mayoría de los casos, el nodo "CSV Reader" detecta automáticamente la estructura del archivo.
- Si este no es uno de la mayoría de los casos y el nodo del lector CSV no ha detectado la estructura del archivo exactamente, entonces presione el botón "Autodetect format" y / o habilite / deshabilite todas las casillas de verificación requeridas en las pestañas en la parte superior, de acuerdo con la estructura de datos de archivo.
- Una vista previa de los datos leídos está disponible en la parte inferior de la ventana de configuración y muestra posibles errores de lectura.

A la derecha de los botones OK/Cancel en la parte inferior de la ventana, hay un pequeño botón que lleva un icono de signo de interrogación. Este es el botón de ayuda y conduce a una nueva ventana que contiene la descripción del nodo.

2.14. Ventana de configuración del nodo "CSV Reader"



# Personalizar las propiedades de las columnas

Es posible personalizar la forma en que se leen los datos de cada columna.

Por ejemplo, el archivo adult.csv contiene un campo llamado "**fnlwgt**". Podemos cambiar el encabezado de esta columna a un "**final weight**" más significativo en la pestaña **Transformation** (Fig. 2.15), simplemente escribiendo el nuevo nombre en el campo Nuevo Nombre correspondiente.

En la pestaña **Settings tab** (Fig. 2.14), también podemos establecer: el carácter delimitador que separa la columna, si usar la primera fila de datos para los encabezados de columna, si usar la primera columna como RowID, y si podemos tolerar más filas de datos.

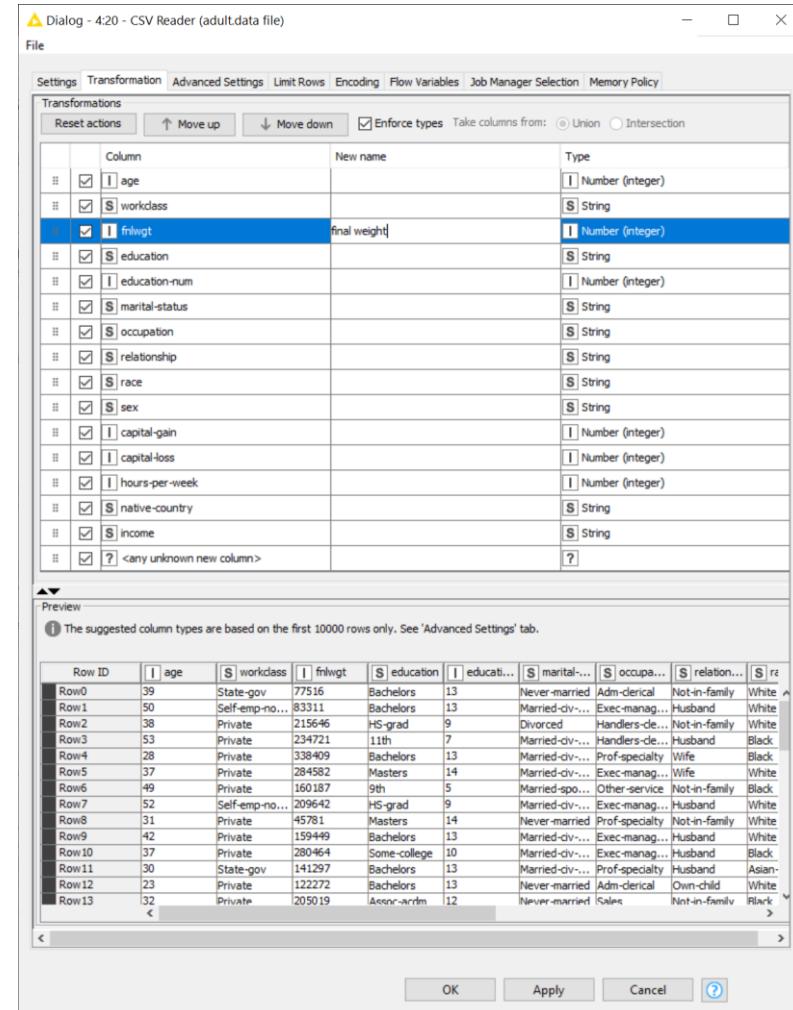
**Tab Limit Rows** permite omitir las primeras líneas en el archivo de texto. Esto suele ser muy útil cuando se trata de archivos con un texto de encabezado.

**Tab Advanced Settings** ofrece algunas configuraciones adicionales y **tab Encoding** permite establecer la codificación correcta para el texto.

Observe los tres puntos en la esquina inferior izquierda del ícono de CSV Reader. Estos puntos indican puerto dinámico. Al hacer clic en los tres puntos, puede agregar uno o más puertos de entrada a este nodo para conectarse a un sistema de archivos externo, como Amazon S3, HDFS, Databricks, Microsoft Azure, etc.

Los tres puntos en un nodo siempre significan puertos dinámicos.

2.15. Personaliza cómo leer columnas



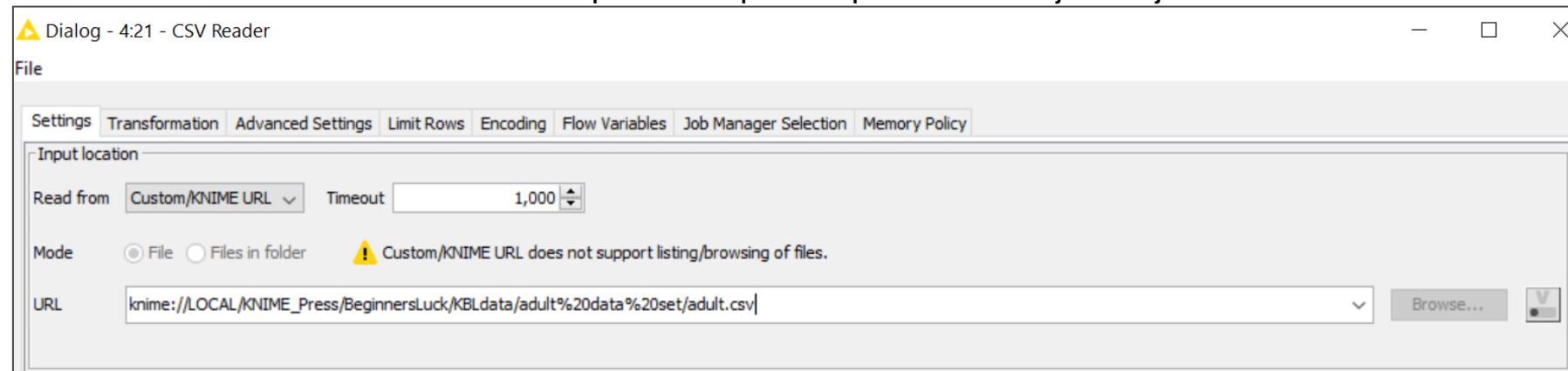
**Nota.** Despues de configurar el nodo “CSV Reader”, su estado cambia de amarillo a verde .

¡Lo que hemos descrito aquí es el camino más largo para configurar un nodo de lector CSV! Dado que esta (creación y configuración de nodos) es una forma que funciona para todos los nodos, no pudimos evitar pasar por ella. Sin embargo, si el archivo tiene una extensión conocida, como .csv o .txt, es posible que exista una forma más rápida.

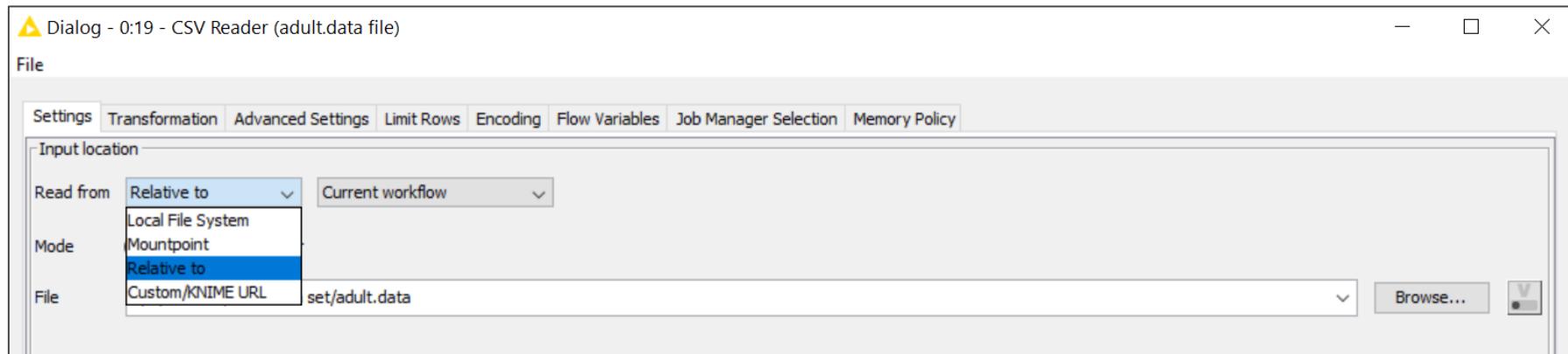
**Nota.** En lugar de escribir manualmente la ruta completa del archivo en el campo Ubicación de entrada / Archivo en la ventana de configuración de CSV Reader, podríamos simplemente arrastrar y soltar el archivo desde el panel KNIME Explorer en el editor de flujo de trabajo. Si el archivo tiene una extensión conocida (como .csv, por ejemplo), esto crea automáticamente el nodo lector apropiado y lo configura.

Tenga en cuenta que cuando cree un lector CSV de esta manera, obtendrá un protocolo diferente en el cuadro de URL. Al buscar un archivo, obtiene un protocolo: *file:// ...*; arrastrando y soltando se obtiene un protocolo *knime://* (Fig. 2.16). Este protocolo especial está emparejado con la opción “Custom/KNIME URL” en el Read From field.

**2.16. Configuración automática de la parte Ubicación de entrada en la pestaña Configuración de un nodo Lector CSV creado mediante la función de arrastrar y soltar un archivo desde el panel KNIME Explorer en el panel del editor de flujo de trabajo.**



## 2.17. Opciones disponibles en el campo "Read from" en la pestaña "settings" de un nodo CSV Reader



Otras opciones disponibles para ubicar un archivo en la parte Ubicación de entrada de la ventana de configuración de un nodo CSV Reader son: desde el Sistema de archivos local, desde un punto de montaje específico, desde una ubicación relativa y la URL personalizada / KNIME que ya hemos visto. Las primeras tres opciones definen el punto de partida de la ruta del archivo. El botón Browser puede ayudarlo a encontrar el archivo, mientras que la opción "Read from" lo escribirá en el formato correcto.

Observe que en los nodos Reader y Writer que aún no tienen el campo "read from", el protocolo knime: // sigue funcionando. Echemos un vistazo a las diferentes opciones de protocolo de knime: //.

## El protocolo *knime://...*

El protocolo *knime://...* es un protocolo especial que le permite hacer referencia al espacio de trabajo local o al flujo de trabajo local en una ruta. Esto permite la creación de rutas relativas que lo hacen independiente de la URL absoluta de la carpeta del espacio de trabajo, pero solo depende de la estructura de la carpeta del espacio de trabajo.

Esta función es particularmente útil cuando se trasladan flujos de trabajo a otros espacios de trabajo o incluso a otras máquinas. Siempre que se conserve la estructura de carpetas de datos y el flujo de trabajo, CSV Reader seguirá buscando el archivo y leyéndolo.

El protocolo *knime://* funciona en todos los nodos de lectura y escritura.

### 2.18. Posibles opciones del protocolo *knime://*

<i>knime://LOCAL/</i>	refers to the current workspace location
<i>knime://LOCAL/.../knime-workspace</i>	moves two levels up from the current workspace location to a new workspace folder named knime.workspace
<i>knime://knime.workflow/</i>	refers to the current workflow location
<i>knime://knime.workflow/.../data</i>	moves two levels up from the current workflow location to a new folder named data
<i>knime://&lt;knime-mountID&gt;/</i>	refers to a KNIME Server available in the KNIME Explorer panel
<i>knime://&lt;knime-mountID&gt;/&lt;path&gt;/data</i>	moves to the <path>/data folder on the referenced KNIME Server

También necesitamos asignar a este nodo un comentario significativo para que podamos reconocer fácilmente cuál es su tarea en el flujo de trabajo. El comentario predeterminado en nuestro "Lector CSV" es "Nodo 1" porque fue el primer nodo que creamos en el flujo de trabajo. Para cambiar el comentario del nodo:

- Haga doble clic en la etiqueta "Nodo 1" debajo del nodo "CSV Reader"
- Ingrese el nuevo comentario del nodo (por ejemplo "Adult data set")
- Haga clic en otro lugar

Ahora hemos cambiado el comentario en el nodo "Lector CSV" de "Nodo 1" a "Adult data set".

**Nota.** Observe los tres puntos en la esquina inferior izquierda del nodo. Hacer clic en ellos le permite conectarse a un sistema de archivos externo y acceder a los archivos desde allí.

Después de la configuración, para que el nodo realmente lea el archivo, necesitamos ejecutarlo. Por lo tanto, proceda de la siguiente manera:

- Haga clic derecho en el nodo
- Seleccione "Ejecutar"

**Nota:** Si el proceso de lectura no tiene errores, el nodo cambia su semáforo a verde.

**Nota:** En cada ventana de configuración encontrará una pestaña, llamada "Flow Variables". Estas se utilizan para pasar parámetros externos de un nodo a otro. Sin embargo, no vamos a trabajar con "Flow Variables". en este libro, ya que pertenecen a un curso más avanzado sobre funcionalidades KNIME.

La categoría "IO" → "Read" en "Node Repository" contiene varios nodos adicionales para leer archivos en diferentes formatos, como Excel, CSV, formato propietario de KNIME y más. La categoría IO"/"File Handling" tiene nodos para leer formatos especiales y archivos especiales, como por ejemplo archivos ZIP, archivos remotos, etc.

## 2.4. Estructura de datos y tipos de datos en KNIME

Si la ejecución del nodo fue exitosa, ahora puede ver los datos resultantes.

- Haga clic con el botón derecho en el nodo "CSV Reader"
- Seleccione la opción "File Table"

Aparece una tabla con los datos leídos. Echemos un vistazo a esta tabla para comprender cómo se estructuran los datos dentro de KNIME.

En primer lugar, los datos en KNIME se organizan como una **tabla**.

Cada fila está identificada por un **Row ID**. De forma predeterminada, los **Row ID** son cadenas donde "n" es un número progresivo. Pero los RowID se pueden forzar a ser cualquier cosa, con la única condición de que sean únicos. Los RowID no únicos producen un error

Las columnas se identifican mediante **column headers**. Si no hay **column headers** disponibles, se asignan **column headers** predeterminados como "Col n", donde "n" es un número progresivo. En el archivo adult.data se incluyeron **column headers**. Activamos la casilla de verificación ""Read column headers" en la ventana de configuración del nodo "CSV Reader" y ahora tenemos un encabezado para cada columna en la tabla de datos final. Los **column headers** deben ser únicos. Si un **column headers** aparece más de una vez, KNIME Analytics Platform agrega un sufijo "(#n)" (n = número progresivo) a cada aparición múltiple del **column header**.

Cada columna contiene datos con un tipo de dato. Los tipos de datos disponibles son:

- Double ("D")
- Integer ("I")
- String ("S")
- Date&Time (calendar + clock icon)
- Unknown ("?")
- Otros tipos específicos relacionados con el dominio (como Documento en la extensión de procesamiento de texto, Imagen en la extensión de procesamiento de imagen o Sonrisas en extensiones de química)

Date&Time puede provenir de la importación de datos de una base de datos. No aparece al leer datos de un archivo. En los archivos de texto, las fechas y horas se leen como cadenas. Luego, necesita un nodo "String to Date&Time" para convertir una cadena en una columna de tipo Date&Time.

El tipo de dato Unknown ("?") se refiere a columnas cuyo tipo de datos no se pudo determinar, como por ejemplo con tipos de datos mixtos o con todos los Missing values.

Missing values son celdas de datos con un estado especial de "Missing value" y se muestran de forma predeterminada con un signo de interrogación ("?"), a menos que el carácter de visualización para los valores perdidos se haya establecido de otra manera en la configuración del nodo "CSV Reader"

**Nota:** Los **Missing values** se representan de forma predeterminada con signos de interrogación. No son signos de interrogación, son datos que faltan y se representan con signos de interrogación. Los signos de interrogación en un archivo de texto se leen correctamente como signos de interrogación, pero no como Missing values. Los Missing values pueden representarse con cualquier otra cosa, según se defina en la ventana de configuración del nodo "CSV Reader".

## 2.18. Estructura de datos en KNIME

# Estructura de datos en KNIME

Los datos en Knime están organizados en una tabla con un número fijo de columnas. Cada columna se identifica con un **Row ID**

Las columnas se identifican con un **column headers**.

Cada columna tiene un solo un **data type**:

- Double ("D")
- Integer ("I")
- String ("S")
- Date&Time (calendar + clock icon)
- Unknown ("?")
- Otros

Row ID	age	workclass	fnwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss
Row0	39	State-gov	77516	Bachelors	12	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0
Row1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exe-manag	Husband	White	Male	0	0
Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0
Row3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0
Row4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-spec	Wife	Black	Female	0	0
Row5	37	Private	284582	Masters	14	Married-civ-spouse	Exe-manag	Wife	White	Female	0	0
Row6	49	Private	160187	9th	5	Married-civ-spouse	Other-serv	Not-in-family	Black	Female	0	0
Row7	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exe-manag	Husband	White	Male	0	0
Row8	31	Private	45781	Masters	14	Never-married	Prof-spec	Not-in-family	White	Female	14084	0
Row9	42	Private	159449	Bachelors	13	Married-civ-spouse	Exe-manag	Husband	White	Male	5178	0
Row10	37	Private	280464	Some-college	10	Married-civ-spouse	Exe-manag	Husband	Black	Male	0	0
Row11	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-spec	Husband	Asian-Pac-Islander	Male	0	0
Row12	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0
Row13	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0
Row14	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0
Row15	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Other	Male	0	0
Row16	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0
Row17	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0
Row18	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0
Row19	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exe-manag	Unmarried	White	Female	0	0
Row20	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-spec	Husband	White	Male	0	0
Row21	54	Private	302146	HS-grad	9	Separated	Other-serv	Unmarried	Black	Female	0	0
Row22	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0

Haciendo clic en el encabezado de una columna de datos se pueden ordenar en orden ascendente / descendente. Haciendo clic con el botón derecho en el encabezado de una columna de datos se pueden visualizar los datos utilizando renderizadores específicos.

Para datos Double ("D")/Integer ("I") por ejemplo, el renderizador "Barras" muestra los datos como barras con una longitud proporcional a su valor y en un mapa de calor rojo / verde.

Ud. puede ordenar temporalmente los datos haciendo clic en el **column headers** y seleccionar el tipo de clasificación. Puede cambiar temporalmente el renderizador de datos haciendo clic con el botón derecho en **column headers** y cambiar a un renderizador diferente, ya sea numérico o basado en barras. Ambas operaciones son solo temporales. Si cierra la tabla de datos y la vuelve a abrir, la ventana predeterminada se mostrará.

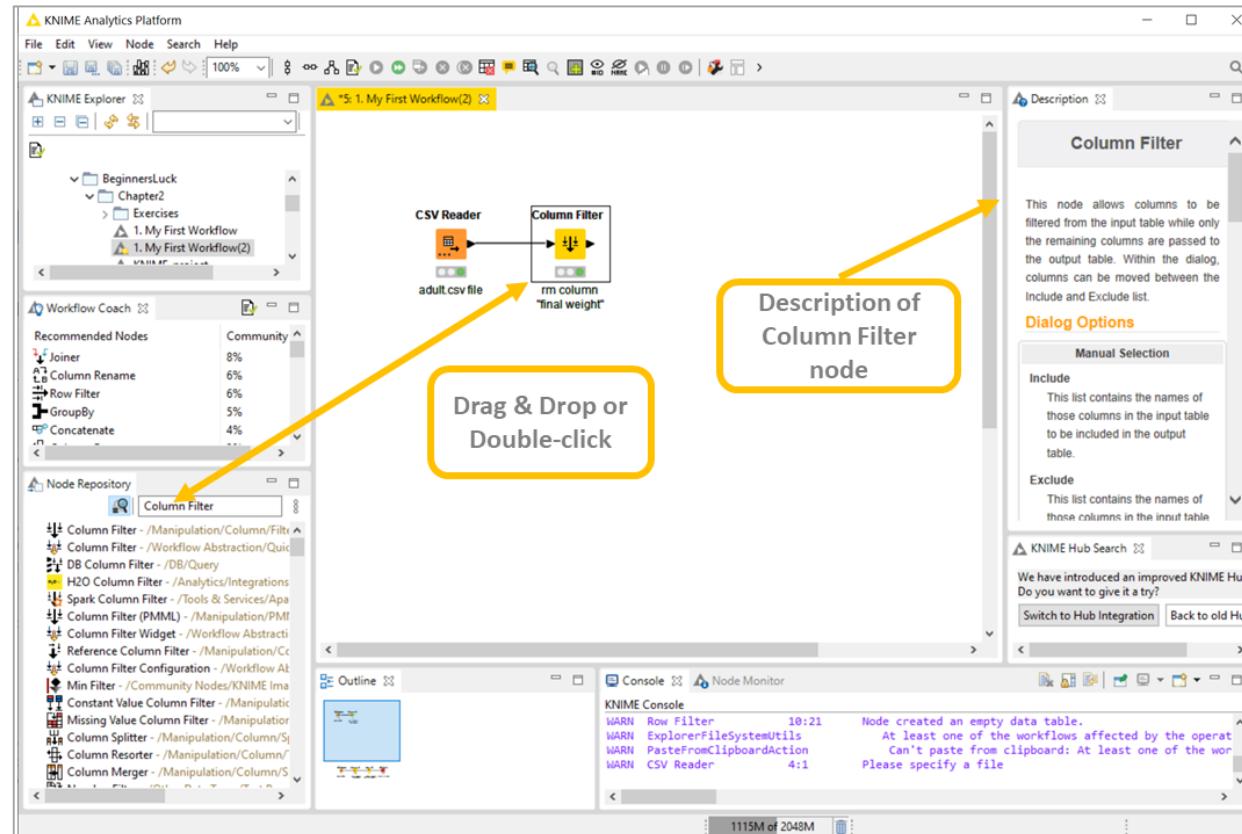
## 2.5. Filtrado de columna de Datos

En el próximo paso, se quiere filtrar la columna “final weight”. En el “Node Repository”, en la parte inferior izquierda, hay un conjunto de categorías denominado “Manipulation” con nodos dedicados a manipular la estructura de los datos. Esta categoría incluy operaciones para columnas, filas y la matriz completa de datos.

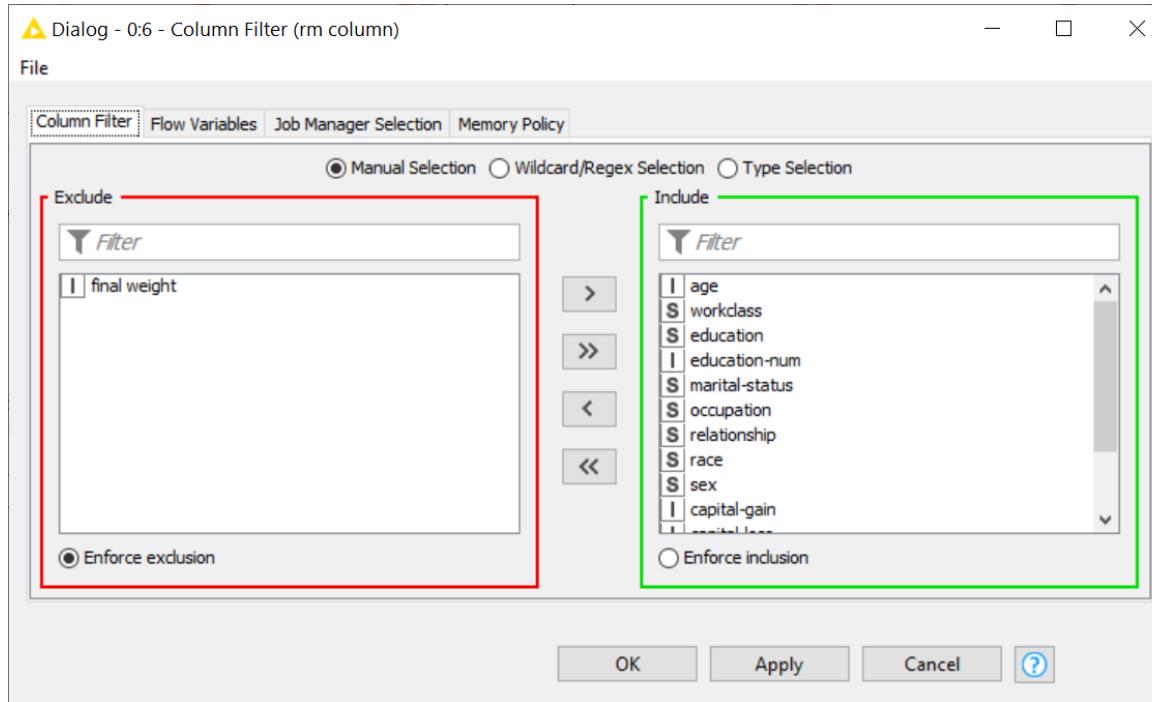
### Crear un “Column Filter” node

- En el panel “Node Repository”, busque el nodo “Column Filter” bajo la ccategoría “Manipulation” → “Column” → “Filter” ó busque “Column Filter” en el cuadro de búsqueda
- Arrastre y suelte el nodo “Column Filter” del “Node Repository”al workflow editor o haga doble clic en el “Node Repository”
- Conecte el nodo “Column Filter” con el nodo anterior (en nuestro workflow, el nodo “CSV Reader”) haciendo clic en la salida del nodo “CSV Reader” y soltando en la entrada del nodo “Column Filter” .

## 2.19. Creando un nodo "Column Filter"



## 2.20. Ventana de configuración de un nodo "Column Filter" : la opción "Manual Selection"



Para configurar el nodo:

- Haga doble clic en el nodo o haga clic con el botón derecho en el nodo y seleccione "Configure"
- Se abre la ventana de configuración. La ventana de configuración del nodo contiene todas las configuraciones para ese nodo en particular.
- Establecer los ajustes de configuración del nodo
- Haga clic en "OK"

# Configure un nodo “Column Filter”

El primer ajuste en la ventana de configuración es el tipo de filtrado. Puede seleccionar y retener columnas **manually**, **by type**, ó **by name**, de acuerdo con los botones de opción en la parte superior de la ventana de configuración (Fig. 2.21).

## Selección “Manual”

Si la opción “Manual Selection” es seleccionada, la ventana de configuración muestra 2 conjuntos de columnas (Fig. 2.21):

- Las columnas que se incluirán en la tabla de datos (“Include” a la derecha)
- Las columnas que se excluirán de la tabla de datos ) (“Exclude” establecido a la izquierda)

Dos botones “Search” permiten buscar una columna específica

Puede agregar y eliminar columnas de un conjunto a otro usando los botones “Add” y “Remove”.

- **“Enforce Inclusion”** mantiene el grupo “Include” fijo. Si se agrega una columna de entrada más del nodo anterior, esta nueva columna se inserta automáticamente en el grupo “Exclude”.
- **“Enforce Exclusion”** mantiene el grupo “Exclude” fijo. Si se agrega una columna de entrada más del nodo anterior, esta nueva columna se inserta automáticamente en el grupo “Include.”

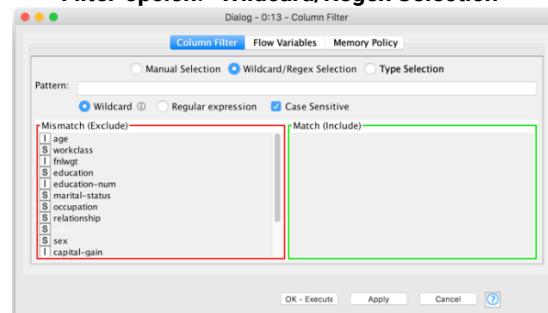
## Selección Wildcard/Regex

En caso de que la opción “Wildcard/Regex Selection” está seleccionada, la ventana de configuración presenta un cuadro de texto para editar el comodín deseado o la expresión regular.

Los botones de opción debajo del cuadro de texto especifican si se trata de una expresión regular o una expresión “wildcard”. Una casilla de verificación adicional habilita una coincidencia sensible a mayúsculas y minúsculas.

Las columnas cuyo nombre coincide con la expresión se incluirán en la tabla de datos de salida.

### 2.21. Configuración del Nodo “Column Filter” opción: “Wildcard/Regex Selection”



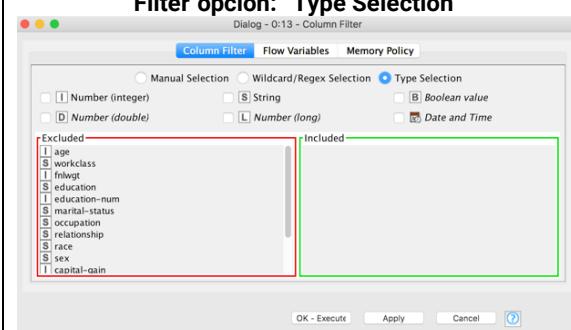
## Selección Type

Si la opción “Type Selection” es seleccionada, se le presenta una serie de casillas de verificación sobre los tipos de columnas que se deben mantener en la tabla de datos de salida.

Por ejemplo, seleccionando la casilla all Numbers, mantendrá todas las columnas numéricas solo en la tabla de salida del nodo.

Seleccionando String mantendrá todas las columnas del tipo I String solo en la tabla de salida del nodo.

### 2.22. Configuración del Nodo “Column Filter” opción: “Type Selection”



Recuerde este marco de selección de columna que viene con la opción “Manual Selection”, porque volverá a aparecer en todos aquellos nodos que requieran selección de columna, porque volverá a aparecer en todos aquellos nodos que requieran selección de columna.

En nuestro workflow de ejemplo “my First Workflow”, queremos remover la columna “final weight”.

Configuramos el modo de filtro de columna en “Manual Selection” y llenamos el panel Exclude con la columna “final weight”. Luego habilitamos “Enforce Exclusion”, porque queríamos mantener todas las columnas de entrada nuevas, si las hubiera, y siempre excluir solo la columna “final weight”.

Después de completar la configuración, hacemos clic con el botón derecho en el nodo “Column Filter” y lo comentamos con “rm column ‘final weight’”

Finalmente hacemos clic derecho en el nodo y seleccionamos “Execute” para ejecutar el filtro de columna.

Para ver los datos procesados finales, hacemos clic derecho en el nodo “rm column “final weight”” y seleccionamos la opción “Filtered Table”. La columna “final weight” no se encontrará en la tabla de datos de salida del Column Filter

## 2.23. La tabla filtrada por columnas no contiene la columna “final weight”

## 2.6. Filtrado de “Data Rows”

Si ha examinado los datos que estamos analizando actualmente, habrá visto que cada registro describe a una persona en términos de edad, trabajo, educación y otra información demográfica general. Hemos visto cómo eliminar una columna de datos de una tabla de datos. Veamos ahora cómo excluir filas de datos (“Data rows”) de una tabla de datos.

Supongamos que queremos conservar todos los registros de personas nacidas fuera de los Estados Unidos. Es decir, queremos conservar solo aquellas filas con “native-country” que no sea “United States”. Necesitamos usar un nodo “Row Filter” .

## Creando un nodo “Row Filter”

En el panel “Node Repository”,abra la categoría “Manipulation” y busque el nodo “Row Filter” in “Manipulation” → “Row” → “Filter” o busque “Row Filter”en el cuadro de búsqueda.

Arrastre y libere o haga doble-click en el nodo “Row Filter”en el “Node Repository”para crar una nueva instancia en el editor de workflow

La descripción de la tarea y la configuración para este nodo se puede encontrar en el panel “Node Description” panel a la derecha o haciendo clic en el botón de ayuda en la ventana de configuración a la derecha del botón “Cancel”.

Conecte el nodo “Row Filter” con el nodo “Column Filter” previamente creado.

## Configurando el nodo “Row Filter”

Haga doble-click en el nodo “Row Filter”para abrir la ventana de configuración del nodo.

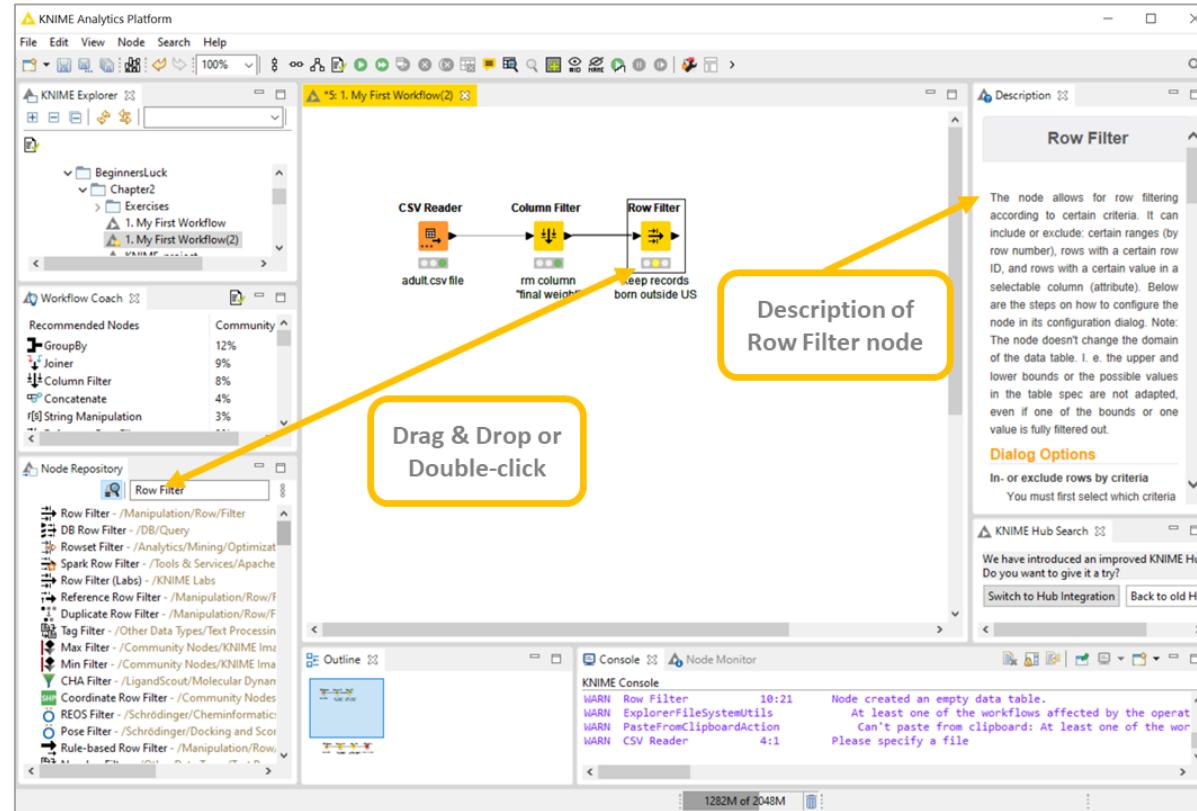
El nodo implementa tres criterios de filtrado:

- Seleccion de rows **attribute value** (coincidencia de patrones )
  - Value matching: el valor de la columna coincide con algún valor de patrón predefinido (se permiten “wild cards”y “regex expressions” en la definición de patrón)
  - Range checking : para columnas numéricas: valor de columna por encima o por debajo de un valor dado
  - Coincidencia de valores faltantes
- Seleccion de rows por **row number**
- Seleccion de rows por **RowID** (coincidencia de patrones en RowID)

Cada uno de estos criterios puede ser usado par **include** o **exclude rows**.

- Implemente su criterio de filtrado
- Click “OK”

## 2.24. Creando un nodo "Row Filter"



A continuación, puede encontrar una descripción más detallada de los criterios de filtrado de filas disponibles en la configuración del nodo "Row Filter"

El nodo Row Filter no es la única forma de realizar el filtrado de filas en KNIME, aunque probablemente sea la más fácil y funcione para el 80% de sus necesidades de filtrado de filas. Otras opciones de filtrado de filas :

El nodo "Nominal Value Row Filter" para la coincidencia de múltiples patrones. "OR mode" (example: native-country = "United Stated" OR native-country="Canada" OR native-country="Puerto Rico");

- El nodo “Rule Based Row Filter” para definir un conjunto arbitrario complejo de reglas de filtrado de filas IF-THEN, incluso abarcando varias columnas
- El nodo ;“Geo-location Row Filter” en KNIME Labs category para filtrado de filas basado en coordenadas geográficas;
- El nodo “Date&Time-based Row Filter” para filtrado de filas en base a Date&Time;
- El nodo “Database Row Filter” para implementar un filtro de filas SQL query que corra directamente en una “database”.

## Criterios de filtrado (Row filter)

### Por attribute value

Todas las filas, para las que el valor de una columna determinada coincide con un patrón predefinido, se filtran o conservan.

Despues que seleccione “select the column to test”, se necesita definir el modelo de coincidencia .

Para **String/Integer/Date&Time values**, “use pattern matching”

Se requiere que el patrón dado se ingrese manualmente o se seleccione de un menú poblado con los valores de columna como posibles valores de patrón.

Para **Integer values**, “use range checking” requiere un límite inferior y / o un límite superior, que coincidirán si la condición es igualdad.

Para **Missing values**, elija la última opción

### Por row number

Si sabe dónde están las filas deseadas o no deseadas, puede ingresar el **row number range** de fila para filtrar.

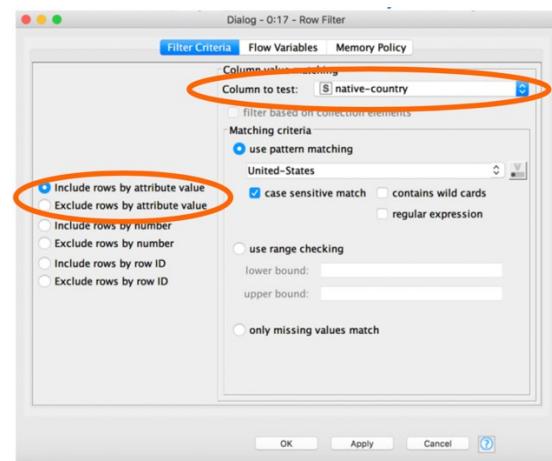
Por ejemplo, si se sabe que las primeras 10 filas son comentarios o simplemente basura, seleccionaría el criterio de filtro “exclude row by number”y establecería el rango de números de fila 1-10.

### Por RowID

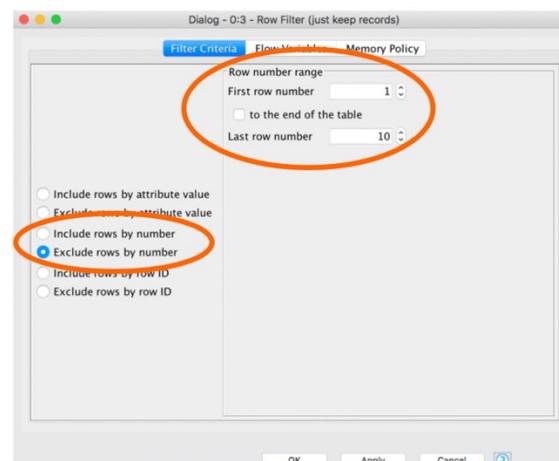
Un filtro de fila especial por valor de atributo se ejecuta en el RowIDs.

Aquí, el patrón de coincidencia viene dado por una expresión regular. La expresión regular tiene que coincidir con el whole RowID

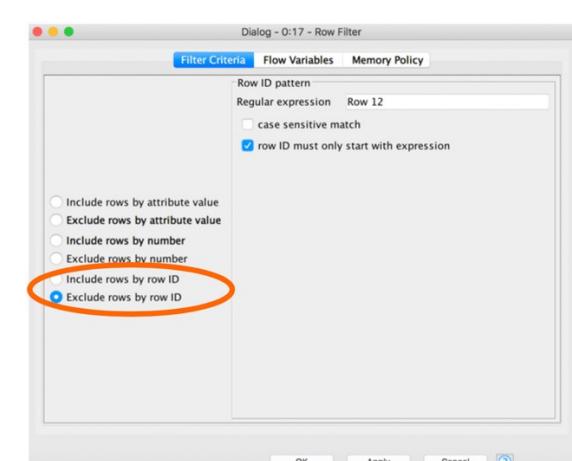
## 2.25. Criterio de Row filter por attribute value



## 2.26. Criterio de Row filter por row number



## 2.27. Criterio de Row filter por RowID



In order to retain all rows with data referring to people born outside of the United States, we need to:

- Seleccione el modo de filtrado “exclude row by attribute value”
- Seleccione la columna “native-country”
- Habilite “use pattern matching” por que es una comparacion de strings
- Seleccione el patron “United States”

Acabamos de implementar el siguiente criterio de filtrado:

`native-country != "United States"`

- Comente el nodo “Row Filter” con un comentario significativo. Lo comentamos con “just keep records born outside US”.
- Give the “Row Filter” node a meaningful comment. We commented it with “just keep records born outside US”. El comentario sobre un nodo es importante para fines de documentación. Dado que KNIME es una herramienta gráfica, es fácil tener una visión general de lo que hace un flujo de trabajo, si el nombre de cada nodo da una indicación clara de su tarea.
- Haga click-derecho y seleccione “Execute” para correr el row filter

2.28. A tablade salida no tiene "United States" en la columna "native-country"

Row ID	sex	capital...	capital-l...	hours-p...	native-country	income
Row4	Female	0	0	40	Cuba	<=50K
Row6	Female	0	0	16	Jamaica	<=50K
Row11	s... Male	0	0	40	India	>50K
Row14	s... Male	0	0	40	?	>50K
Row15	... Male	0	0	45	Mexico	<=50K
Row27	s... Male	0	0	60	South	>50K
Row35	Male	0	0	40	Puerto-Rico	<=50K
Row38	Male	0	0	38	?	>50K
Row51	Female	0	0	30	?	<=50K
Row52	Female	0	1902	60	Honduras	>50K
Row56	Male	0	0	40	Mexico	<=50K
Row57	Male	0	0	40	Puerto-Rico	<=50K
Row61	Male	0	0	40	?	<=50K
Row75	Male	0	0	40	Mexico	<=50K
Row81	Male	0	0	40	Cuba	<=50K
Row93	s... Female	0	1573	35	?	<=50K
Row98	Female	0	0	40	England	<=50K

Para ver los datos procesados haga click-derecho en el nodo “born outside US” y seleccione “Filtered”.

No observara a “United States” en la columna “native-country”.

## 2.7. Guardar en un archivo (Write Data to a File)

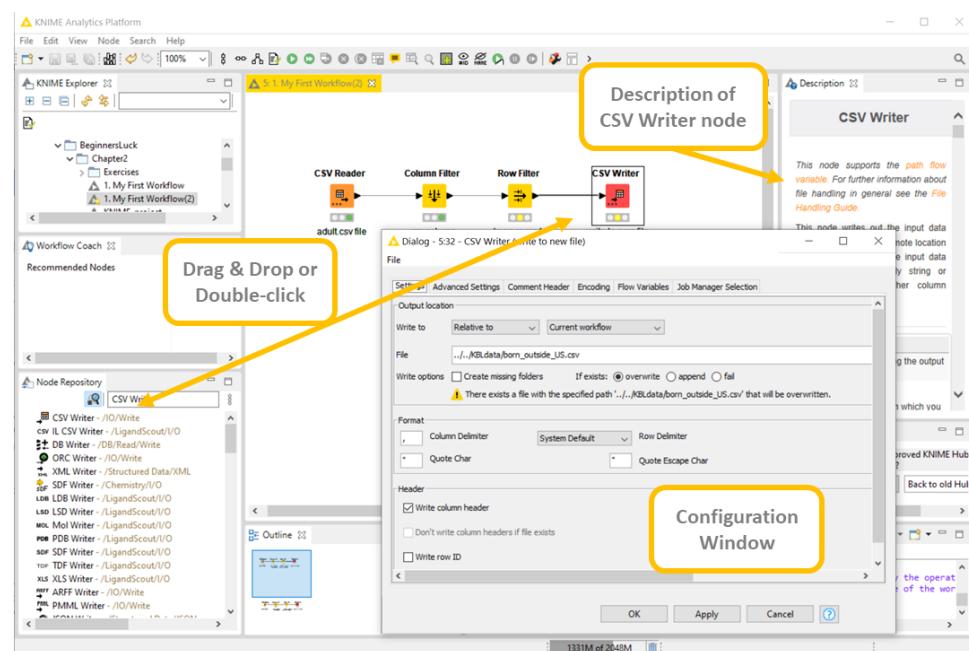
Ahora queremos escribir la tabla de datos procesados en un archivo. Hay muchos nodos que pueden escribir en un archivo. Elijamos el formato más sencillo y estándar por ahora: el formato CSV (valores separados por comas).

# Creando un nodo "CSV Writer"

En el "Node Repository":

- Despliegue la categoría "IO"/"Write" o busque "CSV Writer" en el cuadro de búsqueda
- Arrastre y suelte (o haga doble clic) en el nodo "CSV Writer" para crear un nuevo nodo en el workflow editor
- Haga clic con el botón derecho en el nodo "CSV Writer" y seleccione "Configure" o haga doble clic en él para abrir su ventana de configuración.
- La ventana de configuración tiene cuatro pestañas: Settings, Advanced Settings, Comment Header, and Encoding (Configuración, Configuración avanzada, Encabezado de comentarios y Codificación)

2.29. Crear y configurar un nodo "CSV Writer"



La ventana de configuración del nodo CSV Writer es similar a la ventana de configuración del nodo CSV Reader.

## Configurar el nodo "CSV Writer"

La pestaña “Settings” tab es la más importante de esta ventana de configuración. Requiere:

- La ruta del archivo de salida en Output Location. Observe que Output Location ofrece las mismas opciones que Input Location en el nodo CSV Reader.
- Algunas opciones adicionales sobre la estructura de datos, tales como:
  - o Si se escriben las column headers y/o RowID en el output (archivo de salida)
  - o El carácter delimitador de columna
  - o El modo de escritura si el archivo ya existe
    - Overwrite
    - Append
    - Fail (no escribe el archivo)

La pestaña **“Advanced Settings”** permite algunas especificaciones más, como el carácter de las comillas, el separador decimal o la compresión a un archivo gzip.

La pestaña **“Comment Header”** permite escribir automáticamente una cabecera con comentarios sobre los datos.

La pestaña **“Encoding” tab** elige la codificación adecuada para el texto.

Tab **“Memory Policy”** contains a few options that might speed up the node execution. This tab is common to the configuration window of all nodes.

La pestaña **“Memory Policy”** contiene algunas opciones que pueden acelerar la ejecución del nodo. Esta pestaña es común a la ventana de configuración de todos los nodos.

En este libro no investigamos las pestañas “Flow variables” y “Job Manager Selection”.

**Nota.** Escribir en el modo "Append" puede ser complicado, porque simplemente añade los nuevos datos a un archivo existente sin comprobar la estructura de los datos ni cotejar la columna por su nombre. Por lo tanto, si la estructura de la tabla de datos ha cambiado, por ejemplo debido a columnas nuevas o eliminadas, el archivo CSV de salida ya no será consistente

En algunos casos, es posible que desee seleccionar "Fail" como modo de sobreescritura, para evitar sobrescribir el archivo existente.

Ahora vamos a cambiar el comentario del nodo:

- Haga clic en la etiqueta del nodo bajo el nodo
- Introduzca el nuevo comentario del nodo (por ejemplo "write new file")
- Haga clic en otro lugar
- - Haga clic con el botón derecho del ratón en el nodo y seleccione "Execute"

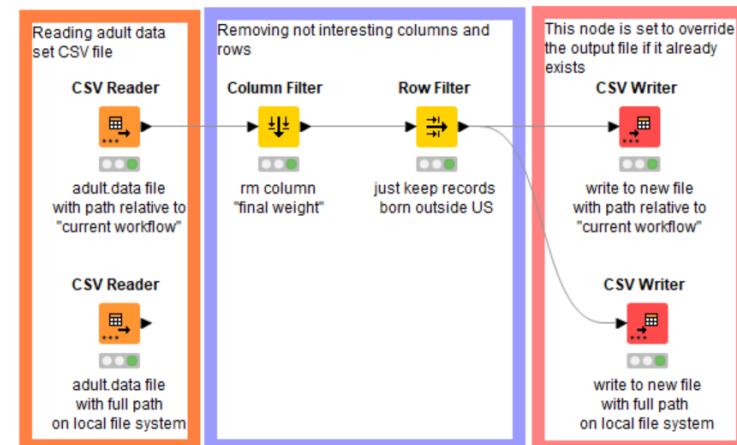
También puedes hacer que los comentarios de los nodos sean más verbales, si quieres añadir más información sobre la configuración del nodo y la tarea implementada.

En este punto también añadimos algunas anotaciones para dejar aún más claro lo que hace cada nodo o grupo de nodos.

Hemos creado nuestro workflow (My First Workflow) para leer datos de un archivo, reorganizar filas y columnas, y finalmente escribir los datos en un archivo de salida.

El workflow final, denominado "My First Workflow", está disponible en la carpeta de workflows que ha descargado del KNIME Hub.

### 2.30. Workflow "My First Workflow"



**Workflow: My first workflow**  
This workflow is my first KNIME workflow.  
It reads data, removes uninteresting columns and rows, and writes the resulting data table to a CSV file.

## 2.8. Ejercicios

### Ejercicio 1

En un workflow group "Exercises" en el grupo de workflows existente "Chapter2", crear un flujo de trabajo vacío "Exercise1". El workflow "Exercise1" debe realizar las siguientes operaciones:

- Leer el archivo data1.txt (de la carpeta KBLdata) con la columna "ranking" como String y denominada "marks";
- Eliminar los comentarios iniciales de los datos leídos del archivo;
- Eliminar la columna "class"
- Escriba los datos finales en el archivo en formato CSV (por ejemplo, con el nombre "data1\_new.csv") utilizando el carácter ";" como separador

Introduzca una breve descripción de todos los nodos del flujo de trabajo.

Guarde y ejecute el flujo de trabajo "Ejercicio1". La ejecución debe realizarse sin errores (luces verdes para todos los nodos).

#### Solution to Exercise 1

El archivo tiene algunos comentarios al principio, que por supuesto no tienen la misma longitud que las otras líneas del archivo.

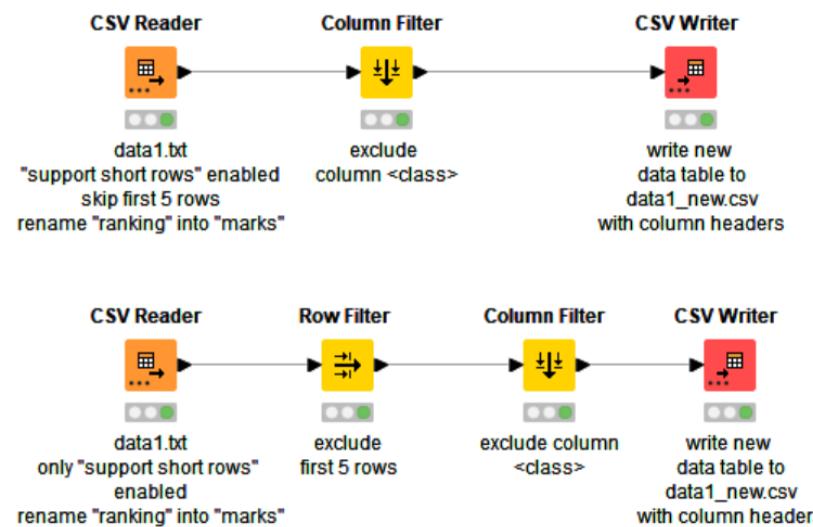
Primero, debe habilitar las opciones "has column headers" y "support short data rows" en la pestaña "Settings" y cambiar el nombre de la columna "ranking" as "marks" en la pestaña "Transformation"

Entonces puedes hacer una de dos cosas:

- 1) puede establecer "Skip first data rows ..." en 5 en "Limit Rows"
- 2) puede utilizar el nodo "Row Filter" para excluir las primeras 5 filas de los datos leídos.

Nuevamente, el flujo de trabajo de la solución está disponible en la carpeta de flujos de trabajo que descargó de KNIME Hub

## 2.32. Ejercicio 1: dos posibles alternativas



### Workflow: Chapter 2/Exercise 1

This exercise applies:

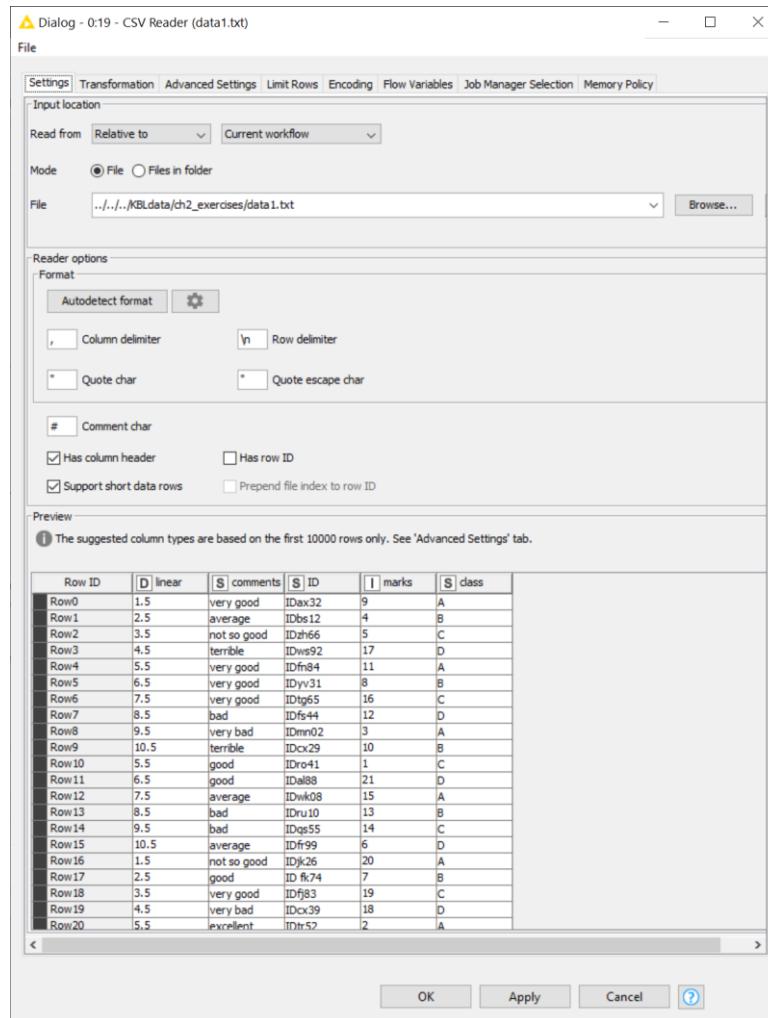
- advanced file reading features
- column and row filtering
- writing file

Note that the data1.txt file has a header with comments. The two workflows apply two different methods to exclude the comment header from the read data:

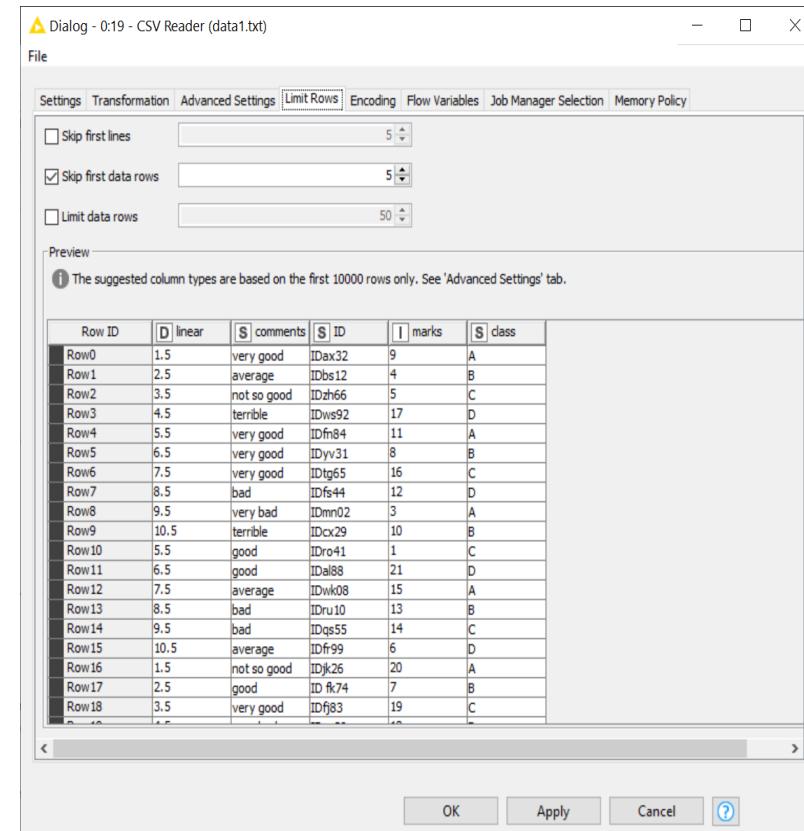
1. File Reader with "allow short lines" on and "skip first 6 lines" in "Limit Rows" tab in "Advanced" page

2. File Reader with only "allow short lines" on in "Advanced" page and subsequent row filter to remove first 5 spurious rows

### 2.33. Ejercicio 1: configuracion del "CSV Reader" "Settings"

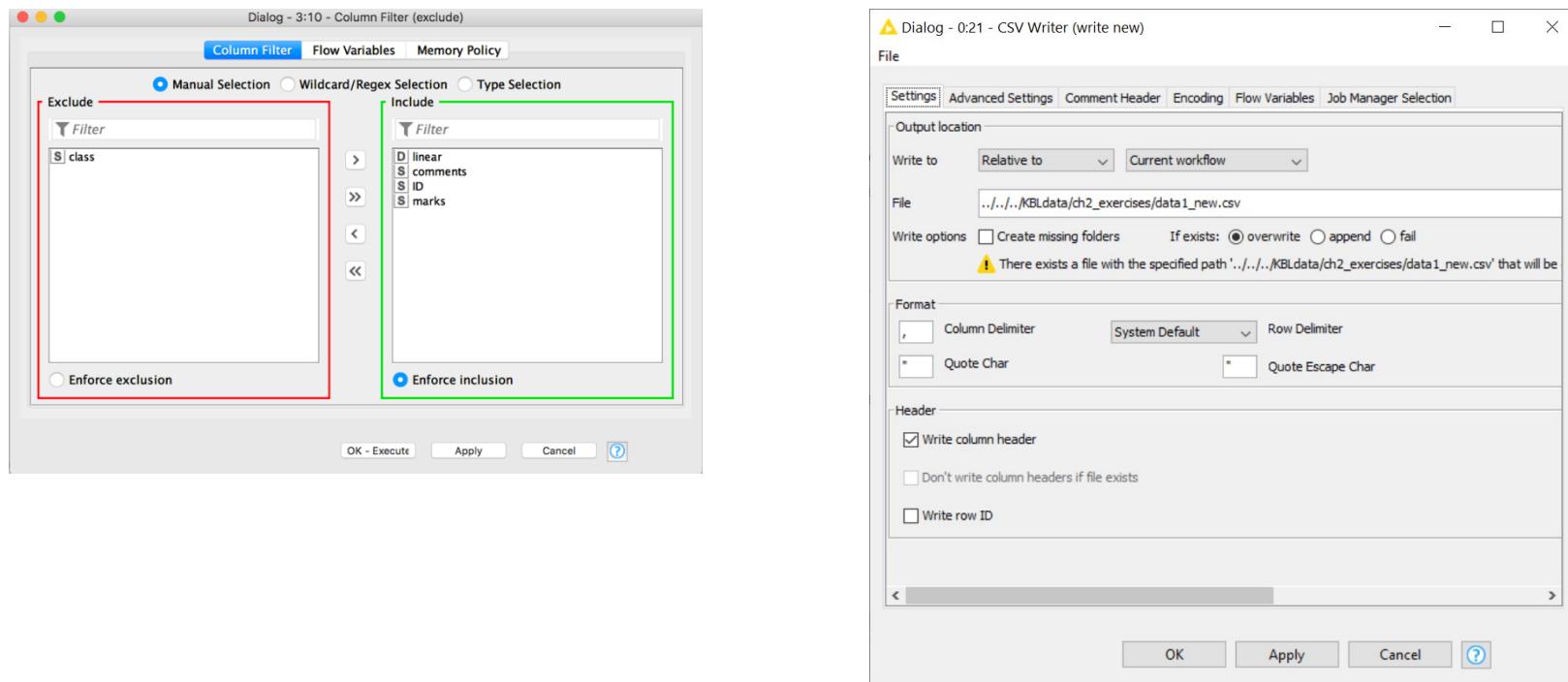


### 2.34. Ejercicio 1: configuracion pestaña "CSV Reader" "Limit Rows"



### 2.35. Ejercicio 1: configuracion "Column Filter" configuration

### 2.36. Ejercicio 1: configuracion "CSV Writer" "Settings"



## Ejercicio 2

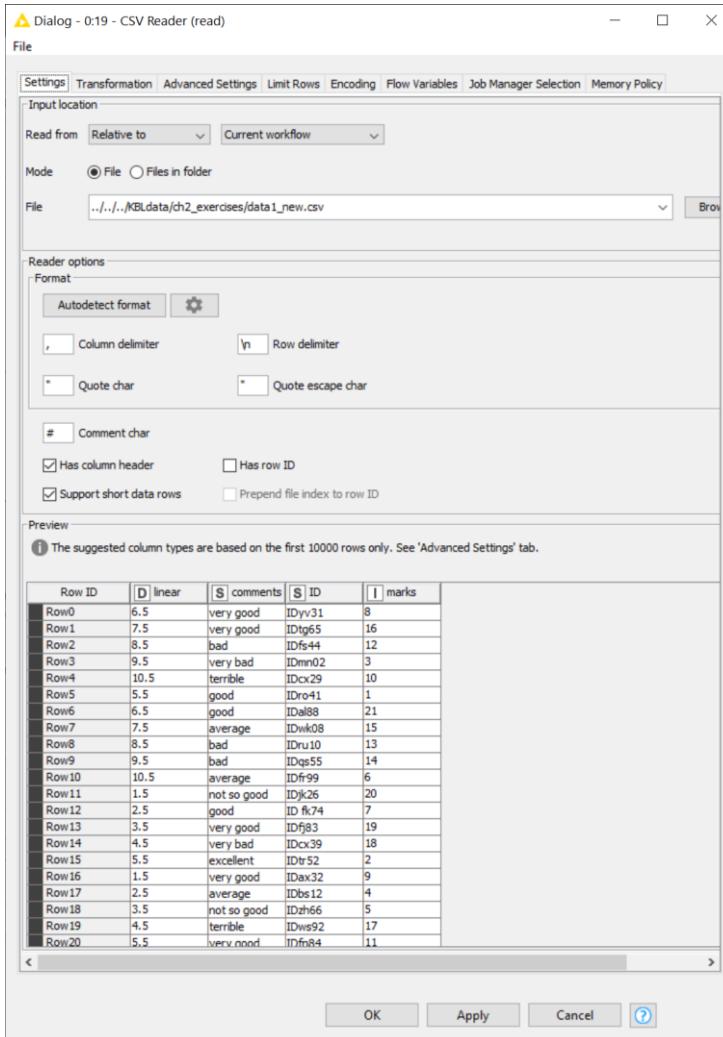
En el workflow group “Chapter2\Exercises” crear un workflow “Exercise2” para realizar las siguientes operaciones:

- Leer el archivo CSV escrito en el Exercise 1 (“data1\_new.csv”) y cambiar el nombre de la columna “marks” a “ranking”
- filtrar las filas con valor ‘average’ en la columna ‘comments’
- Excluir las columnas con datos del tipo Integer
- Escribir los datos finales en el archivo en modo “Append” y con tab como carácter de separación
- Cambie el nombre de todos los nodos cuando sea necesario. Guarde y ejecute el workflow “Exercise2”.

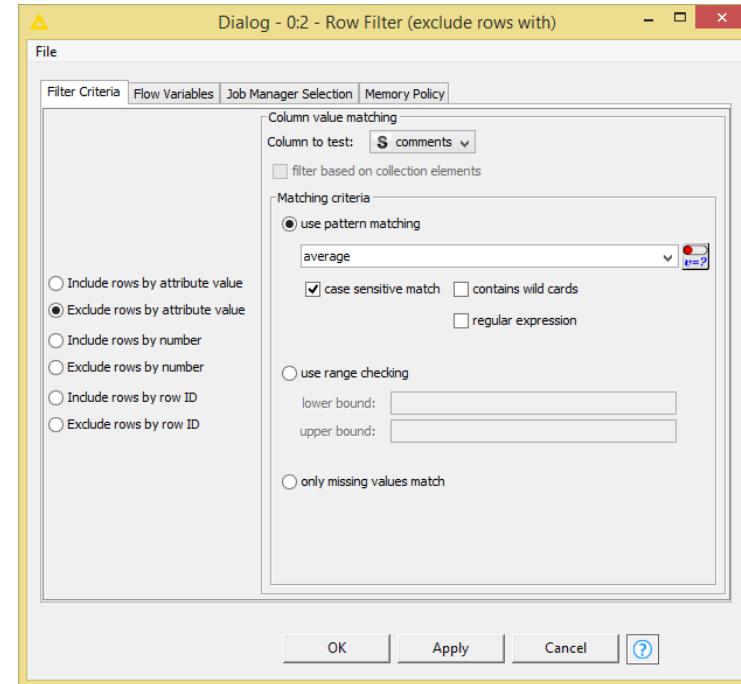
### Solucion al ejercicio 2

Reciclamos la estructura del workflow creado en el Ejercicio 1. Es decir, hicimos una operación de “Copy and Paste” (Ctrl-C, Ctrl-V) en todo el “Exercise 1” workflow desde el “workflow editor” para “Exercise 1” al workflow editor del “Exercise 2”.

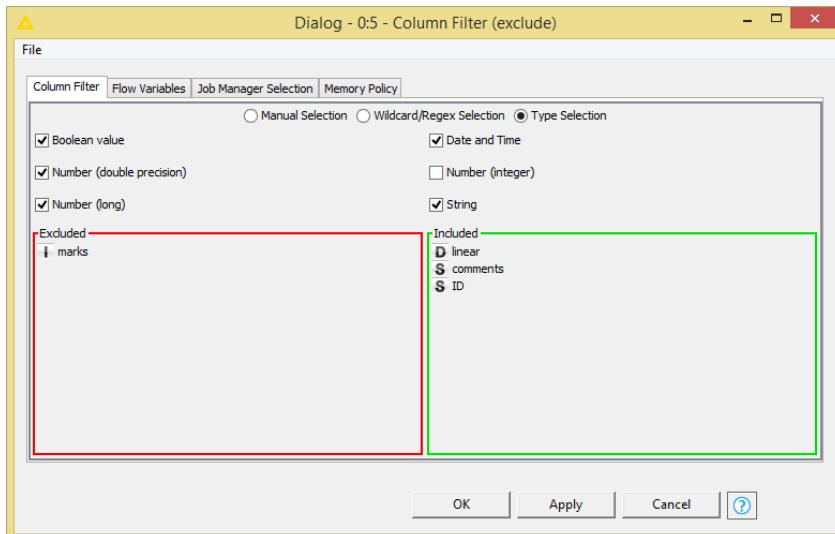
### 2.37. Ejercicio 2: configuración del "CSV Reader"



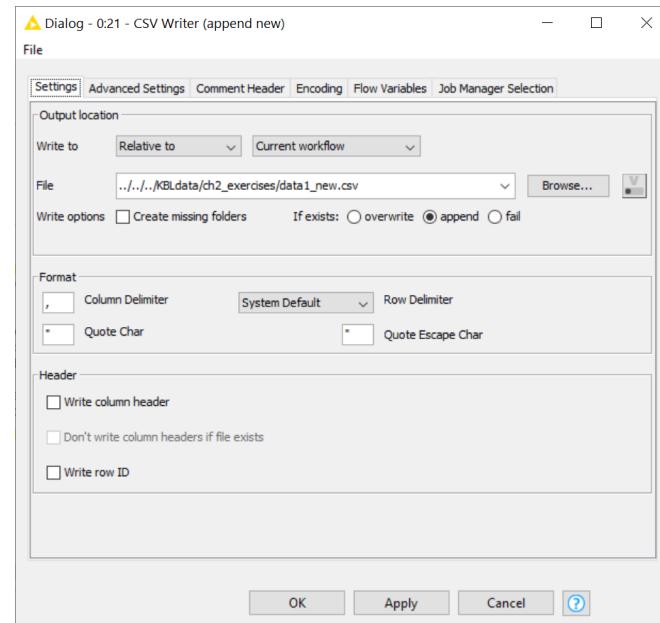
### 2.38. Ejercicio 2: configuración del "Row Filter"



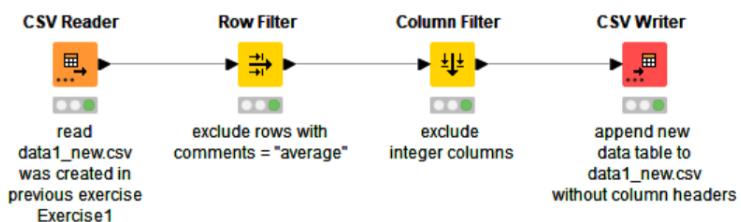
### 2.39. Ejercicio 2: configuración del: "Column Filter"



### 2.40. Ejercicio 2: configuración del "CSV Writer"



### 2.41. Ejercicio 2: workflow



#### Workflow: Chapter 2/Exercise 2

This exercise:  
 - reads the data file data1\_new.csv generated in the previous exercise Chapter2/Exercise 1.  
 - filters rows and columns  
 - writes resulting data table to a new file

The particularity here is in the Column Filter, since we *filter out all Integer columns* using a type based filtering and not just a manual filtering.

**Nota.** Después de copiar el nodo "CSV Reader" del Exercise 1, es necesario desactivar la opción option "**Limit Rows**" en la ventana "Advanced Settings", porque este archivo no tiene comentarios iniciales.

**Nota:** Observe el triángulo amarillo bajo el nodo ""Column Filter". Se trata de un mensaje de advertencia que proviene de la copia del workflow y que permanece incluso cuando el nodo tiene la luz verde. Al pasar el ratón por encima del triángulo amarillo aparece el mensaje de advertencia "*Some columns are not available: marks*". ("Algunas columnas no están disponibles: marcas"). Esto es correcto: la columna "marks" ya no está ahí, porque hemos renombrado la columna "marks" como "ranking". Sin embargo, el filtro de columnas sigue funcionando; sólo se emite un mensaje de advertencia. Si abrimos la ventana de configuración del filtro de columnas, vemos que la columna "ranking" se insertó automáticamente en el conjunto "Exclude". Esto se debe a que se activó la opción "marks" as "ranking". Al hacer clic en el botón "Accept" se aceptan los ajustes de configuración actuales y desaparece el triángulo amarillo de advertencia.

**Nota:** Hemos guardado los datos en modo "Append" en el archivo CSV. Los datos del Exercise 2 sólo tienen 3 columnas, mientras que los datos existentes en el archivo tienen 4 columnas. El nodo "CSV Writer" no comprueba la consistencia del número y la posición de las columnas a escribir con el número y las posiciones de las columnas existentes. Por lo tanto, es posible escribir datos inconsistentes en un archivo. Hay que tener cuidado cuando se trabaja en modo "Append" con un nodo "CSV Writer".

# Capítulo 3. Mi primera exploración de datos

## 3.1. Introducción

En este capítulo se describen algunos nodos de manipulación de datos útiles para darles la forma deseada, que implican la transformación de datos, la manipulación de cadenas, la aplicación de reglas y otras tareas similares. Para ello, utilizaremos tres workflows: "Column Rename Example", "Write To DB" y "My First Data Exploration". "Column Rename Example" es un workflow muy sencillo que muestra el uso del nodo de cambio de nombre de columna, "Write To DB" escribe datos en una base de datos, y "My First Data Exploration" lee los mismos datos de la base de datos y explora gráficamente los datos.

El objetivo de este capítulo es familiarizarse con

- los nodos y las opciones para el manejo de la base de datos
- la categoría "Vistas" que contiene nodos para la exploración gráfica de datos
- algunos nodos más de operación de columnas, como los nodos para la manipulación de cadenas y el manejo de valores perdidos

Empezamos con el muy conocido conjunto de datos Iris, descargado del UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets/Iris>) y disponible en la carpeta KBLdata, para preparar los datos para la siguiente exploración gráfica. El conjunto de datos Iris describe una serie de plantas de iris mediante 4 atributos:

- the sepal length
- the sepal width
- the petal length
- the petal width

Las plantas descritas en el conjunto de datos pertenecen a tres clases

Diferentes de Iris: Iris setosa, Iris versicolor, and Iris virginica.

3.1. Conjunto de datos IRIS					
File Table - 0:43 - CSV Reader (read the iris.data file)					
File Edit Hilite Navigation View					
Row ID	D Column0	D Column1	D Column2	D Column3	S Column4
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

Este conjunto de datos se ha utilizado durante muchos años como estándar de clasificación. Las tres clases no son linealmente separables.

Sólo dos de las tres clases de iris pueden separarse utilizando una función lineal sobre dos de los cuatro atributos numéricos. Para la tercera clase tenemos que utilizar algo más sofisticado que una separación lineal. Las dos primeras clases y su posible separación lineal pueden identificarse claramente utilizando gráficos. Esta es la razón por la que utilizamos este conjunto de datos para ilustrar los nodos KNIME para la exploración visual de datos.

También utilizaremos este capítulo para explorar la manipulación de cadenas y cómo crear nuevos valores basados en reglas a partir de los valores de las columnas existentes.

En el panel "KNIME Explorer", creamos ahora un nuevo grupo de flujos de trabajo llamado "Chapter3", para contener todos los workflows creados en este capítulo del libro. En el grupo de flujos de trabajo "Chapter3", creamos dos workflows vacíos: "Write To DB" y "My First Data Exploration". Como dijimos al principio de esta sección, "Write To DB" mostrará cómo construir un nuevo conjunto de datos y cómo escribirlo en una base de datos, mientras que "My First Data Exploration" describirá cómo realizar una exploración visual de los datos.

Comencemos con la lectura de los datos en el workflow "Column Rename Example". Dado que el archivo del conjunto de datos de Iris (iris.data) no está en un formato de datos estándar, no podemos arrastrarlo y soltarlo directamente en el editor del workflow. En su lugar, lo leemos con un nodo "CSV Reader". Si no cambiamos el nombre de las columnas de datos en la pestaña Transformation, el nodo leerá una tabla de datos como la de la figura 3.1. Escribimos el comentario "read the iris.data file from KBLdata folder" ("leer el archivo iris.data de la carpeta KBLdata") bajo el nodo "CSV Reader", para una rápida visión de la tarea del nodo.

**Note.** El archivo del conjunto de datos del iris no contiene nombres de columnas. El nodo "CSV Reader" entonces asigna a cada columna un nombre por defecto como "Column0", "Column1", "Column2", "Column3", and "Column4". Además de "Column4", donde podemos ver que se trata de la clase iris, tenemos que leer las especificaciones del archivo en el archivo "iris.names" para entender qué columna representa cada atributo numérico.

## 3.2. Reemplazar valores en las columnas

Después de leer la descripción del conjunto de datos del iris en el archivo iris.name, descubrimos que las cinco columnas están organizadas como sigue:

1. Sepal length en cm
2. Sepal width en cm
3. Petal length en cm
4. Petal width en cm
5. clase

y, además, que no hay Missing data en el conjunto de datos. Por lo tanto, el primer paso es renombrar las columnas del conjunto de datos, para poder hablar claramente de lo que estamos haciendo en los datos. KNIME tiene un nodo "Renombrar columnas" que se utiliza exactamente para este propósito.

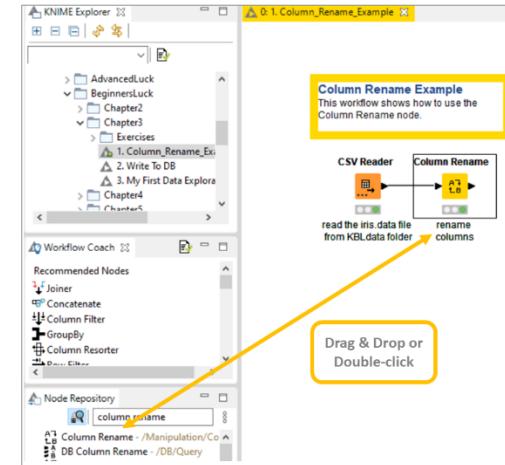
## Column Rename (Renombrar una columna)

El nodo "Column Rename" puede ser encontrado en el panel "Node Repository" :

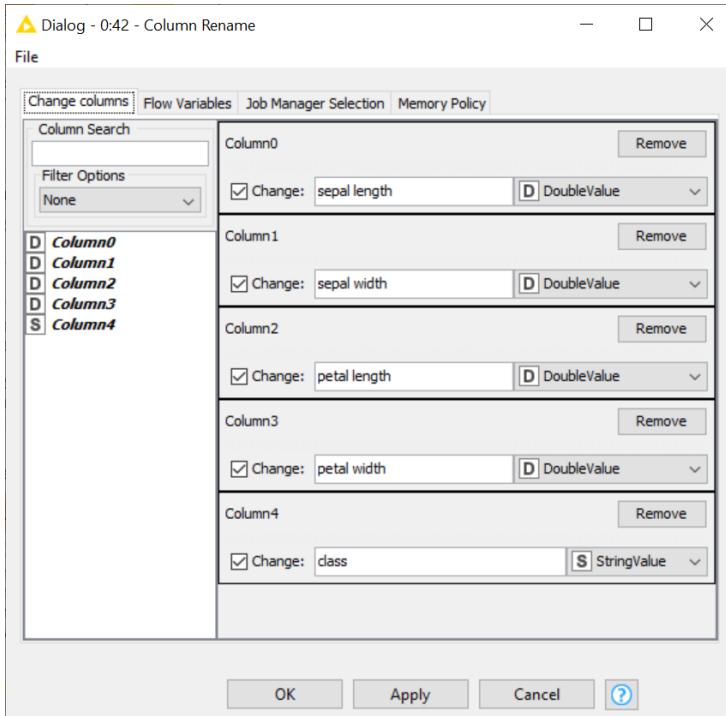
"Manipulation" → "Column" → "Convert & Replace"

El nodo "Column Rename" permite renombrar las columnas de la tabla de datos de entrada.

3.2. Crear un nodo "Column Rename"



### 3.3. Diálogo de configuración del nodo "Column Rename"



En la ventana de configuración tenemos a la izquierda la lista de columnas candidatas a ser renombradas; a la derecha la lista de las columnas realmente seleccionadas para ser renombradas.

El diálogo de configuración requiere:

- seleccionar las columnas sobre las que se va a operar haciendo doble clic en el panel de la izquierda
- marcar las columnas cuyo nombre o tipo hay que cambiar (casilla de verificación)
- proporcionar los nuevos nombres de las columnas
- y opcionalmente los nuevos tipos de columnas
- El botón "Remove" está ahí para quitar una columna ya seleccionada del panel de renombrado

Creamos un nodo "Column Rename" y lo conectamos al nodo "CSV Reader". A continuación, asignamos nuevos nombres a las columnas de la tabla de datos según el archivo de especificación "Iris.name" y ejecutamos el comando "Execute". La misma operación podría haberse ejecutado en la pestaña de Transformación del nodo "CSV Reader". Esto es lo que hicimos en el segundo flujo de trabajo de este capítulo "Write To DB".

Supongamos ahora que los nombres de los iris en la columna "clase" son demasiado largos o complejos para nuestra tarea y que nos gustaría tener sólo números de clase: "class 1", "class 2", and "class 3". Es decir, nos gustaría añadir una columna en la que "Iris-setosa" de la columna "class" se traduzca en "class 1", "Iris-versicolor" en "class 2", y finalmente todas las instancias restantes pertenezcan a una "class 3".

KNIME cuenta con un nodo muy práctico: el nodo “Rule Engine”. Este nodo define un conjunto de reglas sobre los valores de las columnas de datos de entrada y genera nuevos valores según el conjunto de reglas definido. Los nuevos valores pueden formar una nueva columna que se añada a las existentes en la tabla de datos de entrada o sustituir una columna de datos existente.

El conjunto de reglas que queremos aplicar en este caso es el siguiente:

IF class = "Iris-setosa"	THEN	class 1
IF class = "Iris-versicolor"	THEN	class 2
ELSE		class 3

El nodo “Rule Engine” usa la siguiente sintaxis para expresar el mismo conjunto de reglas:

```
$class$ = "Iris-setosa" => "class 1"  
$class$ = "Iris-versicolor" => "class 2"  
TRUE => "class 3"
```

Donde \$class\$ indica los valores de la columna de entrada “class”, “Iris-setosa” es la cadena de coincidencia para el “=” operator, “=>” introduce la regla , y “class 1” es el valor resultante de ejecutar la regla.

**Nota.** Los valores de las cadenas fijas deben ir entre comillas para que el nodo “Rule Engine” los interprete correctamente como cadenas.

The final keyword “TRUE” represents the ELSE value in our list of rules, i.e. the value that is always true if no other rule is applied first.

**Nota.** Para insertar un valor constante en una columna de datos, basta con utilizar

```
TRUE => <new constant value>
```

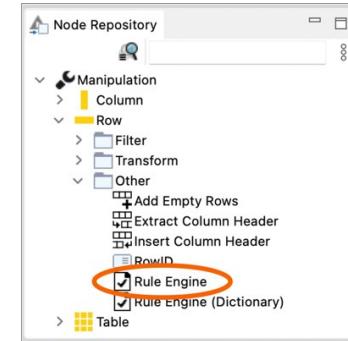
sin ninguna otra regla en un nodo “Rule Engine”. Como alternativa, puede utilizar el nodo “Constant Value Column” node.

# Nodo “Rule Engine”

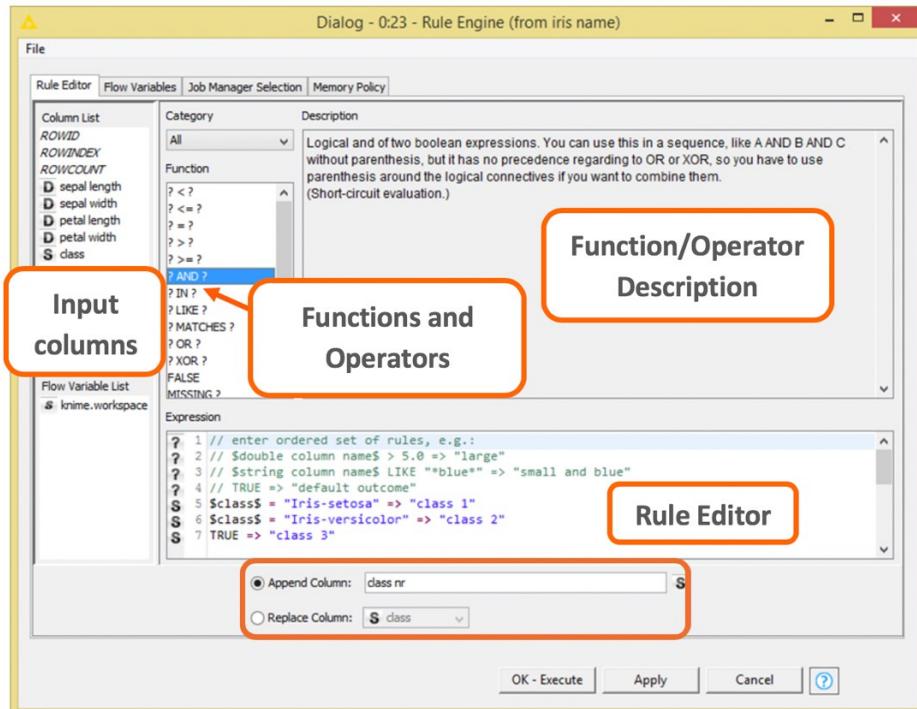
El nodo “Rule Engine” esta ubicado en el “Node Repository” en la categoría “Manipulation” → “Row” → “Other”.

Este nodo define un conjunto de reglas y crea nuevos valores basados en el conjunto de reglas y los valores de la columna de entrada.

3.4. Ubicación del nodo “Rule Engine” en el panel “Node Repository”



3.5. La ventana de configuración del nodo “Rule Engine”



El diálogo de configuración incluye:

- La lista de columnas de datos de entrada disponibles
- La lista de funciones y operadores disponibles
- Un panel de descripción para describir el uso y la tarea de la función/operador seleccionado
- Un editor de reglas donde editar el conjunto de reglas
- La opción de crear una nueva columna en la tabla de datos de salida o de sustituir una existente

**La lista de variables de flujo disponibles. Sin embargo, las variables de flujo se consideran un concepto avanzado y las ignoraremos en este libro.**

### **Column List**

El primer panel en la esquina superior izquierda de la ventana de configuración del "Rule Engine" muestra todas las columnas disponibles de la tabla de datos de entrada. Esas son las columnas sobre las que va a trabajar nuestro conjunto de reglas.

### **Flow Variable List**

El panel que se encuentra justo debajo del panel ""Column List" contiene todas las variables de flujo disponibles para el nodo. Sin embargo, las variables de flujo no se tratan en este libro y las ignoraremos al configurar nuestro conjunto de reglas.

### **Function**

El panel "Function" contiene una lista de funciones y operadores lógicos disponibles para crear el conjunto de reglas. El menú "Category" situado en la parte superior de la lista "Function", permite reducir la lista de funciones a un subconjunto más pequeño.

### **Description**

Si se selecciona una función o un operador en la lista "Function", este panel ofrece una descripción de su tarea y uso.

### **Expression**

El panel "Expression" es el editor de reglas. Aquí puede escribir su conjunto de reglas. Si necesita involucrar una columna de datos o una función, simplemente haga doble clic en el elemento deseado en el panel respectivo y aparecerá en el editor de reglas con la sintaxis correcta.

Toda regla consta de una condición (antecedente), incluida una función o un operador, y de un valor de consecuencia. El símbolo " $=>$ " lleva la condición al valor de la consecuencia, como: <antecedente>  $=>$  <valor de la consecuencia>. "TRUE" en la última regla lleva al valor por defecto, cuando no se aplica ninguna de las condiciones anteriores. La regla puede ser editada y cambiada en cualquier momento.

Para construir nuestro conjunto de reglas, escribimos el conjunto de reglas descritas anteriormente.

### **Append Column / Replace Column**

En la parte inferior de la ventana de configuración están las opciones para elegir si se crea una nueva columna de datos o se sustituye una existente. La opción por defecto es "Append Column" y el nombre por defecto para la nueva columna es "prediction". Seleccionamos la opción por defecto y llamamos a la nueva columna "class nr".

Tras la configuración, comentamos el nodo del motor de reglas como "from iris names to class no" y ejecutamos el comando "Execute".

### 3.3. Separación de cadenas de texto (String Splitting )

En esta sección exploramos cómo realizar la manipulación de cadenas con KNIME. Por ejemplo, ¿cómo podemos dividir la columna "class" de forma que tengamos "Iris" en una columna y "setosa", "versicolor" o "virginica" en otra? Y viceversa, ¿cómo puedo construir una clave para identificar de forma única cada fila de la tabla de datos?

En KNIME hay 3 nodos para dividir las celdas de las cadenas:

- **"Cell Splitter by Position"** divide cada cadena en función de la posición de los caracteres. La columna a dividir contiene valores de cadena. El nodo divide todas las cadenas de la columna en k subcadenas, cada una de ellas de longitud n<sub>1</sub>, n<sub>2</sub>, n<sub>3</sub>,... n<sub>k</sub>, donde n<sub>1</sub>+n<sub>2</sub>+n<sub>3</sub> +... n<sub>k</sub> = L es la longitud de las cadenas originales. Cada subcadena se coloca en una columna adicional. Tenga en cuenta que para este nodo todas las cadenas de entrada deben tener al menos L caracteres.
- **"Cell Splitter [by Delimiter]"** utiliza un carácter delimitador para extraer subcadenas de las cadenas originales. Las cadenas pueden ser de longitud variable. Si se encuentra el carácter delimitador, la subcadena anterior y posterior se colocaran en dos columnas adicionales diferentes. El nombre del nodo es, en realidad, simplemente "Cell Splitter". Sin embargo, dado que utiliza un carácter delimitador, lo llamaré "Cell Splitter [by Delimiter]". "Divisor de celdas [por delimitador]".
- **"Regex Split"** es un divisor de celdas por Regex. Utiliza una regla de Expresión Regular para reconocer las subcadenas. Una vez reconocidas las subcadenas, el nodo divide la cadena original en las subcadenas reconocidas y las coloca en diferentes columnas adicionales.

A diferencia de la operación de división de columnas, sólo hay un nodo para combinar columnas de cadenas: el nodo "Column Combiner"

- El nodo **"Column Combiner"** concatena las cadenas de dos o más columnas y coloca el resultado en una nueva columna anexa..

**Note.** Todos los nodos de manipulación de cadenas, como los nodos "Cell Splitter" y el nodo "Column Combiner", se encuentran en el panel "Node Repository" en "Manipulation" → "Column" → "Split & Combine"

En la columna "class" queremos separar la subcadena "Iris" del resto de subcadenas "setosa", "versicolor" o "virginica".

- Si dividimos por posición, necesitamos dividir en el 4º carácter (al final de "Iris" y antes del resto de la cadena) y en el 5º carácter (antes de "setosa", "versicolor" o "virginica").
- Si dividimos por el delimitador (split by delimiter), tenemos que dividir alrededor del carácter "-".

- Finalmente, si dividimos por RegEx, necesitamos encontrar una regla de Expresión Regular para expresar "Iris", "-", y las letras restantes. Una posible expresión regular podría ser: ((Iris)[\-\-]\*([A-Za-z]\*).

Veamos ahora en detalle cómo utilizar los tres nodos "Cell Splitter" para hacerlo.

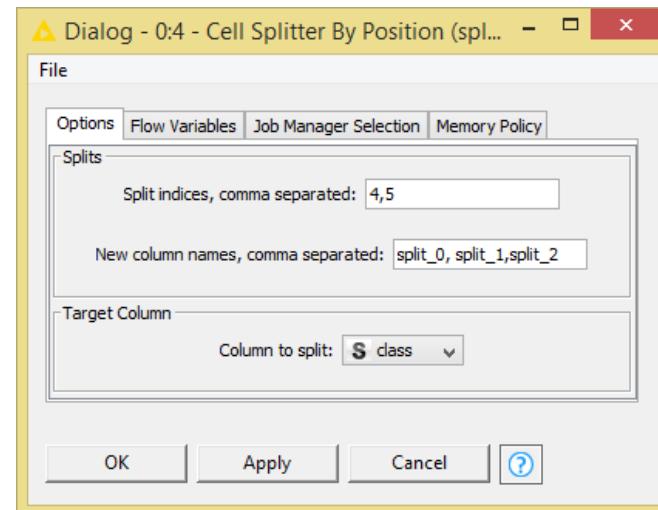
## Cell Splitter por Posición

Este nodo divide los valores de la cadena de columnas en función de la posición de los caracteres. El resultado consiste en tantas columnas nuevas como índices de posición más 1.

La ventana de configuración pide:

- Los índices de división (las posiciones de los caracteres dentro de la cadena en la que se va a dividir) separados por comas.
- Los nombres de las nuevas columnas (los nombres de las nuevas columnas son siempre uno más que el número de índices de división). Los nombres de las nuevas columnas deben estar separados por comas.
- El nombre de la columna de la cadena sobre la que se van a realizar las divisiones.

3.6. Cuadro de diálogo de configuración del nodo "Cell splitter" por posición"



Hemos elegido:

- índice de posición 4 (al final de la palabra "Iris") y 5 (después de "-")
- obtendremos 3 subcadenas: "Iris" in column "split\_0", "-" in column "split\_1", and "setosa"/"virginica"/"versicolor" in column "split\_2"
- la columna a dividir sera la "class"
- The column to perform the split on is "class"

La columna "split\_1" contendrá sólo cadenas "-". Siempre podemos eliminarla posteriormente mediante un nodo "Filtro de columna".

# Cell Splitter [by Delimiter]

Este nodo divide los valores de la cadena de columnas en un carácter delimitador. El resultado será tantas columnas nuevas como caracteres delimitadores se hayan encontrado más uno. La ventana de configuración requiere los siguientes ajustes:

- El nombre de la columna sobre la que se van a realizar las divisiones
- El carácter delimitador
- El tipo de salida:
  - Como nuevas columnas que se añaden a la tabla de datos (aquí es necesario establecer el tamaño de la matriz)
  - Como una sola columna que contiene la lista/conjunto de subcadenas (un conjunto de cadenas es como una lista pero sin valores duplicados)

Si se prevén muchas divisiones, la primera opción puede añadir rápidamente demasiadas columnas nuevas al conjunto de datos de salida y resultar inmanejable. Por el contrario, la segunda opción sólo añade una columna adicional al conjunto de datos de salida, compactando todas las subcadenas en una columna de tipo colección.

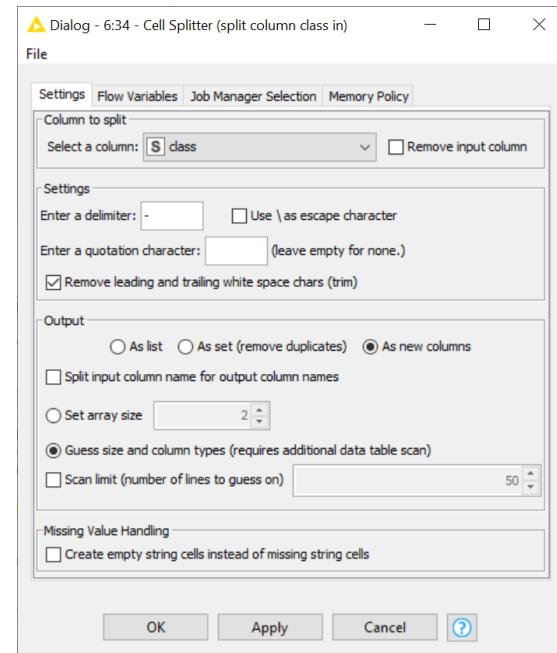
El tamaño de la matriz de subcadenas resultante puede fijarse a priori o, si no lo conocemos, podemos dejar que el nodo adivine el mejor tamaño. Esta última opción funciona para la mayoría de las tareas de división de cadenas. Para tareas más complejas, es posible que tengamos que establecer el tamaño del array manualmente nosotros mismos.

Si establecemos un tamaño de matriz fijo, si el tamaño de la matriz es menor que el número de subcadenas detectadas, se ignorarán las últimas divisiones. Por otro lado, si el tamaño del array es mayor que el número de subcadenas detectadas, las últimas columnas nuevas estarán vacías.

Seleccionamos:

- Columna a dividir = "class"

## 3.7. Diálogo de configuración del nodo "Cell Splitter"



- Carácter delimitador = "-"
  - Tamaño de la matriz = 2 y las subcadenas que deben salir como nuevas columnas
- Las subcadenas seán almacenadas en nuevas columnas con nombre "<original\_column\_name>\_Arr[0]" y "<original\_column\_name>\_Arr[1]", lo que está basado en nuestros parámetros de configuración "class\_Arr[0]" and "class\_Arr[1]".

**Nota.** Aquí no hay ninguna columna con sólo cadenas "-". Todos los caracteres "-" se pierden.

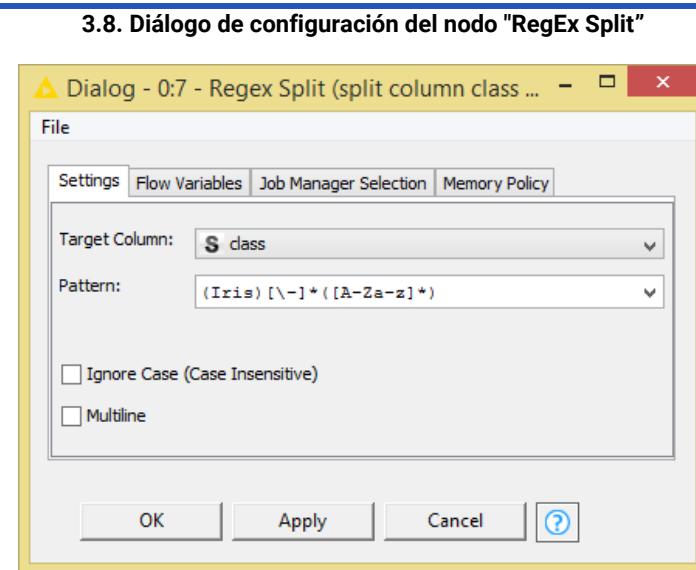
## RegEx Split (= Cell Splitter by RegEx)

Este nodo identifica las subcadenas de una columna de cadenas seleccionada a partir de una Regular Expression.

Las subcadenas se representan como expresiones regulares dentro de paréntesis. Las cadenas originales se dividen en subcadenas identificadas por dichas expresiones regulares. Cada subcadena creará una nueva columna.

La ventana de configuración requiere:

- El nombre de la columna a dividir
- Los patrones de Regular expressions para identificar las subcadenas, con las subcadenas incluidas entre paréntesis
- Algunas opciones adicionales para considerar las cadenas de varias líneas y utilizar una coincidencia sensible a las mayúsculas y minúsculas



Para separar la palabra "Iris" del resto de la cadena en la columna "class" utilizando un nodo "RegEx Split", seleccionamos:

- Column to split (Target Column) = class
- Regular Expression: `((Iris)[\\-]*([A-Za-z]*))`, lo que significa:
  - La primera subcadena entre paréntesis contiene la palabra "Iris"
  - Luego viene un "-" que no debe utilizarse como subcadena, ya que no está entre paréntesis

- La segunda subcadena puede contener cualquier carácter alfabético

El resultado son dos subcadenas denominadas "split\_0" y "split\_1", una de las cuales contiene la palabra "Iris" y la otra contiene la palabra restante "setosa", "versicolor" o "virginica".

El mismo resultado podría haberse obtenido con una Expresión Regular más general, como por ejemplo `([A-Za-z]*)([-]*(.*)$)`, que significa:

- La primera subcadena entre paréntesis contiene cualquier carácter alfabético
- Luego viene un "-" que no debe utilizarse como subcadena, ya que no está entre paréntesis
- La segunda subcadena puede contener cualquier carácter alfanumérico

Todos estos nodos "Cell Splitter" se han denominado "iris + attr", que describe la división entre la palabra "iris" y el siguiente atributo "versicolor", "virginica" o "setosa".

### **3.4. Manipulación de cadenas (String Manipulation)**

Supongamos ahora que queremos reconstruir el nombre de la clase iris pero con una estructura de cadena diferente, por ejemplo "`<attribute>:IRIS`", con la palabra IRIS toda en mayúsculas y `<attribute>` siendo "virginica", "setosa" o "versicolor". Entonces tenemos que sustituir la cadena "Iris" por "IRIS" y recombinarla con la cadena `<attribute>`. En KNIME hay muchos nodos para realizar todo tipo de manipulación de cadenas. Un nodo en particular, sin embargo, puede realizar la mayoría de las tareas de manipulación de cadenas necesarias: el nodo "String Manipulation".

## Nodo “String Manipulation”

El nodo “String Manipulation” puede realizar una serie de tareas de manipulación de cadenas, como calcular la longitud de una cadena, comparar dos cadenas, cambiar una cadena a sólo mayúsculas o minúsculas, reemplazar una subcadena o todas las ocurrencias de un carácter dentro de una cadena, poner en mayúsculas las palabras de la cadena, encontrar las posiciones de una ocurrencia de carácter o subcadena, extraer una subcadena de una cadena, etc.

La ventana de configuración del nodo “String Manipulation” es similar a la del nodo “Rule Engine” .

El “**Expression Editor**” se encuentra de nuevo en la parte central inferior de la ventana de configuración. Aquí se pueden anidar y combinar varias funciones de cadena para obtener la transformación de cadena deseada.

Las funciones de cadena disponibles se enumeran arriba en el panel “**Function List**” . Las funciones también pueden visualizarse en grupos más pequeños, seleccionando una categoría en el menú ““**Category List**” sobre el panel “**Function List**” .

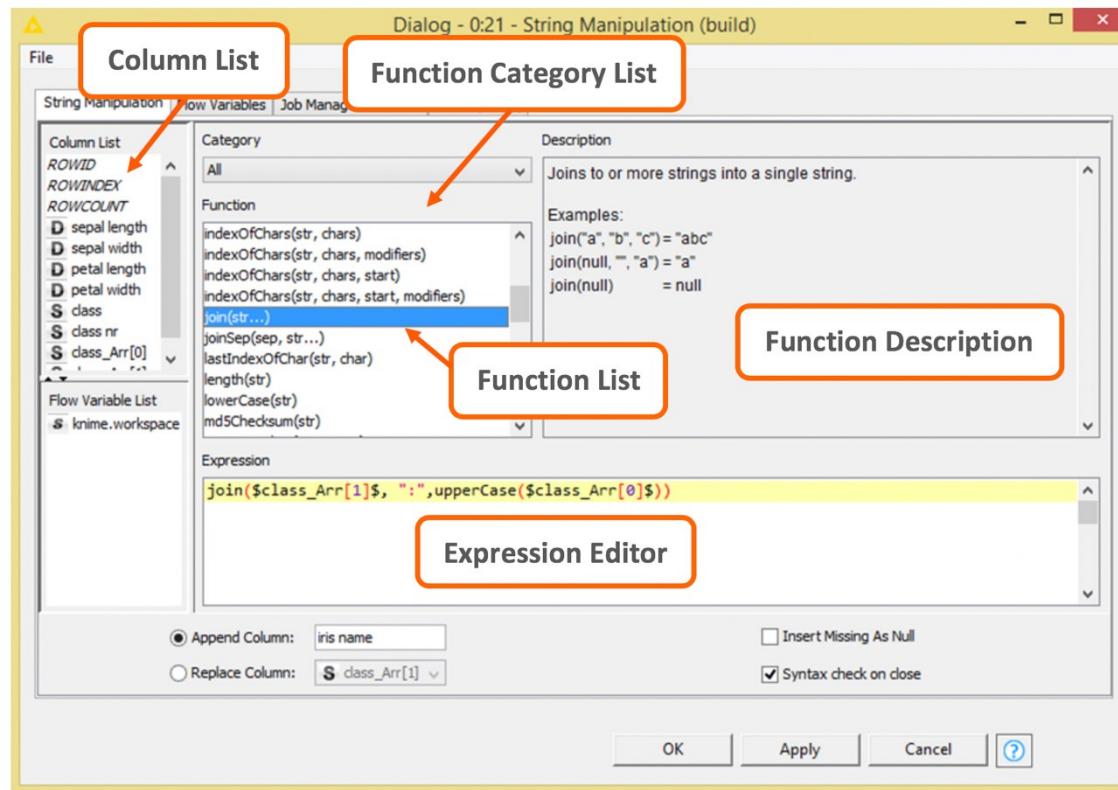
El panel “**Description**” de la derecha explica la tarea de la función seleccionada.

A la izquierda, en el panel ““**Column List**” se muestran todas las columnas de datos disponibles. Al hacer doble clic en una columna o en una función, ésta se inserta automáticamente en el “Expression Editor” con la sintaxis correcta. Los valores de cadena fija tienen que ser informados entre comillas, por ejemplo “abc”, cuando se introducen en el “Expression Editor”.

La bandera “**Insert Missing As Null**” permite la producción de una cadena nula, en lugar de una celda de datos vacía, cuando la función de manipulación de la cadena es de alguna manera infructuosa.

La ventana de configuración requiere finalmente el **nombre de la nueva columna o de la existente**, dependiendo de si la cadena resultante tiene que sobrescribir los datos existentes.

### 3.9. Diálogo de configuración del nodo "String Manipulation"



El nodo "String Manipulation" que introdujimos en el workflow "Write To DB" sigue al nodo "Cell Splitter" y combina (función "join()") la parte <attribute> del nombre de la clase en la columna class\_Arr[1] con la cadena fija ":" y con la versión en mayúsculas (función "uppercase()") de la palabra "iris". El resultado es, por ejemplo, "setosa:IRIS" para la cadena original "Iris-setosa". Observe que la división de la cadena original en las subcadenas contenidas en class\_Arr[] también podría haberse obtenido dentro del nodo String Manipulation utilizando una función substr().

**Nota.** Las funciones "toInt()", "toDouble()", "toBoolean()", "toLong()", "toNull()" convierten una cadena respectivamente en un entero, un doble, etc. Pueden utilizarse para producir una columna de salida que no sea una cadena en el puerto de salida del nodo String Manipulation

El nodo String Manipulation es especialmente útil cuando queremos combinar varias funciones de cadena diferentes en una sola más compleja. Sin embargo, un procesamiento alternativo utiliza una secuencia de nodos individuales dedicados. Este enfoque conduce a un workflow más abarrotado, pero proporciona una interpretación más fácil de todas las funciones de manipulación de cadenas utilizadas.

Para pasar de minúsculas a mayúsculas o viceversa, se puede utilizar el nodo "Case Converter" en la categoría "Manipulation" → "Column" → "Transform".

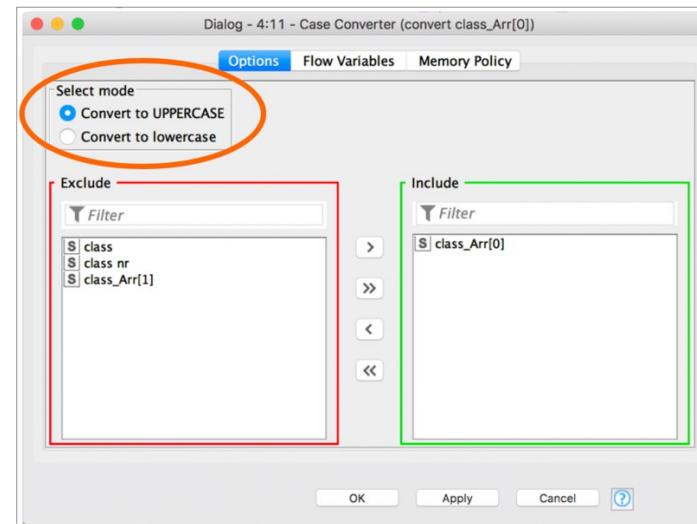
## Case Converter

Este nodo transforma los caracteres de la cadena en minúsculas o mayúsculas dependiendo de la bandera "Select mode"

La ventana de configuración requiere:

- "Select mode": "UPPERCASE" or "lowercase"
- Los nombres de las columnas a transformar. Estas columnas se enumeran en el cuadro "Include". Todas las demás columnas que no se verán afectadas por la transformación se enumeran en el cuadro "Exclude".
- Para pasar del marco "Include" al marco "Exclude" y viceversa, utilice los botones "add" y "eliminate". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "eliminate all".

3.10. Ventana de configuración del nodo "Case Converter"



Conectamos un nodo "Case Converter" al puerto de salida del nodo "Cell Splitter". Por supuesto, podríamos haber conectado el nodo "Case Converter" al puerto de salida de cualquiera de los nodos "Cell Splitter". Elegimos el nodo "Cell Splitter" sólo como ejemplo. Entonces configuramos el nodo "Case Converter" así:

- "Select mode" se fija en "Convert to UPPERCASE"
- Las columnas a cambiar son sólo "class\_Arr[0]", que es la columna que contiene la palabra "Iris", en el conjunto "Include"

Para sustituir una cadena en general, existe un nodo "String Replacer" en "Manipulation" → "Column" → "Convert & Replace".

Este nodo tiene una variante "String Replace (Dictionary)" que realiza los reemplazos de cadena basándose en un archivo de texto de diccionario previamente formateado. Este nodo puede ser útil para reemplazar múltiples cadenas y subcadenas con el mismo valor de cadena.

La función correspondiente en el nodo String Manipulation sería uppercase().

## String Replacer

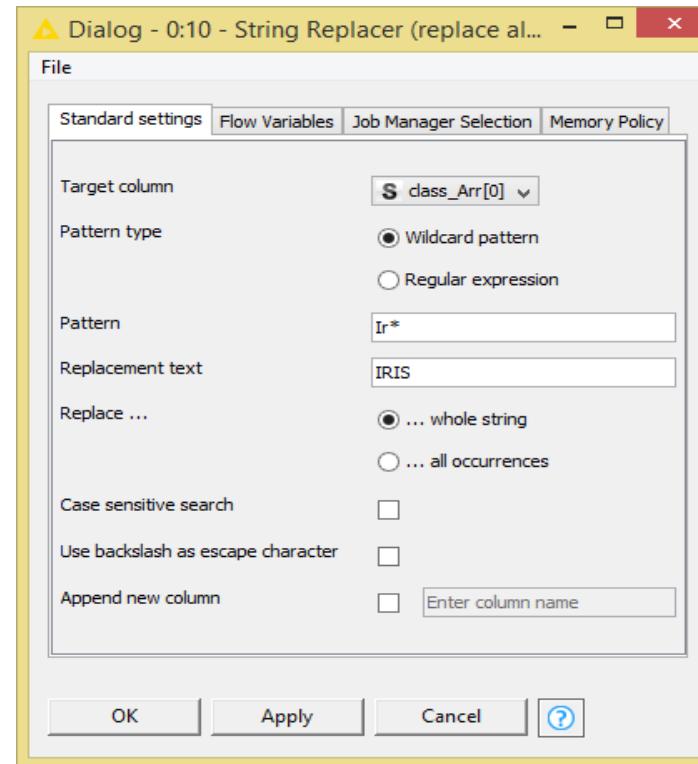
El nodo "String Replacer" sustituye un patrón en los valores de una columna de tipo cadena. La ventana de configuración requiere:

- El nombre de la columna en la que hay que reemplazar el patrón
- El patrón que debe coincidir y reemplazarse (se permiten comodines en el patrón)
- El nuevo patrón a sobreescibir el anterior

Y algunas opciones más:

- Si el patrón que debe ser comparado y reemplazado contiene comodines o es una expresión regular
- Si el texto de sustitución debe reemplazar todas las ocurrencias del patrón como cadenas aisladas o también como subcadenas
- Si la coincidencia del patrón debe distinguir entre mayúsculas y minúsculas
- Si los caracteres de escape se indican mediante una barra invertida
- Si el resultado reemplaza la columna original (por defecto) o crea una nueva columna

3.11. Ventana de configuración del nodo "String Replacer"



Para cambiar la cadena "Iris" por la cadena "IRIS", conectamos un nodo "String Replacer" al puerto de salida del nodo "Cell Splitter" y utilizamos la siguiente configuración:

- Columna objetivo "class\_Arr[0]", que contiene la cadena "Iris"
- El patrón a sustituir puede ser "Iris" o, más generalmente, "Ir\*" con el comodín "\*"
- El nuevo patrón a sobreescribir es "IRIS"

La columna resultante "class\_Arr[0]" contiene todas las cadenas "IRIS", exactamente igual que la columna generada con el nodo "Case Converter". Por último, queremos combinar todas esas subcadenas en una nueva columna de cadenas llamada "iris name" y que contenga cadenas estructuradas como "<attribute>:IRIS". Para combinar dos o más columnas de cadena, existe el nodo "Column Combiner" en "Manipulation" → "Column" → "Split & Combine".

La función correspondiente en el nodo String Manipulation sería replace().

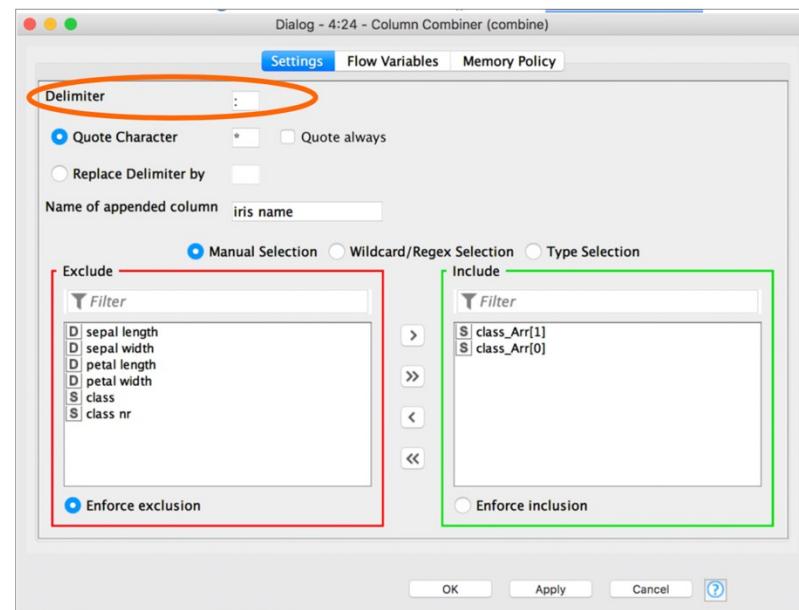
## Column Combiner

El nodo "Column Combiner" combina dos o más columnas de cadena en una sola columna de cadena, uniéndolas opcionalmente mediante un carácter delimitador. La ventana de configuración requiere:

- El carácter delimitador (si lo hay, este campo también puede estar vacío)
- Si queremos incluir las subcadenas originales entre comillas, hay que activar la bandera "Quote always" y suministrar el "Quote character"
- El nombre de la nueva columna
- Los nombres de las columnas a combinar. Estas columnas aparecen en el cuadro "Include". Todas las demás columnas que no se utilizarán para la combinación se enumeran en el cuadro "Exclude".

Para pasar del marco "Include" al marco "Exclude" y viceversa, utilice los botones "add" y "remove". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".

3.12. Ventana de configuración del nodo "Column Combiner"



Para obtener los valores finales de la cadena "<attribute:IRIS>", necesitamos un nodo "Column Combiner" con la siguiente configuración:

- El delimitador es ":"
- Las columnas a combinar en el marco "Include" son "class\_Arr[1]" y "class\_Arr[0]"
- No hay comillas alrededor de las cadenas originales, es decir, la bandera "Quote always" está desactivada
- El nombre de la nueva columna es "iris name". Observe que este nodo no tiene la opción de reemplazar una columna de entrada con los nuevos valores

La función correspondiente en el nodo String Manipulation sería join().

**Nota.** En el nodo "Column Combiner" no es posible organizar el orden de concatenación de las columnas. Las columnas se combinan siguiendo su orden en la tabla de datos de entrada.

Por ejemplo, la columna "class\_Arr[0]" viene antes de la columna "class\_Arr[1]" en la tabla de datos de entrada y, por tanto, las cadenas combinadas resultantes serán "class\_Arr[0]:class\_Arr[1]", es decir "IRIS:<attribute>", que no es exactamente lo que queríamos. Para cambiar el orden de las subcadenas, tenemos que cambiar el orden de las columnas en la tabla de datos de entrada.

Para cambiar el orden de las columnas en la tabla de datos de entrada, utilizamos un nodo "Column Resorter" situado en "Manipulation" → "Column" → "Transform".

## Nodo “Column Resorter”

El nodo “Column Resorter” cambia el orden de las columnas en la tabla de datos de entrada.

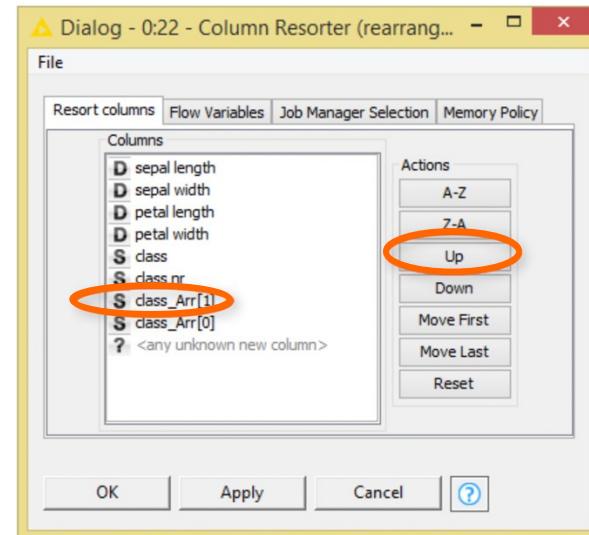
La lista de columnas de entrada con su orden (de izquierda a derecha pasa a ser de arriba a abajo) se presenta en la ventana de configuración.

Para mover una columna hacia arriba o hacia abajo, seleccione la columna en la lista y haga clic en el botón “Up” o “Down”.

Para hacer que una columna sea la primera de la lista, seleccione la columna y haga clic en “Move First”. El mismo procedimiento para hacer que una columna sea la última de la lista con el botón “Move Last”.

Para utilizar un orden alfabético en los nombres de las columnas, haga clic en el botón “A-Z” para el orden descendente y “Z-A” para el orden ascendente.

3.13. Ventana de configuración del nodo “Column Resorter”



Conectamos un nodo “Column Resorter” al puerto de salida del nodo “Case Converter”. Movimos la columna “class\_Arr[0]” una posición hacia abajo en la ventana de configuración, es decir, después de la columna “class\_Arr[1]”. Después de comentar el nodo “Column Resorter” con “rearrange column order for next node column combiner”, conectamos su puerto de salida al nodo “Column Combiner”. Ahora el “Column Combiner” tiene las columnas de entrada en el orden correcto para obtener las cadenas finales estructuradas como “<atributo>:IRIS”.

**Note.** El nodo “Column Combiner” es útil para construir claves únicas para identificar las filas de datos.

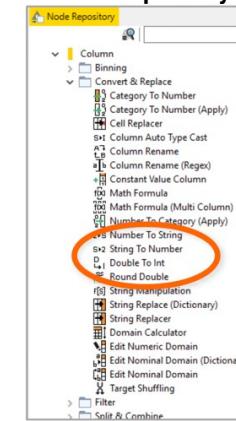
### 3.5. Conversiones de tipos de datos (Type Conversions)

En la sección anterior hemos repasado las funcionalidades de manipulación de cadenas disponibles en KNIME Analytics Platform. Antes de pasar a la sección de la base de datos, me gustaría dedicar un poco de tiempo a mostrar los nodos de "Type Conversion".

En este libro no trabajaremos con el tipo de datos Date&Time. Excluyendo este tipo de datos, hay tres nodos básicos de conversión de tipos: "Number To String", "String To Number", y "Double To Int". Todos estos nodos se encuentran en el panel "Node Repository" en "Manipulation" → "Column" → "Convert & Replace".

Para mostrar cómo funcionan estos nodos de conversión de tipos, vamos a suponer que queremos cambiar una de las columnas de datos, por ejemplo "petal width", de tipo Double a tipo String. Para ello utilizaremos un nodo "Number To String".

3.14. Ubicación del nodo "Type Conversion" en el panel "Node Repository"



# Number To String (número a cadena)

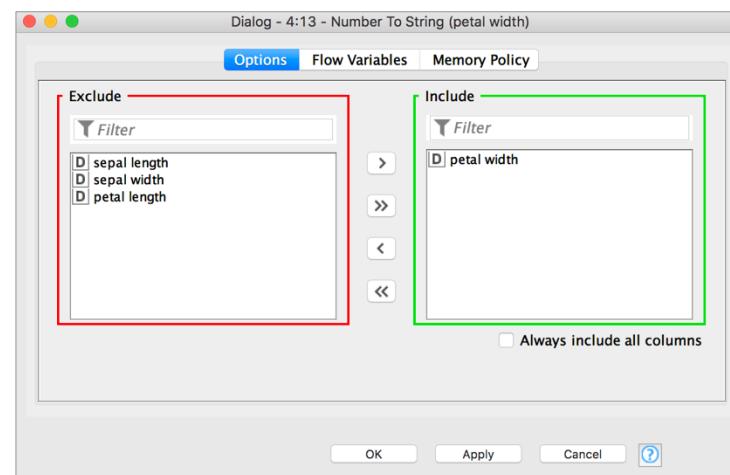
El nodo "Number To String" convierte todas las celdas de una columna de tipo "Double" o "Int" a tipo "String".

La ventana de configuración requiere:

- Los nombres de las columnas a convertir a tipo String. Estas columnas se listan en el marco "Include". Todas las demás columnas se enumeran en el marco "Exclude".
- Para pasar del marco "Include" al marco "Exclude" y viceversa, utilice los botones "add" y "remove". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".

El nodo "Number to String" es equivalente a la función `string()` del nodo String Manipulation.

3.15. Ventana de configuración del nodo "Number To String"



Insertamos sólo la columna "petal width" en el marco "Include" para convertirla de tipo Double a tipo String.

Ahora, para la demostración, supongamos que queremos aislar la parte flotante y la parte entera de la columna "petal width". Como ahora la columna "petal width" es de tipo String, utilizaremos un nodo "Cell Splitter" con carácter delimitador ". ". Llamamos a este nodo "int(petal width)". En este punto tenemos:

- La columna original "petal width"
- La primera subcadena "petal width\_Arr[0]" que contiene la parte entera del valor "petal width"
- La segunda subcadena "petal width\_Arr[1]" contiene la parte flotante del valor de "petal width".

Para convertir valores en la dirección opuesta al nodo Número a Cadena, encontramos el nodo String To Number. Para la demostración, reconvirtamos "petal width" de un tipo String a un tipo Number (Double, Long, o Int). Para ello, podemos utilizar el nodo "String To Number".

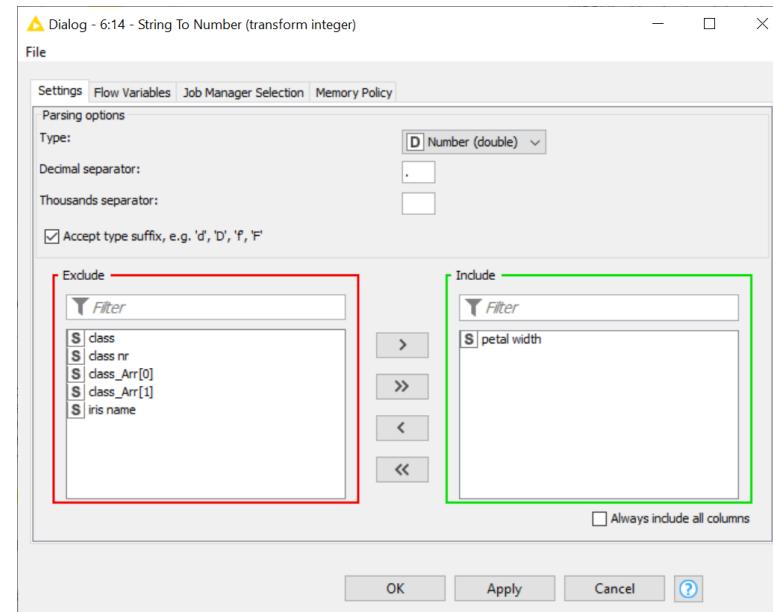
# String To Number (cadena a número)

El nodo "String To Number" convierte todas las celdas de una columna de tipo "String" a tipo "Double", "Long" o "Int". La ventana de configuración requiere:

- El tipo de columna final: Double, Long, or Int
- El separador decimal y el separador de miles (si lo hay)
- Los nombres de las columnas que deben convertirse al tipo seleccionado. Estas columnas se enumeran en el cuadro "Include". Todas las demás columnas aparecen en el cuadro "Exclude".
- Para pasar del marco "Include" al marco "Exclude" y viceversa, utilice los botones "add" y "remove". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".
- El nodo "String to Number" es equivalente a `toInt()`, `toDouble()`, `toLong()`, y funciones similares en el nodo "String Manipulation".

The "String to Number" node is equivalent to `toInt()`, `toDouble()`, `toLong()`, and similar functions in the "String Manipulation" node.

3.16. Diálogo de configuración del nodo "String To Number"



Sigamos suponiendo, para la demostración de los nodos, que hemos convertido las columnas del array "petal width" al tipo Double, pero que en realidad queríamos tenerlas de tipo Int. Ignoremos que bastaría con cambiar la opción "Type" en la ventana de configuración del nodo "String To Number" y experimentemos con un nuevo nodo: el nodo "Double To Int".

# Double To Int

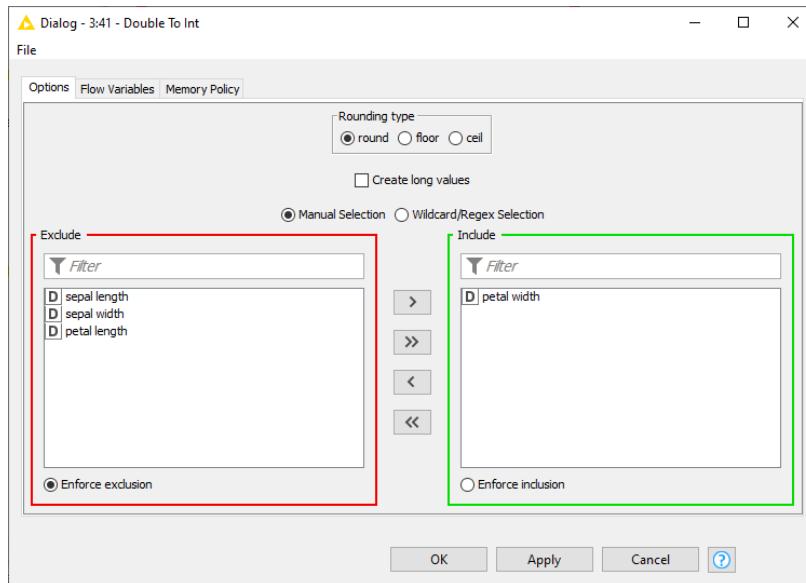
El nodo "Double To Int" convierte todas las celdas de una columna de tipo "Double" a tipo "Int". La ventana de configuración requiere:

- El tipo de redondeo: round, floor, o ceil. "round" es el redondeo estándar, "floor" redondea al siguiente entero menor, "ceil" redondea al siguiente entero mayor.
- La selección de las columnas que se van a convertir al tipo Integer. La selección puede establecerse manualmente o utilizando comodines o regex. Para ambas selecciones:

Las columnas a transformar en tipo Int se listan en el marco "Include". Todas las demás columnas se enumeran en el marco "Exclude".

Para pasar del marco "Include" al marco "Exclude" y viceversa, utilice los botones "add" y "remove". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".

3.17. Ventana de configuración del nodo "Double To Int"

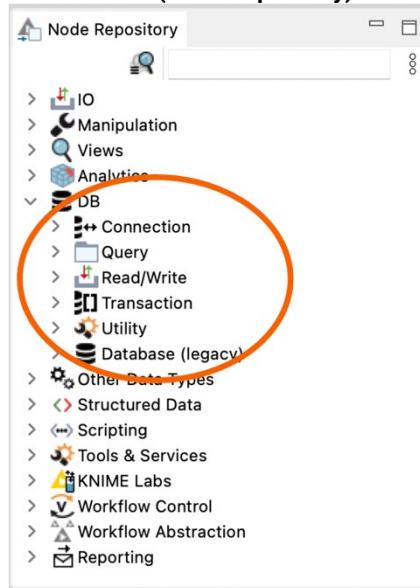


**Nota.** Un nodo que cubre muchas de las operaciones de conversión enumeradas en esta sección es el "Column Auto Type Cast". Este nodo escanea una columna para definir automáticamente su tipo de datos. También es posible una exploración rápida, más rápida pero más arriesgada.

### 3.6. Operaciones de Base de Datos (Database Operations)

Sólo hemos mostrado los nodos de conversión de tipos para ilustrar las potencialidades de KNIME. En realidad no necesitamos estas conversiones de tipo para preparar los datos para la parte de visualización. Las funciones del nodo String Manipulation habrían sido suficientes. En el próximo flujo de trabajo para la visualización utilizaremos de hecho sólo los datos producidos por el nodo String Manipulation.

3.18. Categoría "DB" en el repositorio de nodos (Node Repository)



Ahora necesitamos escribir la tabla de datos generada por el nodo de Manipulación de Cadenas en una base de datos. En el panel "Node repository" hay toda una categoría llamada "DB" que contiene todos los nodos que realizan operaciones en bases de datos.

Para acceder a una base de datos con KNIME Analytics Platform primero establecemos una conexión con la base de datos con un nodo conector y a lo largo de esa conexión utilizamos un "DB Writer" o un "DB Table Selector" seguido por un nodo "DB Reader" para escribir o leer una tabla.

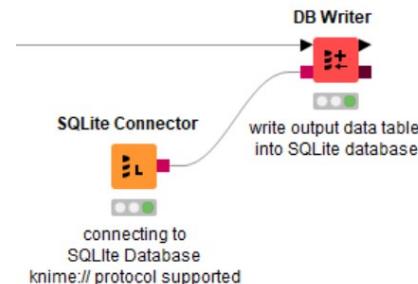
Los nodos conectores que KNIME proporciona para acceder a las bases de datos vienen con controladores JDBC precargados. Los nodos conectores cubren las versiones de bases de datos más utilizadas y más recientes, como MySQL, SQLite, Vertica, Hadoop Hive, H2, PostgreSQL, etc.

Si el nodo conector para su base de datos no está disponible, siempre puede utilizar un nodo conector de BD genérico. Aquí tendrá que proporcionar el archivo del controlador para su base de datos. Si éste no está ya en la lista de archivos de controladores precargados, siempre puedes añadirlo a través de la opción "File" → "Preferences" → "Databases" (ver más adelante en este capítulo).

Para este workflow de ejemplo utilizamos la base de datos SQLite (<https://www.sqlite.org/>). SQLite es una base de datos autónoma, sin servidor, sin configuración y basada en archivos transaccionales que no requiere autenticación. Esto facilita la distribución de los flujos de trabajo asociados a este libro, ya que no se requiere la instalación ni la configuración de una base de datos independiente. La base de datos está contenida en el archivo llamado "KBLBook.sqlite" en la carpeta KBLdata. Sólo hay que recordar que se debe seguir un procedimiento similar, incluyendo la autenticación, con una secuencia de nodos similar cuando se utilicen otras bases de datos.

Primero establecemos la conexión con la base de datos y luego, a lo largo de esta conexión, escribimos la tabla de datos en la base de datos. Para la primera tarea - establecer la conexión con una base de datos - utilizamos un nodo conector. Para la segunda tarea - escribir la tabla de datos en la base de datos - utilizamos el nodo DB Writer.

### 3.19. Node SQLite Connector + nodo DB Writer



En la categoría "DB"/"Conector" encontrará una serie de nodos conectores para establecer una conexión con una base de datos.

Algunos de esos nodos son conectores dedicados, lo que significa que contienen un archivo de controlador JDBC precargado y presentan una interfaz de usuario personalizada. Sólo el nodo "Conector DB" es un conector genérico, que se utiliza cuando el conector dedicado para la base de datos elegida no está disponible.

# SQLite Connector

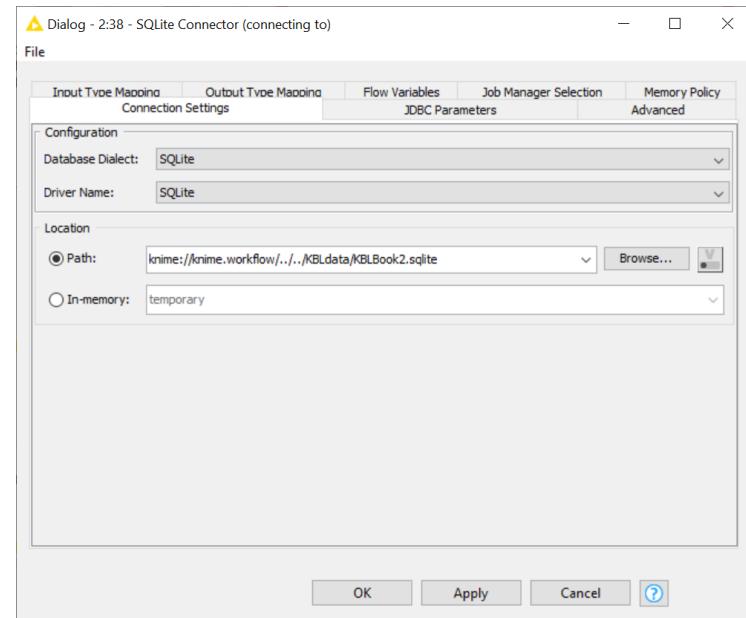
Para la base de datos SQLite, existe un nodo conector dedicado: el nodo Conector SQLite. Su ventana de configuración sólo requiere la ruta del archivo sqlite y ninguna contraseña. El controlador JDBC para la base de datos SQLite ya está precargado.

En la ventana de configuración de todos los nodos del conector, hay cinco pestañas: "Connection Settings", "JDBC Parameters", "Advanced", "Input Type Mapping", "Output Type Mapping".

"Connection Settings" contiene todos los ajustes necesarios para conectarse a la base de datos: Controlador JDBC, nombre de host, puerto, nombre de la base de datos y credenciales completas cuando sea necesario.

Las pestañas "JDBC Parameters" y "Advanced" permiten establecer comandos específicos para conectarse a la base de datos; mientras que las pestañas "Input Type Mapping" y "Output Type Mapping" permiten mapear correctamente todos los tipos de datos de KNIME a la base de datos y viceversa.

3.20. Nodo "SQLite Connector" : pestaña "Connection Settings" en la ventana de configuración



**Nota.** ¿Te has fijado en el cuadrado rojo completo como puerto de entrada? Hasta ahora, sólo hemos visto triángulos negros como puertos de entrada o salida. Un triángulo negro significa datos. Un cuadrado rojo lleno significa una conexión a la base de datos. Un cuadrado rojo vacío significa una conexión de base de datos opcional. Un cuadrado marrón lleno significa una sentencia SQL. Hay muchos tipos diferentes de puertos, cada uno de los cuales exporta o importa un tipo diferente de objeto.

La base de datos SQLite es una simple base de datos basada en archivos, que no requiere credenciales, lo que lleva a una ventana de configuración muy simple. Revisemos la ventana de configuración del nodo conector a una base de datos más compleja: el nodo conector MySQL.

# MySQL Connector

El nodo Conector MySQL se conecta a una base de datos MySQL y requiere:

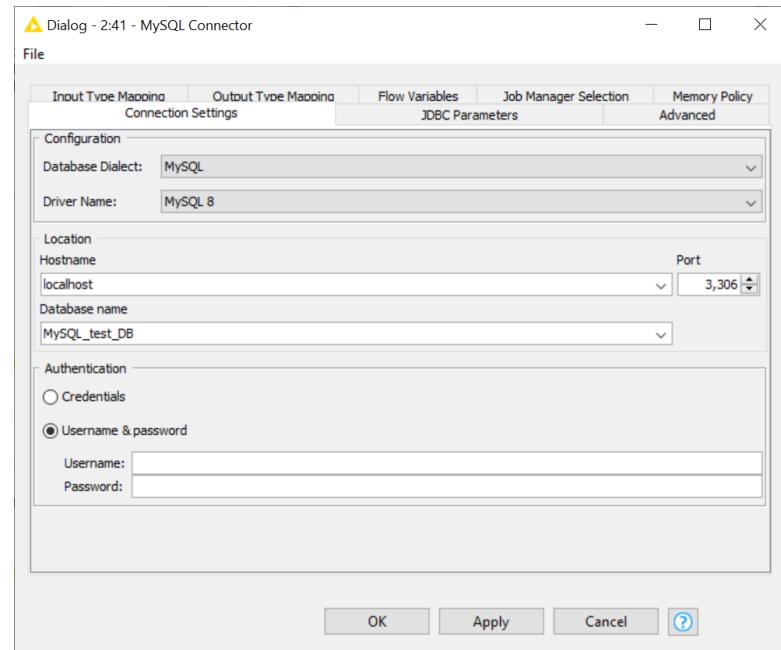
- El controlador de la base de datos (precargado)
- El nombre del host y de la base de datos
- El nombre de usuario y la contraseña para la autenticación

Las credenciales pueden ser suministradas como nombre de usuario y contraseña habilitando la opción “Username & password”.

Otra opción es definirlas como credenciales a nivel de flujo de trabajo.

Las otras pestañas: “JDBC Parameters”, “Advanced”, and “Mappings” incluye las mismas funcionalidades que el nodo SQLite Connector

3.21. Nodo “MySQL Connector” :pestaña “Connection Settings”



Las credenciales en el nivel del flujo de trabajo se cifran automáticamente y, por tanto, son más seguras. El nombre de usuario y la contraseña proporcionados directamente en la ventana de configuración no se cifran automáticamente y requieren un paso adicional para la seguridad: una clave maestra. Esta clave maestra se utilizará para cifrar los nombres de usuario y las contraseñas cuando se proporcionen en las ventanas de configuración.

# Workflow Credentials (credenciales)

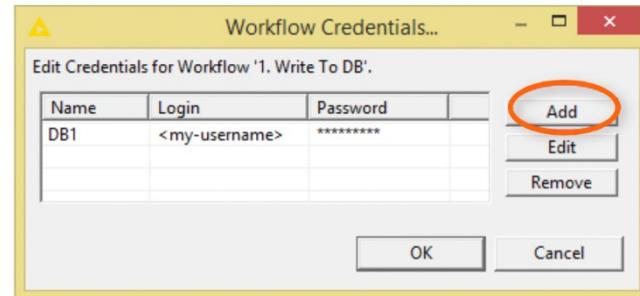
Puede establecer todos los nombres de usuario y contraseñas que necesite para su flujo de trabajo como credenciales del flujo de trabajo desde el menú contextual del flujo de trabajo.

- Haga clic con el botón derecho del ratón en el nombre del flujo de trabajo en el panel "KNIME Explorer"
- Seleccione "Workflow Credentials".

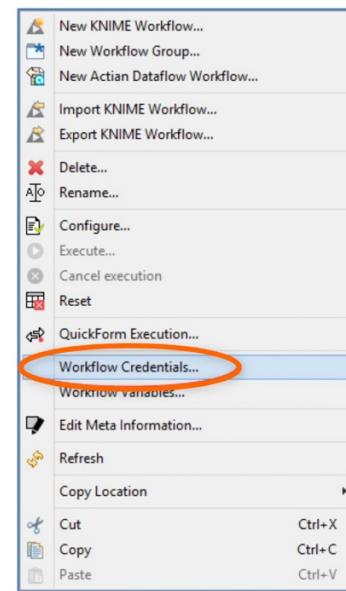
En el "Workflow Credentials ...", añadir una nueva workflow credential – es decir un nuevo par (Username, Password) – :

- Haga click en el botón "Add"
- En la pestaña "Add/Edit Credentials":
  - Establecer ID de credencial, Username (User Login), y User Password
  - haga Click n "OK"
- Introduzca tantas credenciales como necesite para su workflow
- Haga click en el botón "OK"

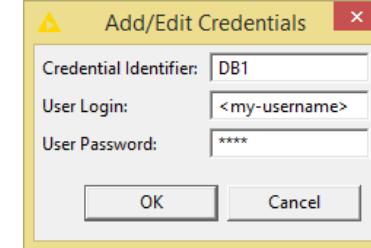
3.23. Ventana "Workflow Credentials..."



3.22. Establecer el "Workflow Credentials" desde el menu workflow's context



3.24. Ventana "Add/Edit Credentials"



**Note.** Las credenciales de Workflow credentials se encriptan automáticamente. El acceso a la base de datos a través de credenciales de flujo de trabajo es entonces más seguro y, por lo tanto, recomendado.

La lista de todos los IDs de credenciales creados está disponible en el menú de la configuración “Workflow Credentials” en la ventana de configuración de todos los nodos del conector de base de datos.

Ahora que nos hemos conectado a la base de datos SQLite a través del nodo SQLite Connector, necesitamos escribir los datos KNIME en ella con el nodo DB Writer.

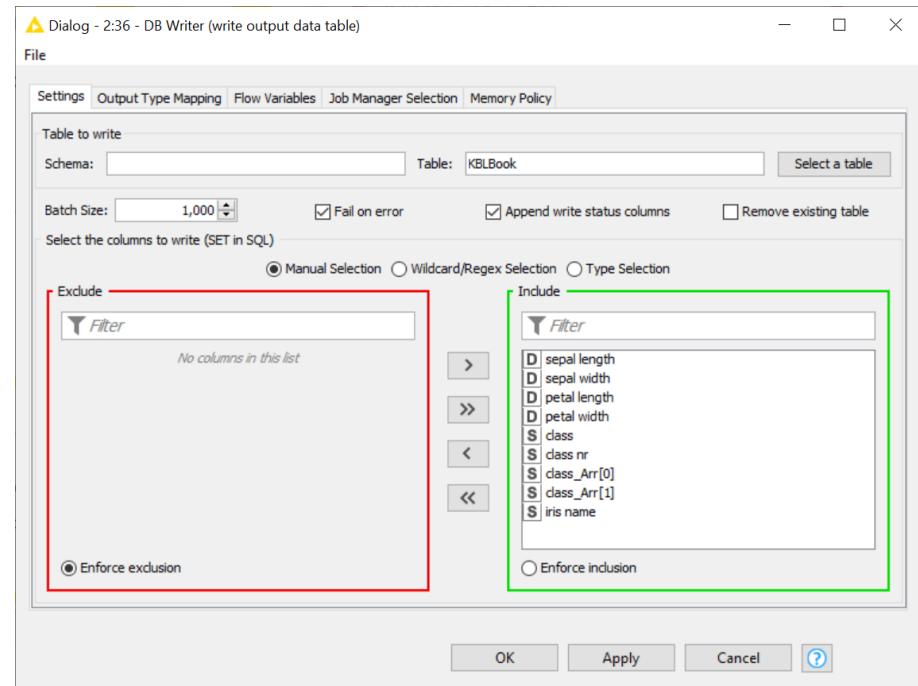
El nodo DB Writer tiene dos puertos de entrada: uno para los datos (triángulo negro) y otro para la conexión desde un nodo conector de base de datos (cuadrado rojo). Dado que el nodo DB Writer recibe toda la información de conexión del nodo conector de base de datos, como el controlador de la base de datos, el nombre del host y el puerto, sólo requiere unos pocos ajustes de configuración sobre los datos y la tabla de la base de datos a utilizar.

# DB Writer

El nodo "DB Writer", situado en la categoría "DB"/"Read/Write", escribe la tabla de datos de entrada en una tabla de la base de datos. Si la tabla no existe, se crea. Si la tabla ya existe, los nombres de las columnas en los datos KNIME tienen que coincidir exactamente con los nombres de las columnas en la tabla. Los únicos ajustes necesarios son:

- El nombre y opcionalmente el esquema de la tabla en la base de datos.
- El botón "Seleccionar una tabla" permite explorar el contenido de la base de datos.
- El tamaño del lote de datos a escribir en cada momento.
- Las columnas a transferir a la base de datos desde los datos KNIME a través de un marco de inclusión/exclusión
- El indicador de fallo en caso de error
- La alerta para añadir el estado de la operación de escritura para cada fila de datos
- La alerta "Remove existing table" permite escribir en modo "Append" o en modo "Overwrite".
- La pestaña "Output Type Mapping" contiene las especificaciones sobre el mapeo de los tipos de datos KNIME en los tipos de datos de la base de datos.

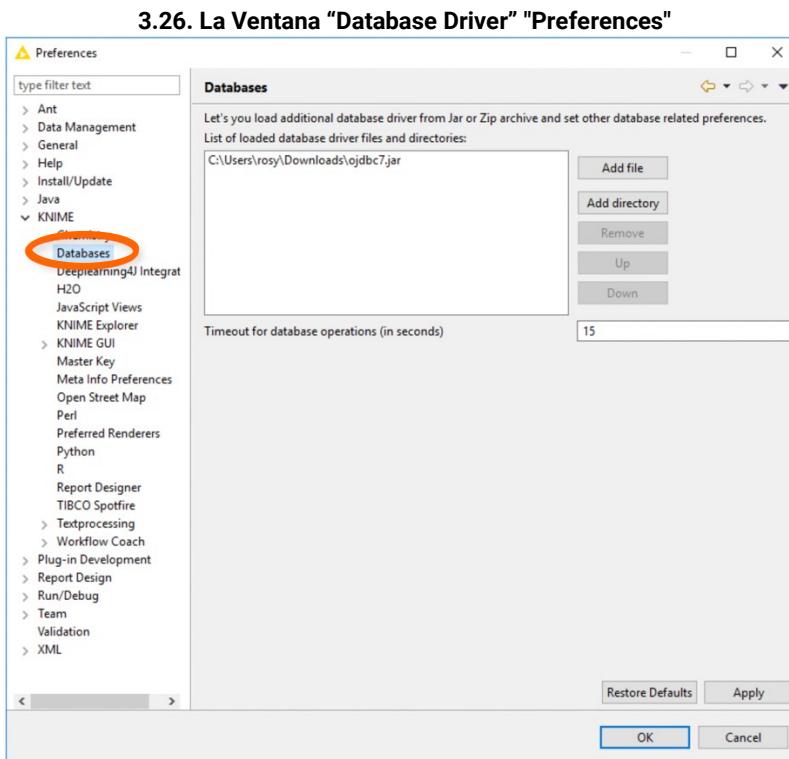
3.25. Configuración del nodo of the "DB Writer"



Para escribir los datos del iris procesados en la base de datos KBLBook.sqlite, se conectó un nodo "DB Writer" al puerto de salida del nodo "String Manipulation" llamado "build <attr>:IRIS". En la ventana de configuración del nodo "DB Writer", establecemos las columnas de datos que queremos transferir desde KNIME a la base de datos SQLite.

# Importar un controlador de base de datos JDBC (JDBC Database Driver)

Los controladores JDBC para las bases de datos más comunes y recientes ya están precargados y disponibles en los nodos del conector. Sin embargo, puede ocurrir que el controlador JDBC para una base de datos específica no esté disponible. En este caso, deberá cargar el controlador de base de datos necesario en KNIME Analytics Platform. Por lo general, el archivo del controlador JDBC (\*.jar) se encuentra en la instalación de la base de datos o puede solicitarse en el sitio del proveedor como accesoario a la instalación de la base de datos. Para cargar un controlador de base de datos en KNIME, la ubicación del archivo del controlador debe ser especificada en la ventana de "Preferencias" de KNIME.



En el Top Menu, seleccione "File" → "Preferences".  
La ventana de "Preferences" se abre.

La ventana "Preferencias" establece los valores de una serie de elementos generales, como "Ayuda", "Plug-in", "Java", etc. Todos estos elementos están agrupados en la lista de la izquierda de la ventana "Preferencias".

- Despliegue el elemento "KNIME".
- Seleccione el subapartado "Bases de datos".

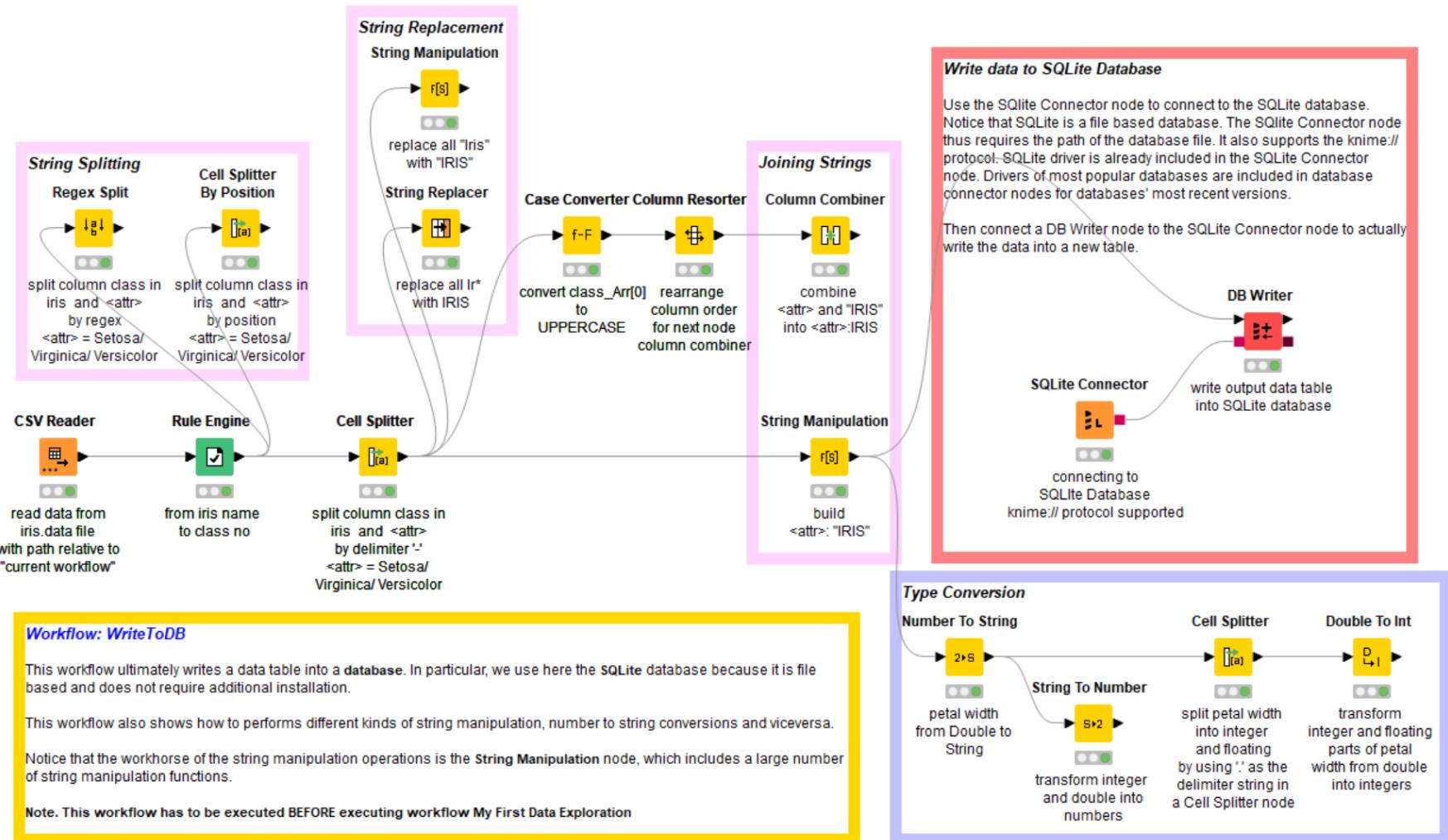
El panel de la derecha muestra la configuración de las "Bases de datos".

Para añadir un nuevo controlador de base de datos:

- Haga clic en el botón "Añadir archivo" o "Añadir directorio".
- Seleccione el archivo \*.jar o \*.zip que contiene el controlador de base de datos JDBC.
- El nuevo controlador JDBC aparece en la "Lista de archivos y directorios de controladores de bases de datos cargados" en el centro y pasa a estar disponible para todos los nodos de bases de datos.

Acabamos de completar el flujo de trabajo "Write To DB", donde hemos leído el conjunto de datos Iris y hemos realizado una serie de manipulaciones de cadenas y algunas conversiones de tipo. La tabla de datos que surge del nodo de manipulación de cadenas se ha escrito en una base de datos SQLite.

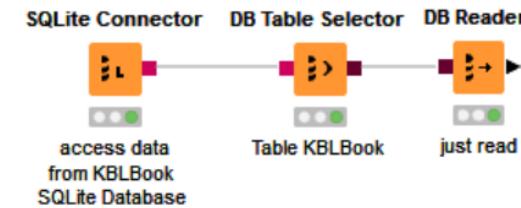
### 3.27. Workflow "Write To DB"



Ahora vamos a leer de la base de datos SQLite los datos que acabamos de escribir, para realizar una exploración visual de los datos. Junto al nodo "DB Writer" en el panel "Node Repository", encontramos el nodo "DB Reader", que utilizaremos para leer los datos de la tabla de la base de datos creada en el flujo de trabajo anterior.

Creamos un nuevo flujo de trabajo “My First Data Exploration” y establecemos la conexión con la base de datos con un nodo conector (en este caso un nodo “Conector SQLite”), seleccionamos la tabla de la base de datos de la que leer con un nodo Selector de Tabla de BD, y luego leemos los datos con un nodo “DB Reader”.

### 3.28. Nodo SQLite Connector + Nodo DB Table Selector + Nodo Database Reader para leer datos de una base de datos SQLite

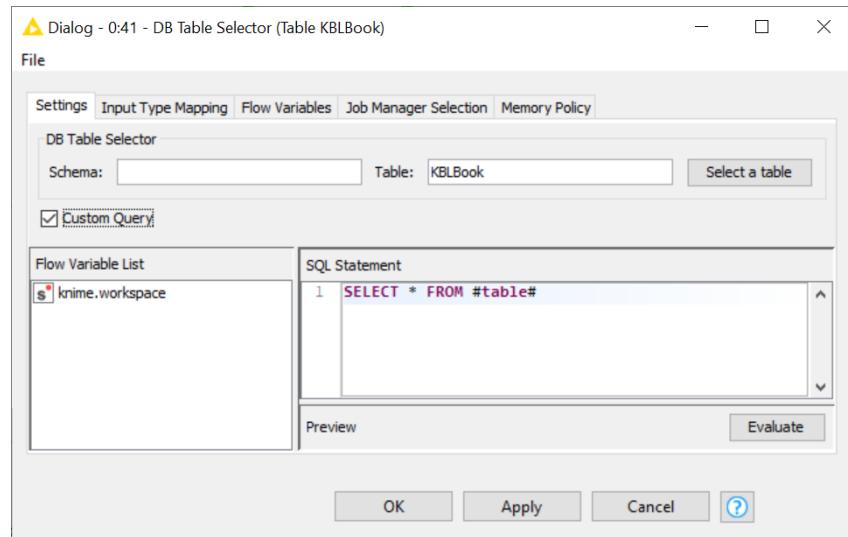


## DB Table Selector

El nodo "DB Table Selector", situado en "DB"/"Query", selecciona una tabla de la base de datos proporcionada con la conexión a la base de datos en su puerto de entrada. Se necesitan los ajustes de configuración siguientes:

- El nombre y opcionalmente el esquema de la tabla de la base de datos. El botón "Select a table" permite navegar por el contenido de la base de datos para seleccionar la tabla deseada.
- La consulta por defecto es "SELECT \* FROM #table#" donde #table# es la tabla seleccionada, que incluye todas las filas de datos y todas las columnas de datos de la tabla.
- También es posible extraer una parte o una transformación de la #table# original, creando una consulta personalizada. El editor de consultas aparece cuando se selecciona "Custom Query".
  - En el editor SQL (si está disponible) puede entonces escribir su consulta personalizada para extraer los datos de la #table#

3.29. Configuration of the "Database Writer" node when following a Database Connector node



Después del nodo SQLite Connector, el nodo DB Table Selector selecciona la tabla llamada KBLBook dentro de la base de datos utilizando la consulta por defecto. A partir de ahora trabajaremos con los datos de esta tabla.

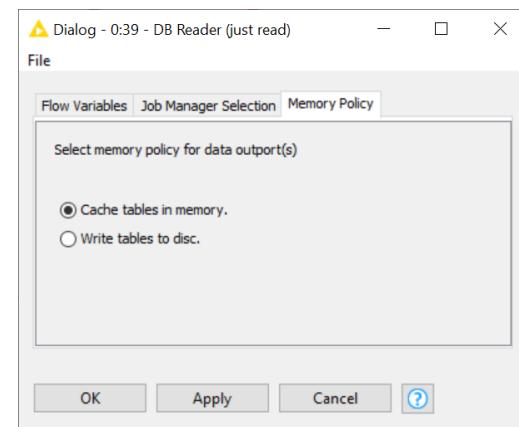
## Lector de DB (DB Reader)

El nodo "DB Reader", situado en la categoría "DB"/"Read/Write", lee una tabla de una base de datos según la consulta SQL de entrada y la importa a un flujo de trabajo.

Dado que el nodo "DB Reader" recibe toda la información (conexión a la base de datos y consulta SQL) en el puerto de entrada, no requiere ninguna configuración adicional para ejecutarse.

Así que los únicos ajustes en la ventana de configuración se refieren al almacenamiento de la tabla de datos resultante

### 3.30. Configuración del nodo "DB Reader"



El nodo lee todas las columnas de la tabla KBLBook en la base de datos KBLBook.sqlite:

- 4 columnas -- "sepal width", "sepal length", petal width", y "petal length" -- tipo de datos "Double" provienen directamente del conjunto de datos Iris
- 1 columna -- "class" -- representa la clase del iris y procede del conjunto de datos Iris
- 1 columna especifica el número de clase ("class 1", "class 2", and "class 3") y se introdujo anteriormente para mostrar cómo funciona el nodo "Rule Engine"
- Las 3 columnas restantes son subcadenas o combinaciones de subcadenas (substrings or combination of substrings) de la columna llamada "class". Se introdujeron como ejemplos de operaciones de manipulación de cadenas.

## 3.7. Agregaciones y agrupaciones (Aggregations and Binning)

Como ejemplo, investiguemos la distribución de la característica sepal\_length en todo el conjunto de datos. Vamos a aproximar esta distribución visualmente con un histograma. El histograma necesita rangos de valores (bins) en los que contar el número de ocurrencias. Por lo tanto, antes de proceder a dibujar el histograma, definimos dichos intervalos en el rango de valores de sepal\_length. Para ello, utilizamos un nodo "Numeric Binner".

Elegimos construir el histograma sólo con los valores de sepal\_length. Definimos 9 intervalos de bins: "< 0", "[0,1[", "[1,2[", "[2,3[", "[3,4[", "[4,5[", "[5,6[", "[6,7[", y ">= 7". Un corchete en el exterior del intervalo significa que el punto delimitador no pertenece al intervalo. También

decidimos crear una nueva columna para los valores de los bins. La columna que contiene los intervalos se denominó "sepal\_length\_binned".

Ahora queremos contar el número de plantas de iris para cada especie y con la medida "longitud\_de\_sépalos" cayendo en uno de los bins; es decir, queremos contar el número de plantas de iris por "sepal\_length\_binned" y por "class".

En KNIME podemos producir una agregación de valores basada en grupos y podemos reportar los valores finales de la agregación en tablas con diferente estructura utilizando dos nodos diferentes: el nodo **GroupBy** y el nodo **Pivoting**. Ambos nodos ("GroupBy" and "Pivoting")

In KNIME we can produce an aggregation of values based on groups and we can report the final aggregation values on tables with different structure by using two different nodes: the **GroupBy** node and the **Pivoting** node. Both nodes ("GroupBy" and "Pivoting") se encuentran en el panel "Node Repository" en "Manipulation" → "Row" → "Transform" category.

Ambos nodos son bastante importantes en el panorama de los nodos KNIME, ya que son bastante flexibles y permiten una serie de operaciones de agregación diferentes, desde el simple recuento de filas hasta el cálculo de medidas estadísticas, desde la correlación hasta la concatenación de valores.

Ambos nodos agrupan los datos de entrada en función de los valores de algunas columnas seleccionadas y sobre los grupos definidos calculan una serie de medidas de agregación. La única diferencia está en la forma de la tabla de datos de salida agregados. En los resultados del nodo "GroupBy" cada grupo de agregación se identifica por los valores de las primeras columnas, mientras que la columna final contiene la medida de agregación relativa a ese grupo. En la tabla resultante del nodo "Pivoting", cada celda contiene la medida de agregación para el grupo identificado por los valores en su cabecera de columna y en su RowID. Dada la importancia de ambos, utilizamos los dos.

Establecemos "sepal\_length\_binned" y "class" para identificar los diferentes grupos y utilizamos "count" como medida de agregación en la columna "sepal\_length". "count" cuenta las filas del grupo definido, es decir, todos los iris de la "clase" iris-virginica con "sepal\_length" entre 6 y 7.

# Contenedores numéricos (Numeric Binner)

El nodo "Numeric Binner" – localizado en "Node Repository" : "Manipulation" → "Column" → "Binning" category - define una serie de intervalos (es decir, bins) y asigna cada valor de columna a su bin.

La ventana de configuración requiere lo siguiente:

- La columna numérica que se va a procesar
- La lista de los intervalos de las ubicaciones
- Un indicador que señale si los valores de los recipientes deben aparecer en una nueva columna o reemplazar la columna original

Para definir un nuevo intervalo de contenedores:

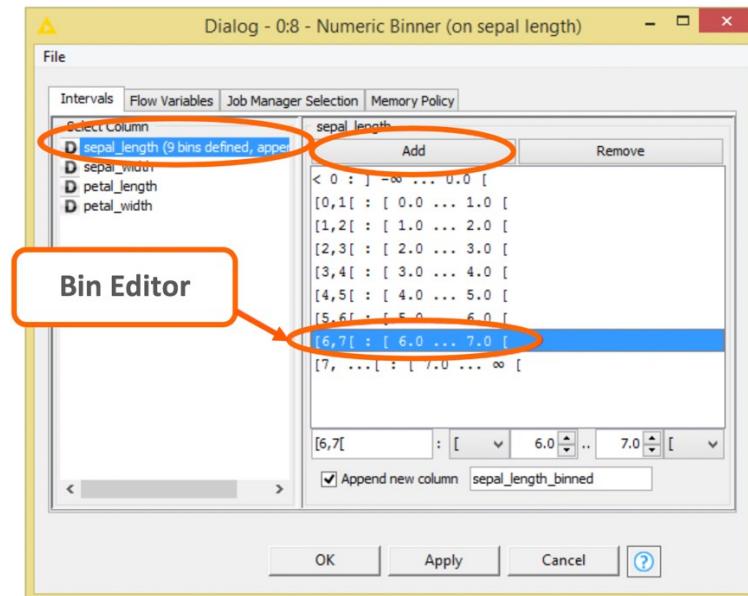
- Haga Click en "Add"
- Personalice el rango de contenedores en el Bin Editor

Para editar un nuevo intervalo de contenedores:

- Seleccione el intervalo de recipientes en la lista de intervalos de recipientes
- Personalice el intervalo de recipientes en el Editor de recipientes

Puede construir una nueva representación de recipientes seleccionando otra columna y repitiendo el procedimiento de agrupación de recipientes.

3.31. Ventana del nodo Numeric Binner"



**Note.** El método de agregación "count" sólo cuenta las filas del grupo. Es indiferente la columna que utiliza para contar las filas, si no excluimos las que tienen valores perdidos. Sin embargo, este es el único método de agregación con esta particularidad. Todos los demás métodos, como la media o la suma o la desviación estándar, producirán, por supuesto, resultados diferentes cuando se apliquen a distintas columnas.

# Agrupando (GroupBy: “Groups” tab)

El nodo "GroupBy" encuentra grupos de filas de datos utilizando la combinación de valores en una o más columnas (**Group Columns**); posteriormente agrega los valores en otras columnas (**Aggregation Columns**) a través de esos grupos. Los valores de las columnas pueden agregarse en forma de suma, media, sólo un recuento de ocurrencias, o utilizando otros métodos de agrupación (**Aggregation Method**).

La ventana de configuración del nodo "GroupBy" consta de varias pestañas. Aquí marcamos la pestaña denominada "**Groups**".

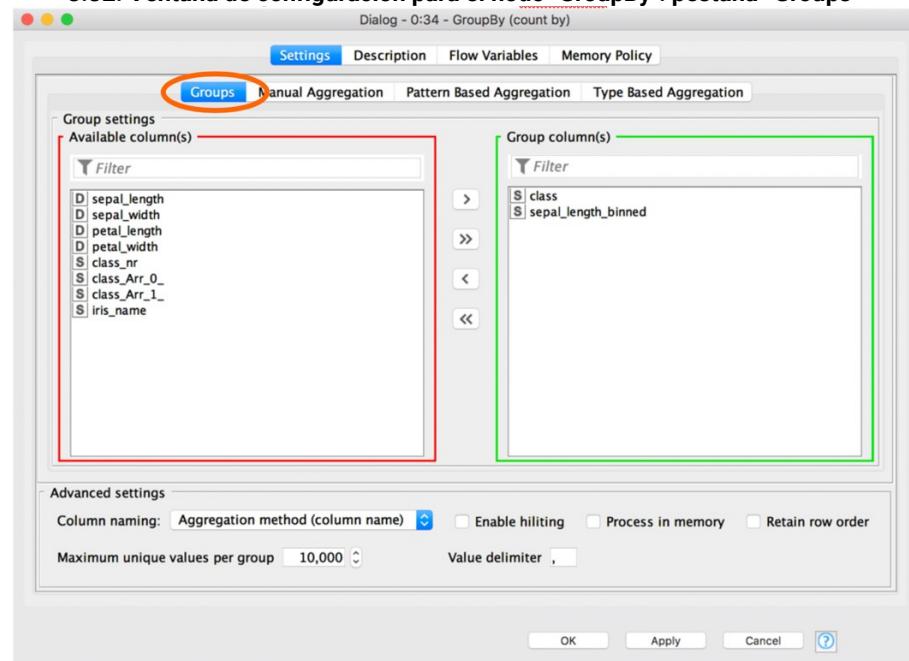
La pestaña "**Groups**". define las opciones de agrupación. Es decir, selecciona la(s) columna(s) del grupo mediante un cuadro "Excluir"/"Incluir":

- Las columnas aún disponibles para la agrupación aparecen en el cuadro "Available column(s)". Las columnas seleccionadas aparecen en el marco "Group column(s)".
- Para pasar del marco "Available column(s)" al marco "Group column(s)" y viceversa, utilice los botones "add y "remove". Para mover todas las columnas a un marco u otro, utilice los botones ""add all" y "remove all".

La parte inferior de la ventana de configuración

- establece el nombre de la nueva columna
- mantiene el orden de las filas o las resitúa en orden alfabético
- rechaza las columnas con demasiados valores distintos (por defecto 10000),
- la opción "Enable hiliting" hace referencia a una función disponible en el antiguo nodo "Data Views".

3.32. Ventana de configuración para el nodo "GroupBy": pestaña "Groups"



# GroupBy: Aggregation tabs

El resto de pestañas de la ventana de configuración definen los ajustes de agregación, es decir

- *La(s) columna(s) de agregación*
- *El método de agregación (uno para cada columna de agregación)*

Las diferentes pestañas seleccionan las columnas sobre las que realizar la agregación utilizando diferentes criterios:

- Manualmente, una por una, a través de un cuadro "Exclude"/"Include": todas las columnas seleccionadas se utilizarán para la agregación
- Basado en un patrón regex o comodín: todas las columnas cuyo nombre coincide con el patrón se utilizarán para la agregación
- Basado en el tipo de columna: todas las columnas del tipo seleccionado se utilizarán para la agregación

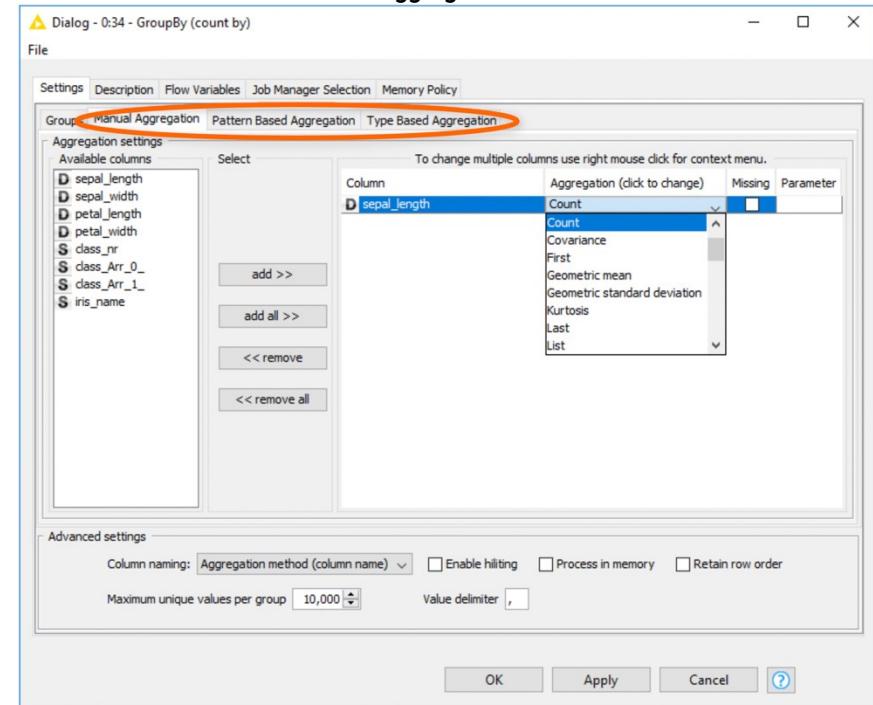
Hay varios métodos de agregación disponibles en todas las fichas de agregación. Todos los métodos de agregación disponibles se describen en detalle en el "Description" tab.

Los métodos de agregación difieren para las columnas numéricas (incluyendo aquí las medidas estadísticas, como la media, la varianza, la asimetría, la mediana, etc...) y para las columnas de cadenas (incluyendo el recuento único, por ejemplo).

Observe que los métodos de agregación "Count" y "Percent" sólo cuentan el número de filas de datos de un grupo y su valor porcentual con respecto al conjunto de datos. Esto significa que cualquiera que sea la columna de agregación asociada a estos dos métodos de agregación, los resultados no cambiarán, ya que el recuento de filas de datos de un grupo y su porcentaje no depende de la columna de agregación, sino sólo del grupo de datos.

Los métodos de agregación "First" and "Last" extraen respectivamente la primera y la última fila de datos del grupo actual.

3.33. Ventana de configuración del nodo "GroupBy": pestaña "Manual Aggregation"



Los métodos de agregación más utilizados para las columnas numéricas son: Maximum, Minimum, Mean, Sum, Variance, and Sum. Los métodos de agregación más utilizados para las columnas nominales son: Concatenate, [Unique] List, and Unique Count.

## Pivoting (Pivotando)

El nodo "Pivoting" encuentra grupos de filas de datos utilizando la combinación de valores de dos o más columnas: las columnas "**Pivot**" y las columnas "**Group**". Posteriormente, agrega los valores de un tercer grupo de columnas (**Aggregation Columns**) a través de esos grupos. Los valores de las columnas pueden agregarse en forma de suma, media, sólo un recuento de ocurrencias o una serie de otros métodos de agregación (**Aggregation Methods**).

Una vez realizada la agregación, las filas de datos se reorganizan en una matriz con los valores de la columna "Pivot" como cabeceras de columna y los valores de la columna "Group" en las primeras columnas.

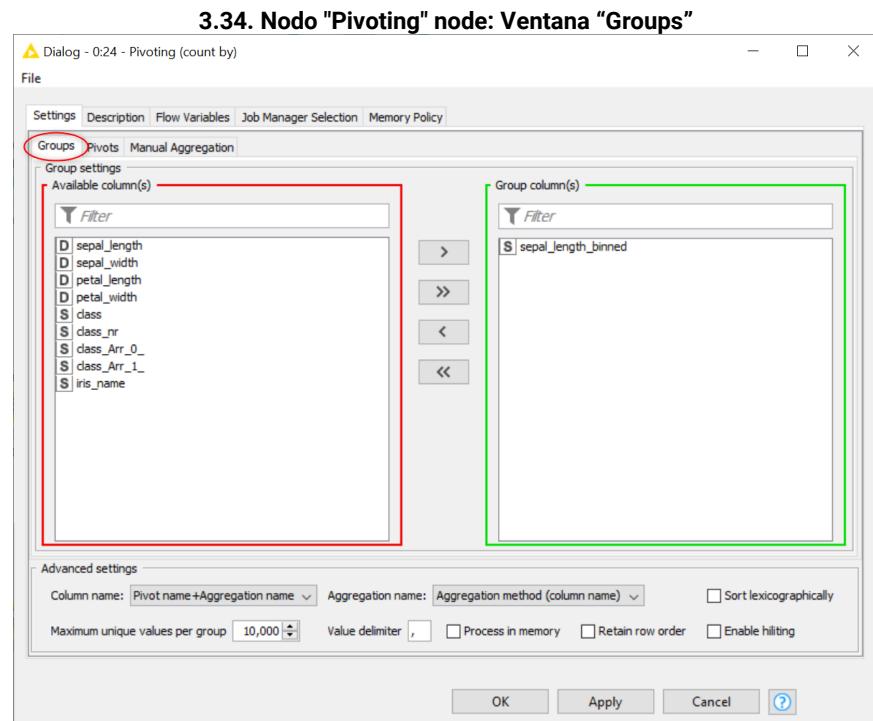
El nodo "Pivot" tiene un puerto de entrada y tres puertos de salida:

- El puerto de entrada recibe los datos
- El primer puerto de salida produce la tabla pivotante
- El segundo puerto de salida produce los totales por columna de grupo
- El tercer puerto de salida presenta los totales por columna pivotante

El nodo "Pivoting" se configura mediante tres pestañas: "**Groups**", "**Pivots**", and "**Manual Aggregation**".

La pestaña "**Grupos**" define las columnas del grupo mediante un cuadro "Exclude"/"Include":

- Las columnas aún disponibles para la agrupación aparecen en el cuadro "Available column(s)". Las columnas seleccionadas aparecen en el cuadro "Group column(s)".
- Para pasar del cuadro "Available column(s)" al cuadro "Group column(s)" y viceversa, utilice los botones "add" y "remover". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".

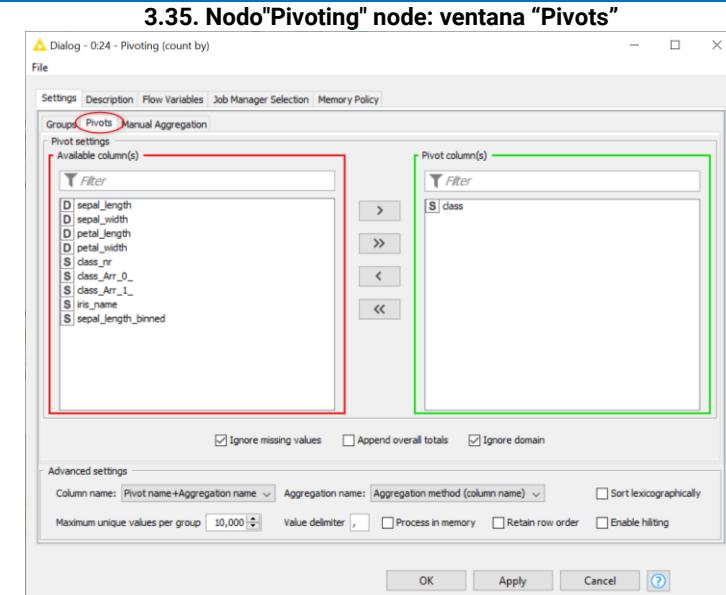


La parte inferior de la ventana de configuración

- establece el nombre de la nueva columna
- mantiene el orden de las filas o las resitúa en orden alfabético
- rechaza las columnas con demasiados valores distintos (por defecto 10000), generando así demasiados grupos distintos
- la opción "Enable hiliting" hace referencia a una función disponible en los antiguos nodos "Data Views"

Al final de esta pestaña hay tres casillas para marcar:

- “**Ignore missing values**” ignora los valores perdidos al agrupar las filas de datos
- “**Append overall totals**” añade el total general en la tabla de salida “Pivot totals”
- “**Ignore domain**” agrupa las filas de datos sobre la base de los valores reales de las celdas de grupo y pivote y no sobre la base del dominio de datos. Esto puede resultar útil cuando hay una discrepancia entre los valores reales de los datos y sus valores de dominio (por ejemplo, después de utilizar un nodo para la manipulación de cadenas).



La ventana “**Manual Aggregation**” selecciona las columnas de agregación y el método de agregación para cada columna de agregación. La selección de la columna se realiza de nuevo mediante un cuadro “Exclude”/“Include”:

Para cada columna de agregación seleccionada, hay que elegir un método de agregación. Hay varios métodos de agregación disponibles. Todos ellos se describen en la sección “Description” tab.

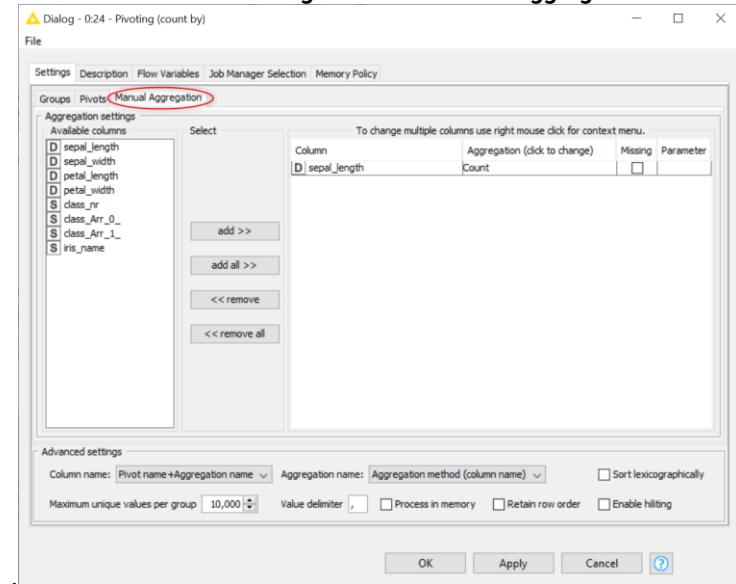
Aggregation methods “Count” and “Percent” just counts the number of data rows in a group and therefore they are independent of the associated aggregation column.

Aggregation methods sólo cuenta el número de filas de datos de un grupo y, por tanto, son independientes de la columna de agregación asociada.

Una vez realizada la agregación, las filas de datos se reorganizan en la tabla pivotante de la siguiente manera:

- Column headers = <pivot columns distinct values> + <aggregation variable name selected criterion>
- First columns = valores distintos en las columnas del grupo

3.36. Nodo “Pivoting” ventana “Manual Aggregation”



## 3.8. Nodos para Visualización de Datos (Nodes for Data Visualization)

Pasemos ahora a la parte de exploración y visualización de datos a través de las funcionalidades gráficas de KNIME Analytics Platform. Por razones históricas, hay tres formas posibles de representar gráficamente los datos en KNIME: Nodos de vistas de datos, nodos basados en JFreeChart y nodos de vistas basados en Javascript.

**Data Views** se encuentran en la categoría “Views” del “Node Repository”. Estos nodos reciben una tabla de datos como entrada y producen una representación gráfica temporal de los datos, es decir, una vista. Son los nodos gráficos más antiguos de KNIME Analytics Platform, creando una representación gráfica menos potente y menos detallada.

**JFreeChart** se encuentran en la categoría “Views”/“JFreeChart” del “Node Repository”. Estos nodos se basan en las bibliotecas gráficas Java JFreeChart. Son similares en contenido y tareas a los nodos de Vistas de Datos, pero producen una imagen estática en lugar de una vista temporal de la representación gráfica de los datos. La imagen estática se exporta al flujo de trabajo KNIME y puede utilizarse posteriormente para los informes, pero no para la exploración interactiva de la estructura de datos.

El nuevo bebé en los conjuntos de nodos de visualización de datos en la Plataforma Analítica KNIME consiste en los nodos basados en Javascript.

Estos nodos, situados en "Vistas"/"Javascript", se basan en bibliotecas gráficas de Javascript y, por tanto, permiten mejores gráficos y un mayor nivel de interacción que los anteriores nodos de Vistas de Datos. Estos nodos producen una tabla de datos y una imagen estática. La tabla de datos de salida es una copia de la tabla de datos de entrada más una columna que contiene la bandera de selección para cada punto de datos. La imagen de salida es una captura de pantalla de la vista gráfica del nodo. Se puede exportar al flujo de trabajo para la elaboración de informes u otros usos. Debido a sus mejores gráficos y mayor interactividad, en esta sección nos centraremos en los nodos basados en Javascript para la visualización de datos.

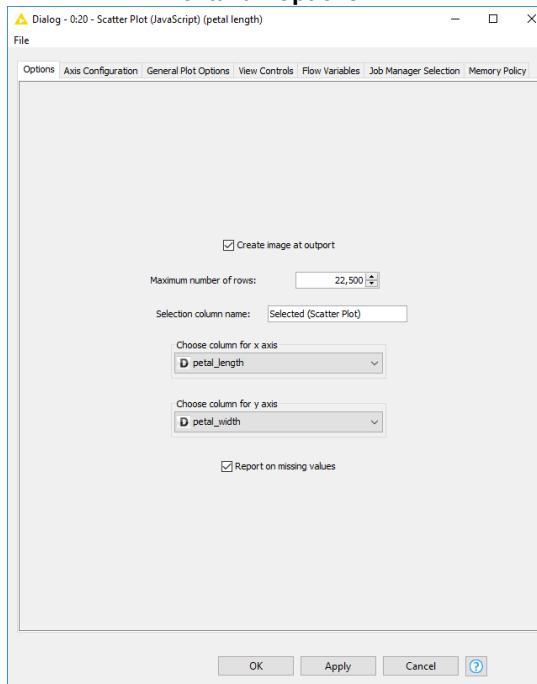
### 3.9. Gráfico de dispersión (Scatter Plot )

Comencemos nuestra exploración de datos con un clásico gráfico de dispersión. El nodo a utilizar aquí es el nodo "Scatter Plot".

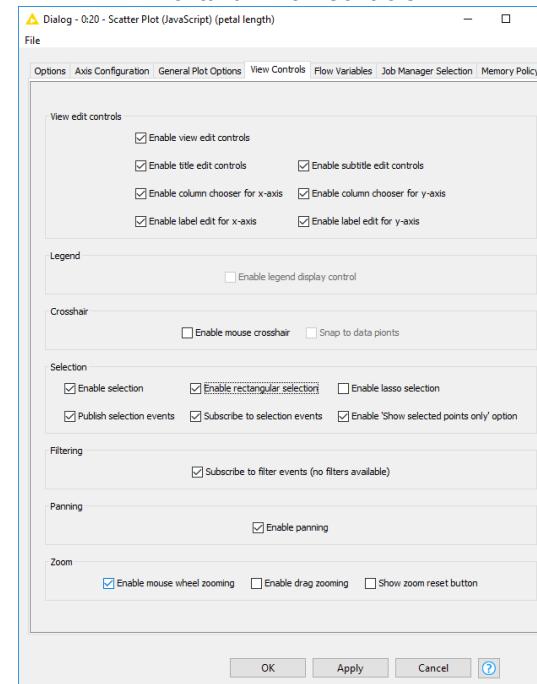
El nodo "Scatter Plot" traza cada fila de datos como un punto, utilizando dos de sus atributos como coordenadas en el eje X y en el eje Y. Después de leer el conjunto de datos del iris de la base de datos KBLBook.sqlite, queremos producir un gráfico de dispersión de la longitud de los pétalos frente a la anchura de los pétalos, que es la vista en la que se reconocen mejor los tres grupos de flores del iris.

La ventana de configuración de un nodo "Scatter Plot" abarca 4 pestañas de opciones "Options", "Axis Configuration", General Plot Options", and "View Controls". La pestaña "**Options**" define las columnas a informar en los ejes x e y, el nombre de la columna de salida para los puntos seleccionados, el criterio de emergencia para el número máximo de filas de datos a visualizar, una bandera para reproducir la vista en una imagen en el puerto de salida, y una bandera para producir una advertencia en caso de que falten valores. La pestaña """**General Plot Options**" especifica las opciones de la imagen, como el tamaño, el título, las características, los colores y el fondo. La pestaña "**Axis Configuration**" define las opciones de los ejes para la vista y la imagen, como las etiquetas, el rango y el formato. La pestaña "**View Controls**" define la interactividad permitida en la vista final, como la posibilidad de editar el título y las etiquetas, cambiar las columnas mostradas, hacer zoom y seleccionar puntos.

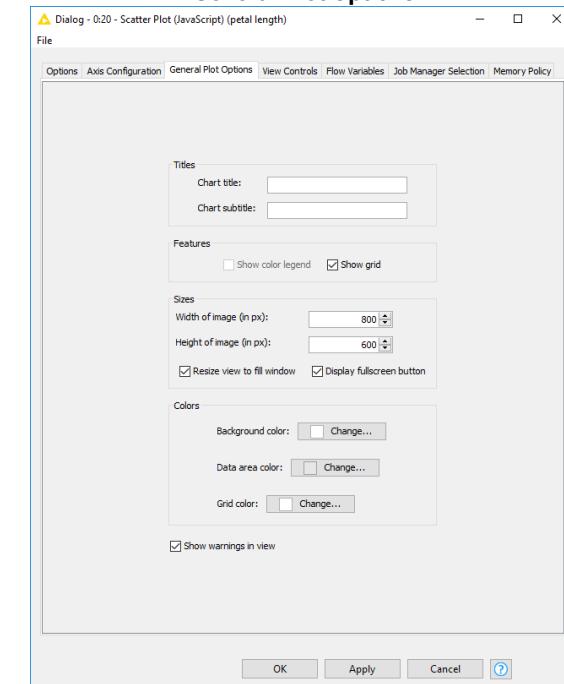
### 3.37. Configuración del nodo "Scatter Plot": Ventana "Options"



### 3.38. Configuración del nodo "Scatter Plot": Ventana "View Controls"



### 3.39. Configuración del nodo "Scatter Plot": Ventana "General Plot Options"



Después de la ejecución, el nodo produce una vista interactiva. Haga clic con el botón derecho del ratón en el nodo y seleccione "Interactive View: Scatter Plot". El nivel de interactividad de esta vista se decidió en los ajustes de la pestaña "View Controls" de la ventana de configuración del nodo. Exploraremos esta vista y veamos el tipo de interactividad que permite.

La vista del nodo "Scatter Plot" se abre utilizando los ajustes de la ventana de configuración. En nuestro caso, se abre sobre la longitud de los pétalos frente a la anchura de los pétalos, con dicha etiqueta de eje, sin título, zoom de rueda y selección simple y rectangular habilitados, como se define en la pestaña "View Controls". Dependiendo de las opciones que haya habilitado en la pestaña "View Controls" de la ventana de configuración, la vista del nodo "Scatter Plot" será más o menos interactiva.

# Gráfico de dispersión: vista interactiva

Esta es la vista del nodo "Scatter Plot", donde se pueden ver los puntos del gráfico de dispersión.

Hay tres botones en la esquina superior derecha. Son los botones de interactividad.

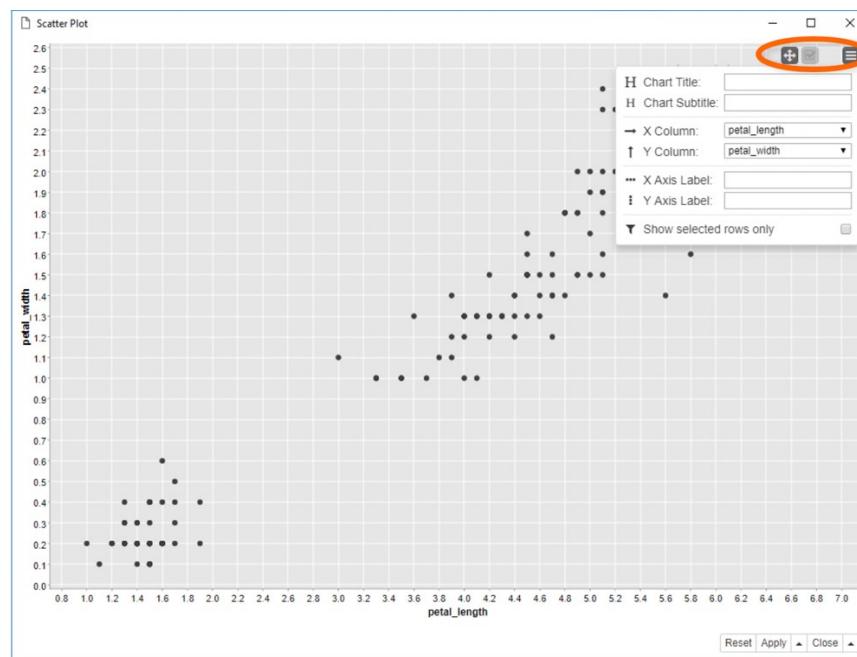
Empezando por el extremo derecho, tenemos el botón que permite cambiar la configuración del gráfico, como las etiquetas de los ejes, las columnas para el eje x y el eje y, y el título.

El segundo botón desde la derecha pone el clic del ratón en modo de selección. Cuando está activado, al hacer clic en un punto o dibujar un rectángulo en el gráfico se seleccionan los puntos correspondientes.

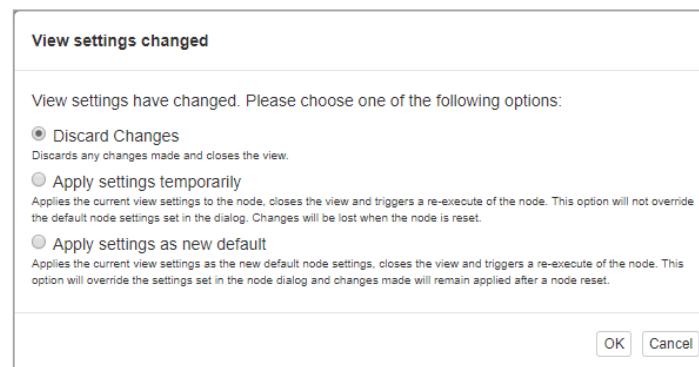
Después de seleccionar los puntos o cambiar los ajustes, si hacemos clic en el botón "Close" de la esquina inferior derecha, aparece una ventana que nos pregunta si queremos mantener los nuevos ajustes, es decir, los puntos seleccionados, de forma temporal o permanente. Con esta última opción prácticamente sobrescribimos los ajustes del nodo.

El último botón de la derecha permite hacer un paneo, es decir, hacer zoom y desplazarse por el gráfico.

3.40. Vista del nodo "Scatter Plot"



3.41. Confirmación de los cambios tras pulsar el botón "Cerrar"



**Note.** La casilla "create image at output port" en la ventana "Options" puede ralentizar la ejecución del nodo si la imagen se construye sobre un número mayor de registros de entrada. En este caso, podría considerar la posibilidad de desactivar esta casilla

Tras seleccionar algunos puntos del gráfico, cerrar la vista y aceptar los cambios, en la tabla de datos de salida la columna adicional denominada "Gráfico de dispersión seleccionado" muestra una serie de valores "verdadero" y "falso". "verdadero" se asocia a todos los registros seleccionados, "falso" a todos los demás.

**3.42. Valores "true" and "false" en la columna adicional "Selected Scatter Plot" y producida por el nodo "Scatter Plot" . "true" se asocia con los datos seleccionados. "false" indica los datos no seleccionados y es el valor por defecto**

Row ID	D sepal_l...	D sepal_...	D petal_l...	D petal_...	S class	S class_nr	S class_A...	S class_A...	S iris_name	B Sel...
Row133	6.3	2.8	5.1	1.5	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row137	6.4	3.1	5.5	1.8	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row138	6	3	4.8	1.8	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row139	6.9	3.1	5.4	2.1	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row142	5.8	2.7	5.1	1.9	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row146	6.3	2.5	5	1.9	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row147	6.5	3	5.2	2	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row149	5.9	3	5.1	1.8	Iris-virginica	class 3	Iris	virginica	virginica:IRIS	true
Row0	5.1	3.5	1.4	0.2	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row1	4.9	3	1.4	0.2	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row2	4.7	3.2	1.3	0.2	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row3	4.6	3.1	1.5	0.2	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row4	5	3.6	1.4	0.2	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row5	5.4	3.9	1.7	0.4	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row6	4.6	3.4	1.4	0.3	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row7	5	3.4	1.5	0.2	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row8	4.4	2.9	1.4	0.2	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false
Row9	4.9	3.1	1.5	0.1	Iris-setosa	class 1	Iris	setosa	setosa:IRIS	false

La interactividad es agradable, pero la vista del gráfico de dispersión parece un poco triste en su simplicidad en blanco y negro.

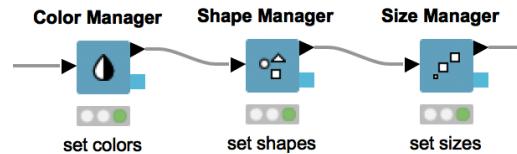
**Nota.** No es posible definir las propiedades gráficas como el color, el tamaño y la forma de los puntos a través del propio nodo del gráfico de dispersión. Se necesita un nodo gestor de propiedades, como "Color Manager", "Size Manager", or "Shape Manager".

## 3.10. Propiedades gráficas (Graphical Properties)

Los gráficos de las vistas de nodos pueden personalizarse con el color, la forma y el tamaño de los marcadores del gráfico. KNIME Analytics Platform tiene tres nodos, en "Views" → "Property" en el "Node Repository", para personalizar la apariencia del gráfico: "Color Manager", "Size Manager", and "Shape Manager". Estos nodos toman una tabla de datos como entrada y producen dos objetos en dos puertos de salida distintos.

- El primer puerto de salida contiene la misma tabla de datos del puerto de entrada, con las propiedades gráficas adicionales como el color, el tamaño y/o la forma asignados a cada fila de datos.
- El segundo puerto de salida contiene el modelo gráfico; es decir, el color, la forma o el tamaño adoptados para cada registro. Este modelo gráfico puede pasarse a un nodo "Appender" y aplicarse a otro conjunto de datos.

### 3.43. Los tres nodos de propiedades de las vistas, para establecer el color, la forma y el tamaño de los marcadores de parcela



Veamos el nodo "Color Manager" como ejemplo de cómo funcionan estos nodos de propiedades gráficas.

## Gestor de color "Color Manager"

El nodo "Gestor de colores" asigna un color a cada fila de una tabla de datos en función de su valor en una columna determinada.

Si se selecciona una columna nominal en el diálogo de configuración, se asignan colores a cada uno de los valores nominales.

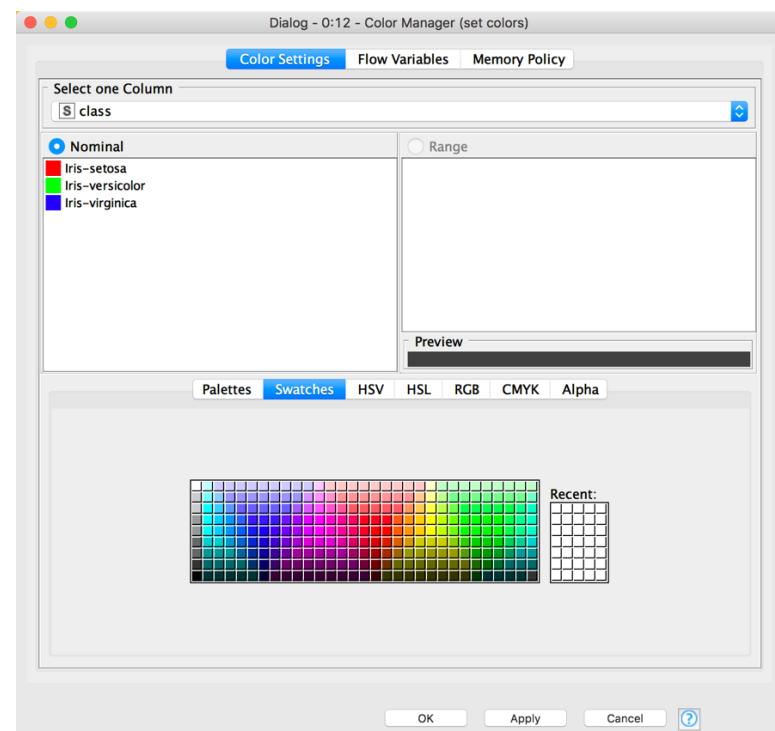
Si se selecciona una columna numérica, un mapa de calor de colores abarca el rango numérico de la columna.

La ventana de configuración requiere:

- La columna de la que extraer valores (columnas nominales) o rangos (columnas numéricas)
- El mapa de colores para cada lista de valores o rango de valores

Se asigna por defecto un mapa de colores a la lista / rango de valores. Esto puede cambiarse seleccionando el valor / rango y luego asignando un color diferente del mapa de colores que aparece en la parte inferior de la ventana de configuración.

3.44. Ventana de configuración del nodo "Color Manager"



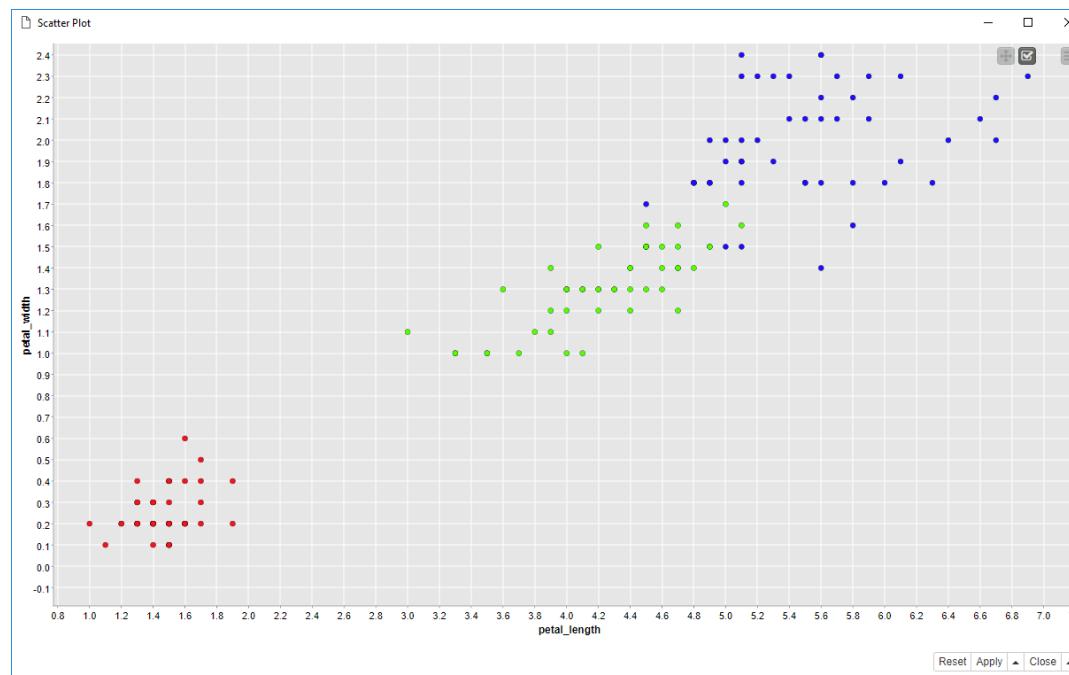
Al igual que en el nodo "Color Manager", en la ventana de configuración del nodo "Shape Manager", la forma puede cambiarse haciendo clic en la fila con el valor de columna deseado y asignando una forma de la lista del menú de la derecha.

El nodo "Gestor de tamaño", situado en el lado opuesto, utiliza un múltiplo de una columna numérica de entrada para escalar el tamaño de los marcadores de la parcela. Su ventana de configuración requiere la columna numérica y el factor a utilizar para la operación de escalado.

**Warning.** A partir de KNIME Analytics Platform 4.1, el nodo "Size Manager" y el nodo "Shape Manager" no son compatibles con los nodos de visualización basados en Javascript.

Esta vez, se aplicó un nodo "Gestor de color" a los datos originales del iris antes de alimentar el nodo del gráfico de dispersión. En la ventana de configuración seleccionamos la columna "clase" para la asignación de marcadores y asignamos colores diferentes a cada una de las tres etiquetas del iris que se encuentran en la columna "clase". La introducción de esta propiedad gráfica transforma el diagrama de dispersión -informado anteriormente en blanco y negro- en el siguiente diagrama de dispersión.

3.45. Vista en el nodo "Scatter Plot" con colores personalizados para los puntos de la parcela



### **3.11. Gráficos de Líneas (Line Plots) y Coordenadas paralelas (Parallel Coordinates)**

Otro gráfico útil es el gráfico de líneas, para dibujar series temporales y otros fenómenos en evolución a lo largo de una sola dimensión. Un gráfico de líneas conecta los valores de los atributos de forma secuencial, es decir, siguiendo su orden en la tabla de datos de entrada. La secuencia de filas representa el eje X, mientras que los valores de atributos correspondientes se trazan en el eje Y. En el gráfico pueden aparecer varias líneas, es decir, varias columnas.

Un gráfico de líneas suele desarrollarse en el tiempo, es decir, la secuencia de filas representa una secuencia temporal. Este no es el caso del conjunto de datos del iris, en el que las filas representan sólo diferentes ejemplos del iris y no tienen ninguna relación temporal. No obstante, vamos a utilizar este flujo de trabajo para mostrar cómo funciona un nodo "Line Plot".

# Line Plot

El nodo "Line Plot" muestra un gráfico de líneas, utilizando una columna como eje X y uno o más valores de columna como eje Y.

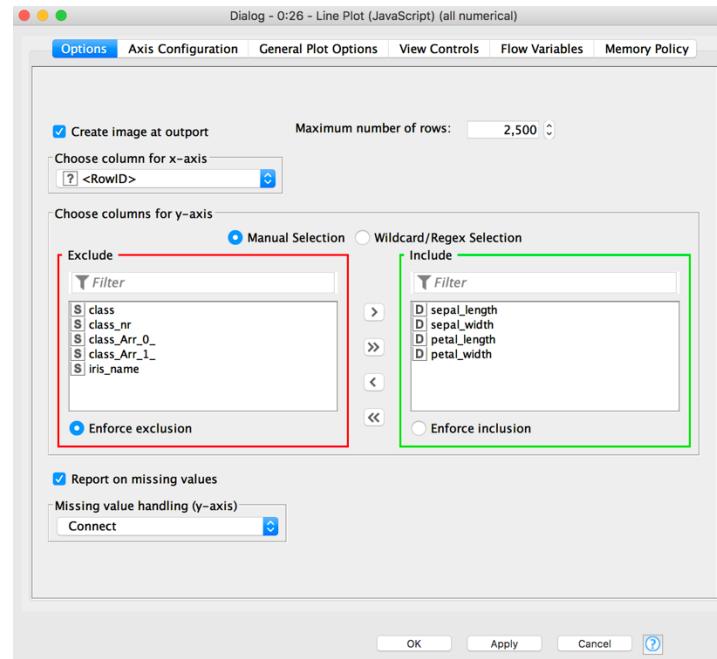
Al igual que en los anteriores nodos de visualización basados en Javascript, la ventana de configuración del nodo "Line Plot" tiene cuatro pestañas: "**Options**" para los datos; "**Axis Configuration**" y "**General Plot Options**" para los detalles del trazado; y "**View Controls**" para las características de interactividad.

La principal diferencia está en la pestaña "Options", y "Include"/"Exclude" en donde se permite seleccionar las columnas para el gráfico.

También hay disponibles varias estrategias de gestión de valores perdidos: ignorar el valor perdido y conectar los dos más cercanos; dejar un hueco vacío; o eliminar toda la columna si contiene valores perdidos.

A diferencia de otros nodos de visualización basados en JavaScript, el nodo "Line Plot" tiene un segundo puerto de entrada opcional para el esquema de colores. En este mapa de entrada los nombres de las columnas se asocian a los colores. En el gráfico final, cada columna se dibujará utilizando el color asociado en el mapa.

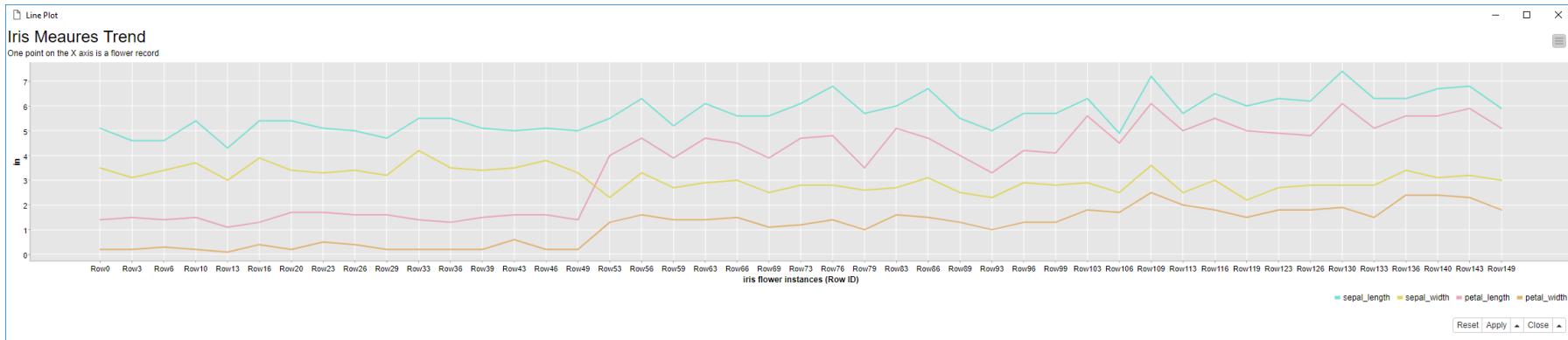
3.46. Ventana de configuración, nodo "Line Plot" node: "Options"



La vista final del nodo "Line Plot" se muestra en la siguiente figura, donde los RowIDs se muestran en el eje X y las medidas del iris se muestran en el eje Y. Aquí hemos utilizado RowIDs para el eje X, pero podríamos haber utilizado cualquier otra columna para ello. Sea cual sea la columna elegida para el eje X, el gráfico seguiría dibujando los valores de las columnas en secuencia, por orden de aparición en la tabla de datos de entrada.

**Warning.** A partir de KNIME Analytics Platform 4.1, el nodo "Line Plot" no permite mucha interactividad.

### 3.47. Vista del gráfico del nodo “Line Plot”



Otro gráfico interesante es el de coordenadas paralelas. Los gráficos de visualización de coordenadas paralelas son útiles para tener una idea de los grupos de patrones a través de las columnas. Por ejemplo, para nuestro conjunto de datos del iris, podemos ver que una de las clases de iris se separa fácilmente de las otras dos a lo largo de las coordenadas de “petal length” and “petal width”.

En el gráfico de coordenadas paralelas, una columna es una coordenada, es decir, un eje Y. Los valores de varias columnas pueden visualizarse en varias coordenadas, es decir, en varios ejes Y. La disposición de los datos a lo largo de cada eje puede contarnos algunas historias sobre los grupos del conjunto de datos. El nodo que produce un gráfico de coordenadas paralelas es el nodo “Parallel Coordinates”.

# Coordenadas paralelas (Parallel Coordinates)

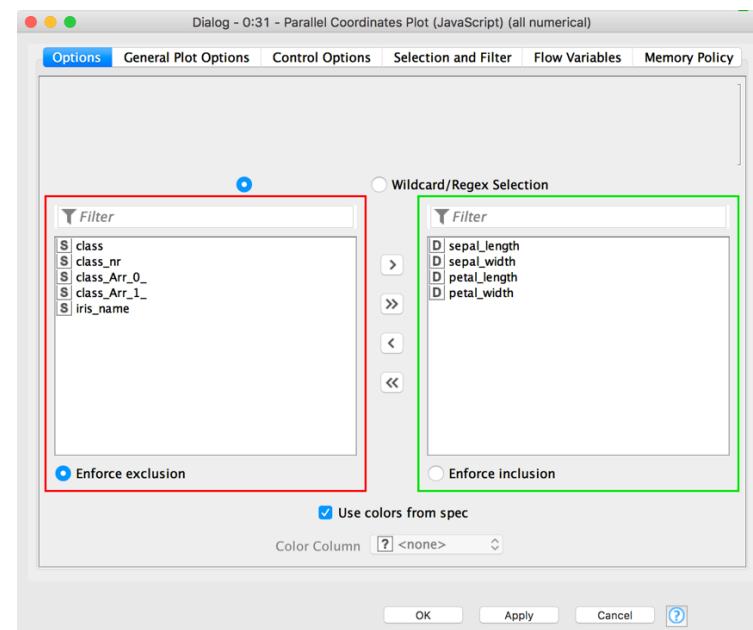
El nodo "Coordenadas paralelas" muestra la tabla de datos de entrada en un gráfico de coordenadas paralelas. Un gráfico de coordenadas paralelas despliega los nombres de las columnas a lo largo del eje X y muestra cada valor de la columna en un eje Y separado. Como resultado, un punto de datos se mapea como una línea que conecta los valores a través de los atributos.

La ventana de configuración de este nodo tiene tres pestañas.

- "**Options**" contiene un marco de "Exclude/Include" para insertar/remover mas columnas(por ejemplo. Y-axis) en/desde el gráfico de coordenadas paralelas.
- "**General Plot Options**" La pestaña define los ajustes generales para el trazado y la imagen de salida
- "**Control Options**" establece el nivel de interactividad de la vista final

Los colores de las líneas pueden proceder de una columna específica que contenga el color como propiedad gráfica (que es el resultado del nodo "Extraer color") o simplemente de la propiedad gráfica asociada a cada fila (flag "use color from spec").

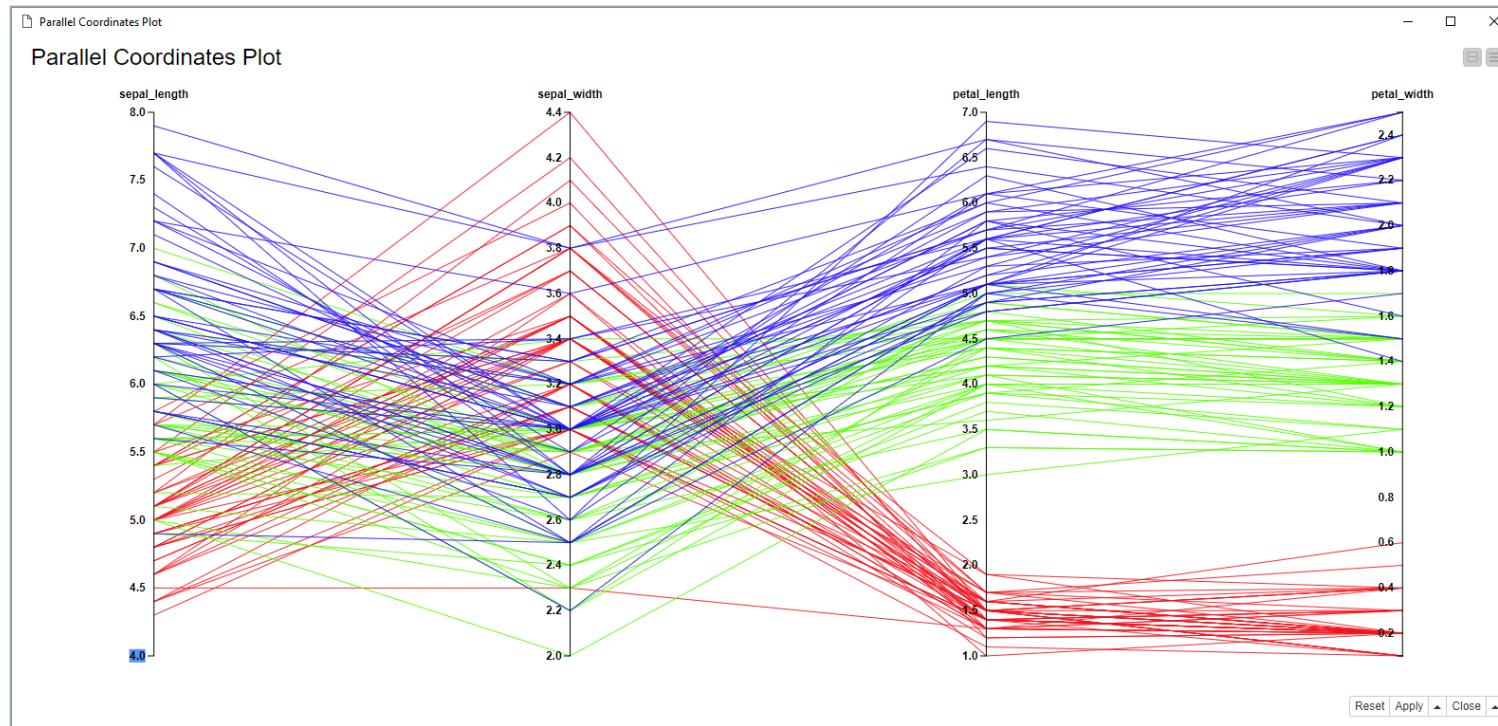
3.48. Vista del nodo "Parallel Coordinates"



A continuación se muestra la vista del nodo "Coordenadas paralelas". Como eje Y encontramos: sepal\_length, sepal\_wodth, petal\_length, petal\_width. Cada planta de iris se describe entonces por la línea que une su sepal length, sepal\_width, petal\_length, and petal\_width . Los colores de las líneas están determinados por el color asociado a cada fila de datos -es decir, a cada planta del iris- por el nodo precedente "Color Manager".

La interactividad en el nodo "Coordenadas paralelas" también es reducida con respecto, por ejemplo, al nodo "Diagrama de dispersión".

3.49. Vista del gráfico de "Coordenadas paralelas", en el que las 4 medidas del iris se muestran en los 4 ejes Y. Una línea corresponde a una planta de



## 3.12. Gráficos de Barras e Histogramas (Bar Charts and Histograms)

De todos los gráficos disponibles para investigar visualmente la estructura de los datos, no podemos dejar de lado el histograma. El histograma visualiza la frecuencia con la que se encuentran los valores en un rango determinado (bin) en la serie de valores. En esta sección se analizan brevemente los histogramas y los gráficos de barras.

Propiamente hablando, no hay un nodo basado en Javascript dedicado a dibujar un gráfico de histograma. La funcionalidad de dibujar histogramas está oculta en el nodo "Gráfico de barras".

Ya hemos colocado el "sepal\_length" en 9 contenedores. Ahora, cada fila de datos de la tabla de datos de entrada se asigna a un bin determinado según el valor de su "sepal\_length". Para construir el histograma del atributo "sepal\_length", basta con contar el número de ocurrencias en cada intervalo "sepal\_length\_binned" con un nodo "Pivoting".

# Bar Chart

El nodo "Gráfico de barras" crea un gráfico de barras genérico. Para ello, necesita

- Una columna de categoría, que en el caso de un histograma es la columna de la jerarquía.
- Una columna de agregación y un método de agregación.

En el caso de un histograma, el método de agregación es el "Recuento de ocurrencias", que sólo cuenta las filas de datos que caen en cada casilla y, por lo tanto, no requiere una columna de agregación específica.

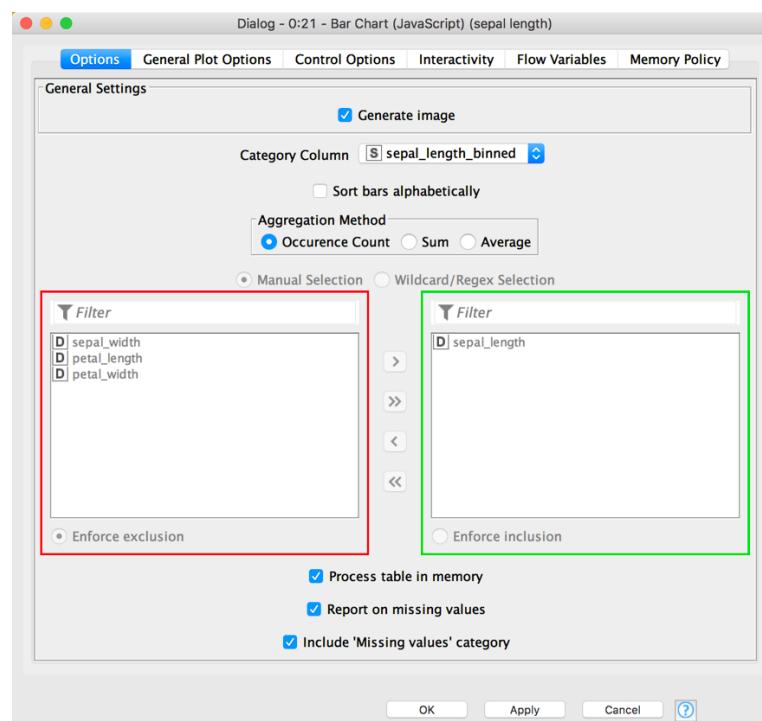
Todos estos ajustes se definen en la pestaña "Opciones" de la ventana de configuración. Dos pestañas adicionales "Opciones generales de ploteo" y "Opciones de control" definen respectivamente los detalles gráficos del ploteo y los controles de vista habilitados.

La pestaña "Opciones generales de ploteo" incluye las preferencias para el título, las etiquetas de los ejes, la orientación del ploteo, la leyenda y el tamaño de la imagen de salida.

La pestaña "Opciones de control" incluye el zoom, el cambio de orientación del gráfico, la edición de títulos y etiquetas, el apilamiento/agrupamiento de barras y el apilamiento de etiquetas.

El nodo "Gráfico de barras" tiene un puerto de entrada opcional para un mapa de colores.

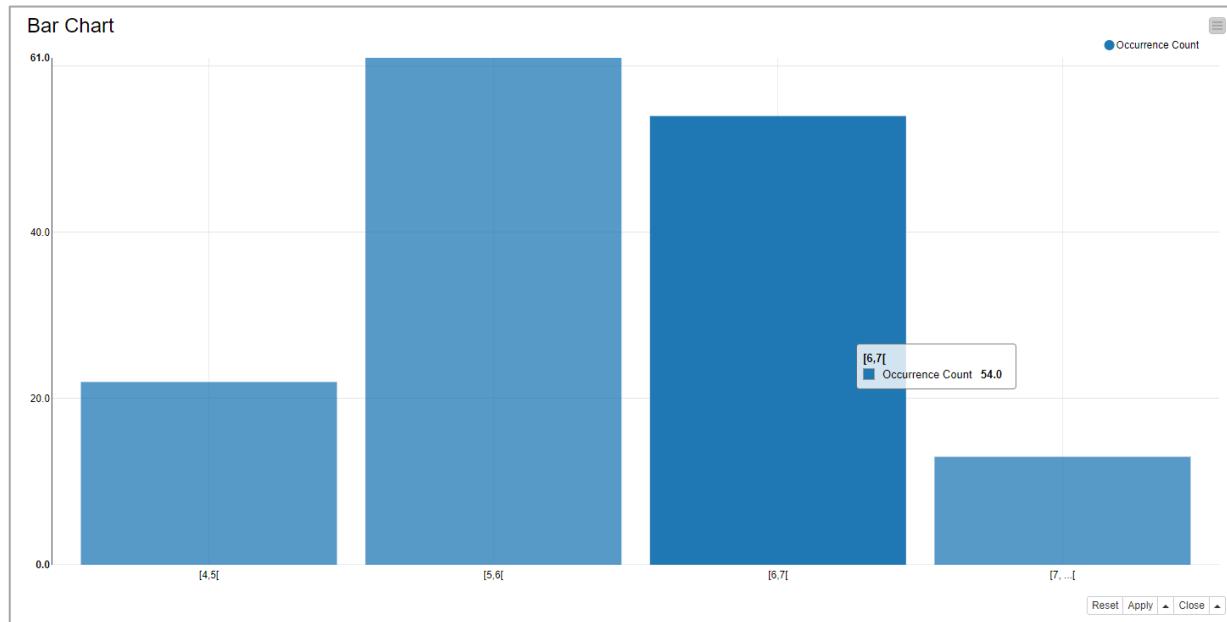
3.50. Ventana de configuración del nodo "Bar Chart": Pestaña "Opciones" configurada para dibujar un histograma



La vista "Histograma" muestra el número de veces que los valores de una determinada columna aparecen en un intervalo determinado (bin). La vista final del histograma se muestra a continuación.

**Nota.** El nodo "Gráfico de barras" no ordena las categorías de cadenas en el eje X. Se muestran en orden de aparición. Si queremos que se ordenen, como en nuestro caso de intervalos de binning, un nodo "Sorter" debe preceder al nodo "Bar Chart".

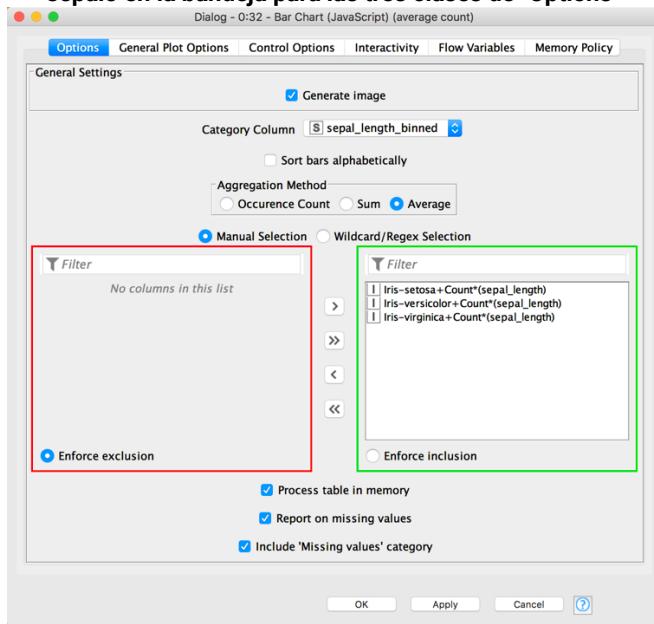
**3.51. Vista del Histograma de la longitud de los sépalos obtenido con un nodo "Bar Chart" con "Occurrence Count" como método de agregación**



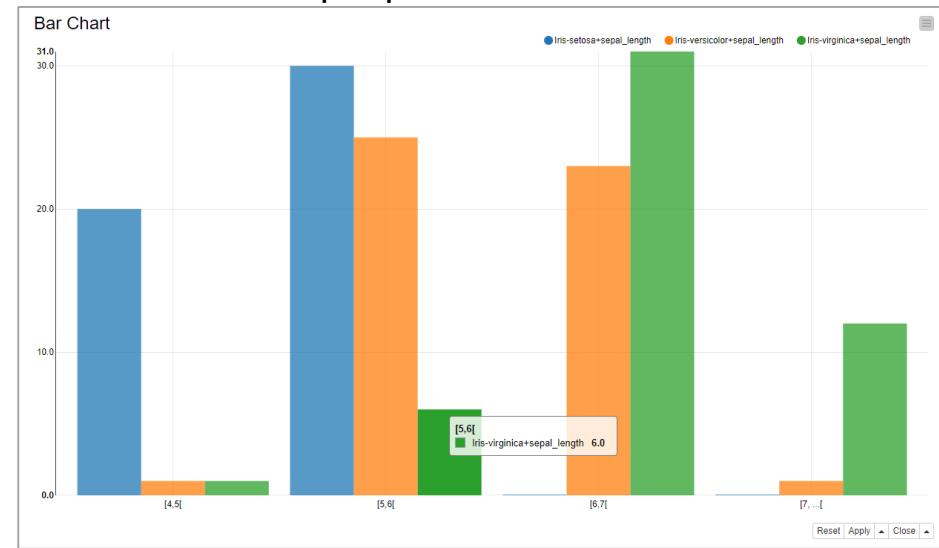
Este histograma abarca todos los casos de plantas de iris representados en el conjunto de datos de entrada. Sin embargo, supongamos que queremos aislar y comparar el mismo histograma para las tres clases separadas: iris-setosa, iris-versicolor e iris-virginica.

En primer lugar, tenemos que separar los tres grupos y contar el número de ocurrencias para cada grupo y para cada casilla en `sepal_length` (nodo "Pivoting"); finalmente, tenemos que dibujar los recuentos en un gráfico de barras (nodo "Bar Chart" con el método de agregación "Average" en las tres clases). La ventana de configuración del nodo "Gráfico de barras" y la consiguiente vista de los histogramas se muestran a continuación.

**3.52. Ventana de configuración del nodo "Bar Chart": Pestaña "Options"**  
**con el método de agregación "Average" en el recuento de la longitud del sépalo en la bandeja para las tres clases de "Options"**



**3.53. Vista del nodo "Bar Chart" que muestra los histogramas de longitud de los sépalos para las tres clases de iris**



El último nodo que nos gustaría considerar en esta sección es el nodo "Table View". Este nodo sólo muestra los datos de entrada en una tabla.

## Nodo “Table View”

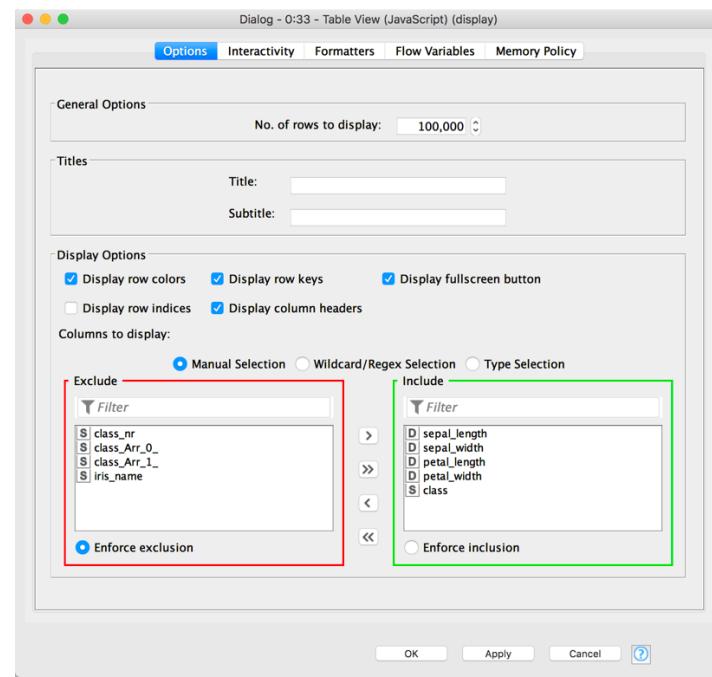
El nodo "table view" muestra los datos de entrada en una tabla.

La ventana de configuración consta de tres pestañas:

- La pestaña "Options" define la selección de datos, por ejemplo, las columnas que se mostrarán
- La pestaña "Interactivity" contiene los ajustes habituales para determinar el nivel de interactividad en la vista producida
- La pestaña "Formatters" proporciona algunas opciones de formato para números, cadenas y fechas.

Dependiendo de la configuración de la pestaña "Interactivity", las filas de la vista de tabla presentan un cuadro de selección a la izquierda. De este modo, es posible seleccionar sólo algunas de ellas. Las filas seleccionadas mostrarán la bandera "true" en la columna anexa "Selected Javascript Table View" en el puerto de salida del nodo.

3.54. Vista de la tabla del nodo "Interactive Table"

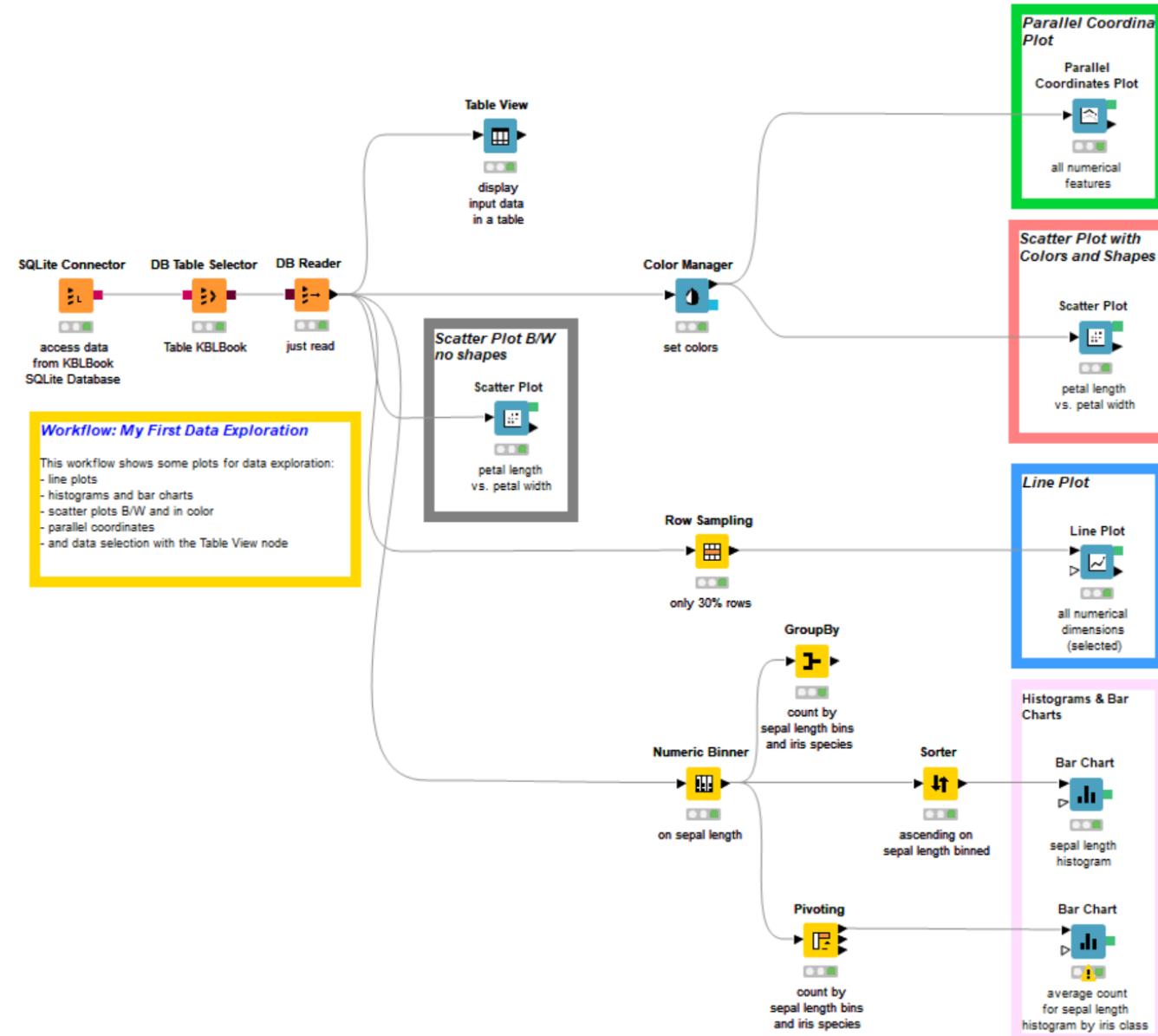


Aquí terminamos nuestra descripción de los nodos disponibles en KNIME Analytics Platform para la visualización de datos. Hay algunos nodos adicionales de visualización interesantes, como "Lift Chart", "Box Plot", "ROC Curve", "Pie Chart", etc ....

En particular, el nodo "Vista genérica de Javascript" permite utilizar código Javascript libre. Si usted es un experto en Javascript y/o prefiere utilizar algunas librerías específicas de Javascript, este es el nodo que permite crear gráficos basados en Javascript arbitrariamente complejos.

Este es el flujo de trabajo final "My First Data Exploration".

### 3.55. Versión final del workflow "My First Data Exploration"



## 3.13. Ejercicios

### Ejercicio 1

Lea el archivo "yellow-small.data" del conjunto de datos de los globos (puede encontrar este archivo en la carpeta KBLdata o puede descargarlo de: <http://archive.ics.uci.edu/ml/datasets.html>).

Este archivo tiene 5 columnas: "Color", "Size", "Act", "Age", and "Inflated. Cambie el nombre de las columnas.

Añada la siguiente columna de clasificación y nómbrela "class":

```
IF Color = yellow AND Size = Small => class =inflated  
ELSE                                     class = not inflated
```

Añade una columna final llamada "Final sentence" que diga:

"inflated is T"

OR

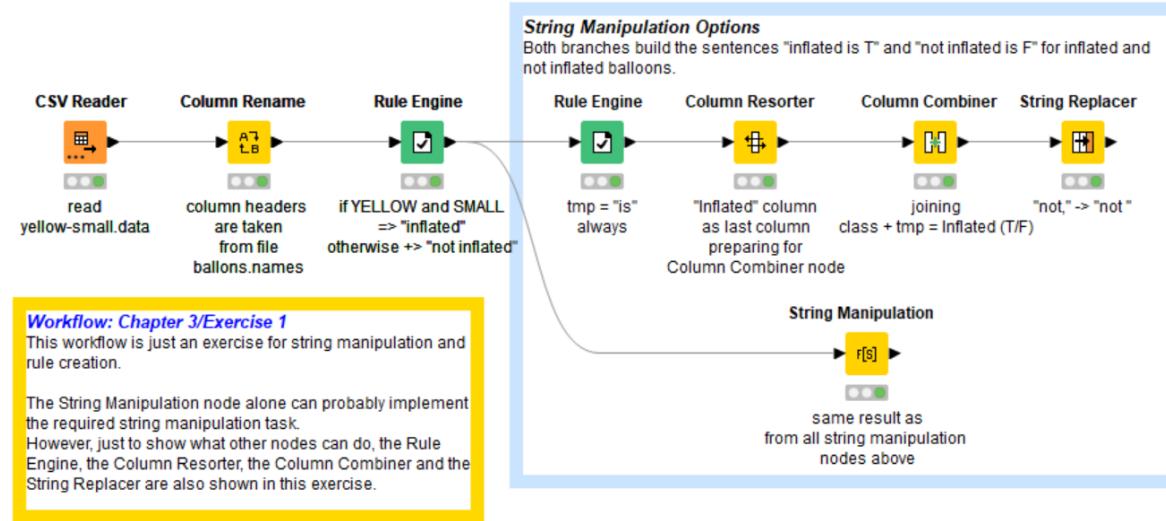
"not inflated is F"

En donde "inflated/not inflated" proviene de la columna "class" y "T/F" de la columna "inflated".

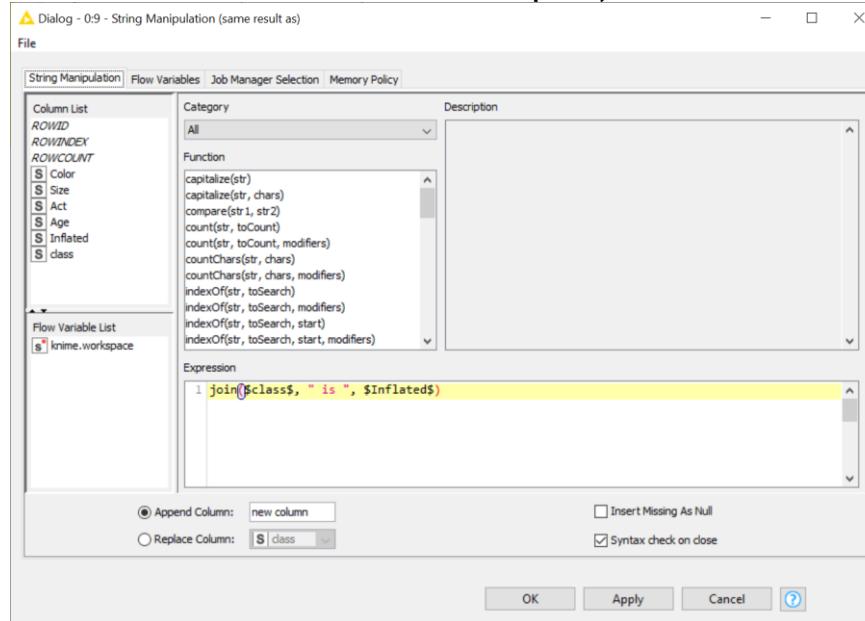
#### Solución al Ejercicio 1:

Hay dos maneras de proceder en este ejercicio.

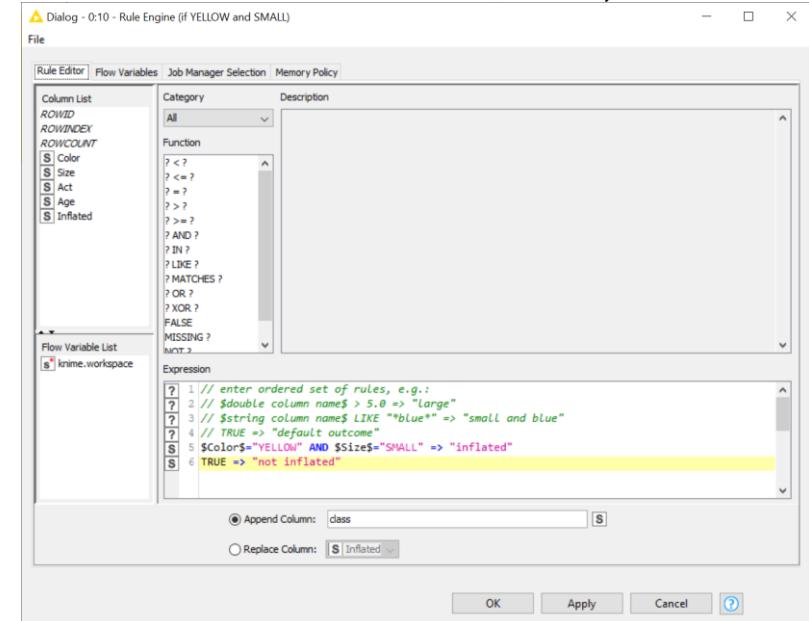
- Con una serie de nodos "String Manipulation" y "Rule Engine" dedicados
- Con un nodo "Rule Engine" y un nodo "String Manipulation" con sus funciones



### 3.57. Ejercicio 1: Configuración del nodo "String Manipulation" (nodo comentado con "mismo resultado que ...")



### 3.58. Ejercicio 1: Configuración del nodo The "Rule Engine" (nodo comentado con "if YELLOW and SMALL ...")

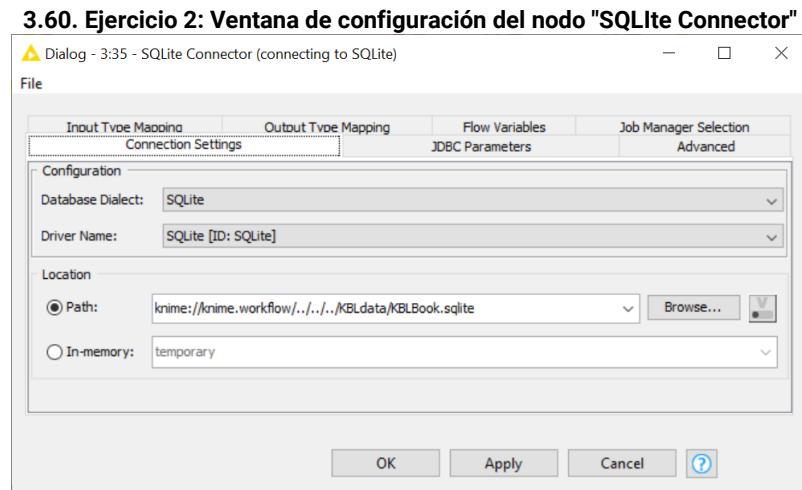
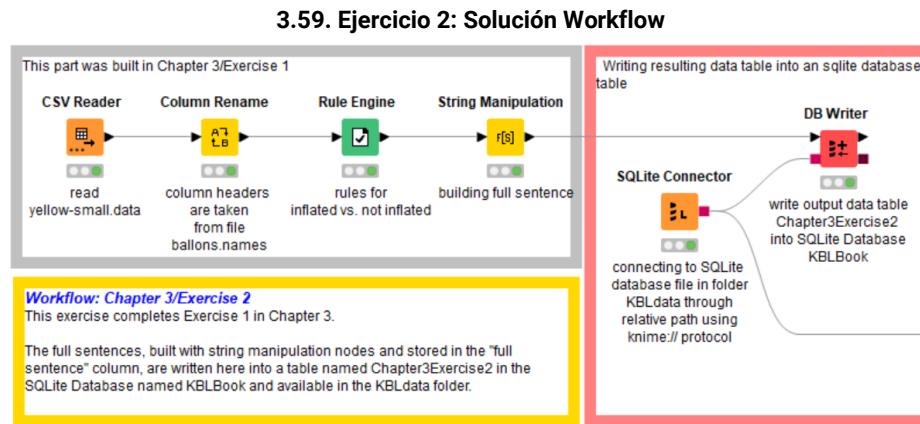


## Ejercicio 2

Este ejercicio es una extensión del Ejercicio 1 anterior.

Escriba la última tabla de datos del flujo de trabajo del Ejercicio 1 en una tabla llamada "Chapter3Exercise2" en la base de datos SQLite "KBLBook.sqlite", utilizando el "Conector SQLite" y el nodo "Database Writer".

Solución al Ejercicio 2



## Ejercicio 3

Lea el archivo adult.data. A partir de este conjunto de datos, muestre tres gráficos:

- "Age" Histograma por sexo en 10 franjas de edad
- "Work class" Gráfico de barras como número de ocurrencias para cada valor de clase de trabajo
- Gráfico de dispersión de "Age" frente a "hours per week" .

Construya el histograma y el gráfico de barras utilizando un nodo "Bar Chart" y el gráfico de dispersión utilizando un nodo "Scatter Plot".

En el gráfico de dispersión "age" frente a "hours per week", seleccione todos los puntos con "age" = 90 y extráigalos con un nodo "Row Filter" en la columna "Seleccionado"(...) = "true".

¿Cuántas personas de 90 años están incluidas en el conjunto de datos?

### Solución al ejercicio 3

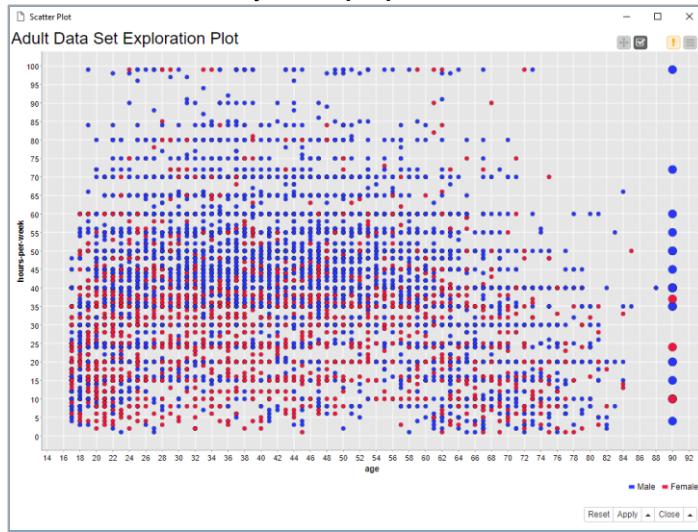
Gráfico de dispersión "age" frente a "hours per week".

- Para asegurarnos de que todos los registros se grafican necesitamos cambiar el valor por defecto del ajuste "Maximum Number of Rows" en la pestaña "options" de la ventana de configuración del nodo "Scatter Plot". Debemos asegurarnos de que este número es mayor que el número de registros del conjunto de datos de entrada. El trazado de todos los registros en lugar de sólo el número predeterminado requerirá, por supuesto, un mayor tiempo de ejecución.
- En la pestaña "Views Control" debemos habilitar la selección rectangular. Abrimos la vista de nodos, activamos el botón de selección en la esquina superior derecha y dibujamos un rectángulo alrededor de nuestras personas de 90 años a la derecha del gráfico de dispersión (si la "edad" se ha colocado en el eje x). Luego hacemos clic en el botón "Close" en la esquina inferior derecha de la vista y aceptamos los cambios.
- Un nodo A "Row Filter" extrae finalmente los registros con la columna "Selected (...)" column = true. Se han seleccionado 43 puntos que representan a personas de 90 años.
- Opcionalmente, coloreamos los puntos en azul para los registros masculinos y en rojo para los femeninos con un nodo "Color manager".

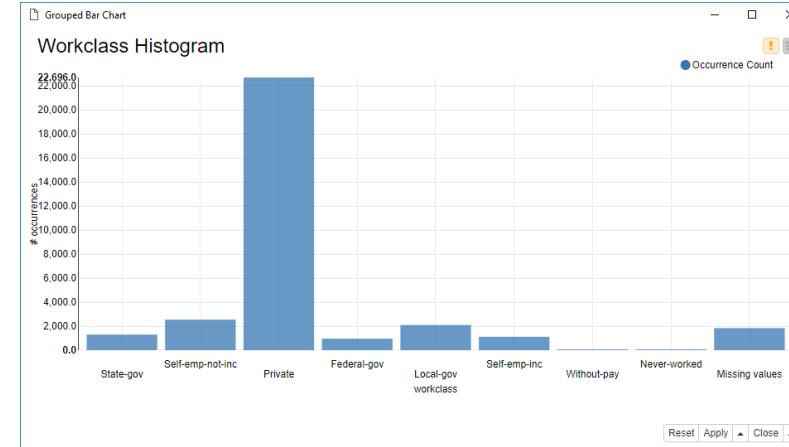
Gráfico de barras sobre el número de ocurrencias en cada clase de trabajo.

- Aquí hemos utilizado un nodo "Gráfico de barras" que cuenta el número de ocurrencias en la categoría "clase de trabajo". Age Histogram for Males and Females
  - Primero construimos automáticamente 10 franjas de edad utilizando el nodo "Auto-Binner".
  - A continuación, utilizamos un nodo "Pivoting" para contar el número de apariciones de hombres y mujeres en las diferentes franjas de edad
  - Usando un nodo de "Manipulación de cadenas" cambiamos "[" por "(" para clasificar y luego ordenamos las franjas de edad en orden ascendente
  - Por último, un nodo "Gráfico de barras" muestra los dos números uno al lado del otro para mujeres y hombres. El efecto lado a lado se obtuvo seleccionando "Agrupado" como "Tipo de Gráfico" en la pestaña "Opciones Generales de Gráfico" en la ventana de configuración del nodo.

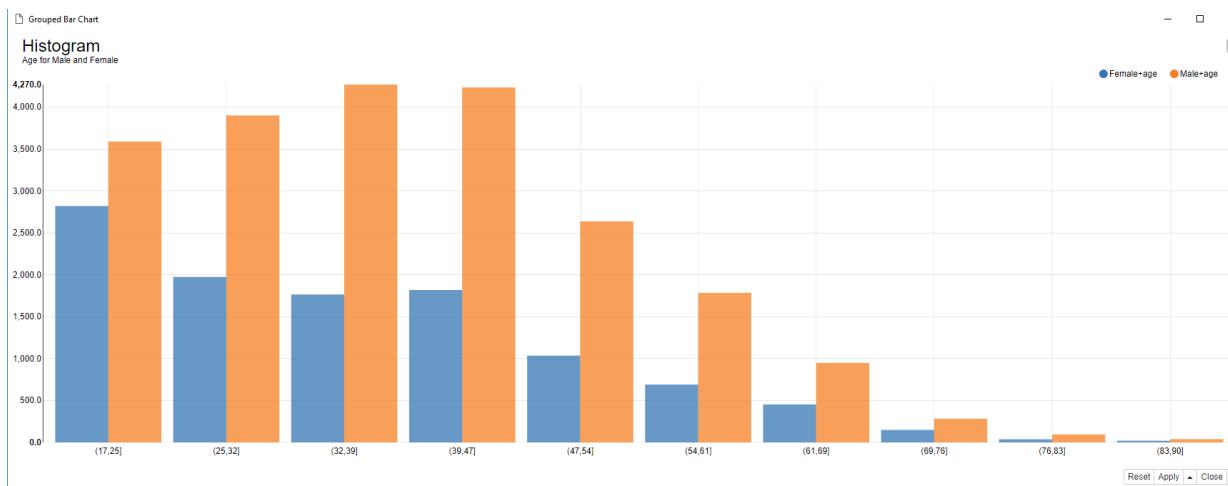
**3.61. Diagrama de dispersión de "age" vs. "hours per week" para el adult data set. 90-year old people ha sido seleccionado.**



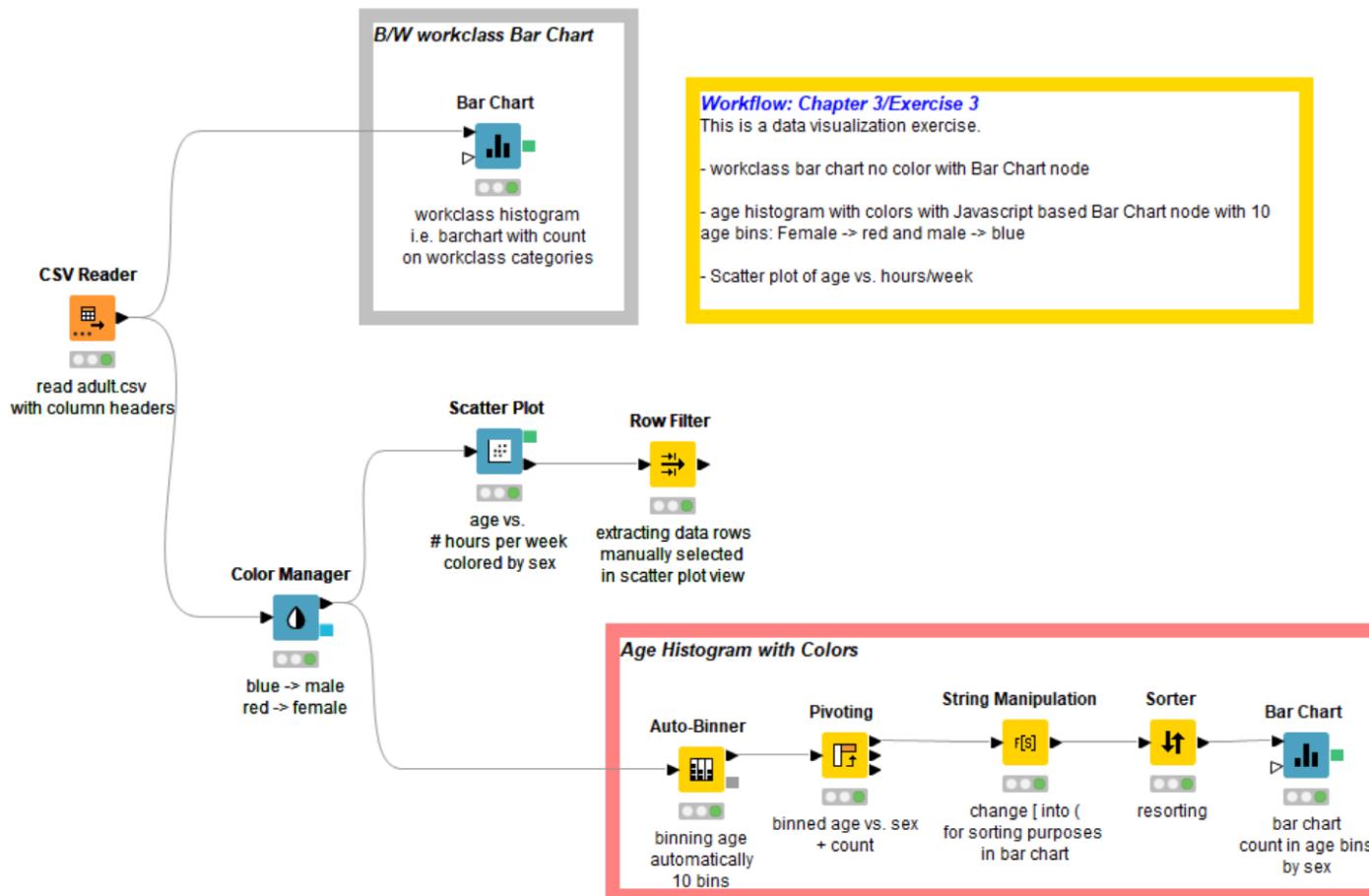
**3.62. Gráfico de barras del numero de cocuurencias de los valores de "work class" en el data set adult.**



**3.63. Histograma de edad para hombres y mujeres a partir de un nodo "Gráfico de barras" ..**



### 3.64. Ejercicio 3: Workflow



# Chapter 4. Mi Primer Modelo

## 4.1. Introducción

Por fin hemos llegado al corazón de la plataforma KNIME Analytics: el modelado de datos. Hay dos categorías de nodos en el panel "Node Repository" totalmente dedicados al modelado de datos: "Analytics" → "Statistics" and "Analytics" → "Mining". La categoría "Estadísticas" contiene nodos para calcular parámetros estadísticos y realizar pruebas estadísticas. La categoría "Minería" contiene principalmente algoritmos de aprendizaje automático, desde redes neuronales artificiales hasta clasificadores bayesianos, desde clustering hasta máquinas de vectores de apoyo, y mucho más.

El modelado de datos consta de dos fases: el entrenamiento del modelo sobre un conjunto de datos (el conjunto de datos de entrenamiento) y la aplicación del modelo a un conjunto de datos nuevos (datos vivos o un conjunto de datos de prueba). Cumpliendo con estas dos fases, los algoritmos de modelado de datos en KNIME Analytics Platform se implementan con dos nodos: un nodo "Learner" para entrenar el modelo y un nodo "Predictor" para aplicar el modelo. El nodo "Predictor" adquiere otro nombre cuando se trata de algoritmos de entrenamiento no supervisado.

El nodo "Learner" reproduce la fase de entrenamiento o aprendizaje del algoritmo en un conjunto de datos de entrenamiento específico. El nodo "Predictor" clasifica nuevos datos desconocidos utilizando el modelo producido por el nodo "Learner". Por ejemplo, la categoría "Mining" → "Bayes" implementa clasificadores bayesianos ingenuos (naïve Bayesian classifiers). El nodo "Naïve Bayes Learner" construye (aprende) un conjunto de reglas de Bayes sobre el conjunto de datos de aprendizaje (o entrenamiento) y los almacena en el modelo. El nodo "Naïve Bayes Predictor" lee entonces las reglas de Bayes del modelo y las aplica a los datos entrantes.

Todos los algoritmos de modelado de datos necesitan un conjunto de datos de entrenamiento para construir el modelo. Normalmente, después de construir el modelo, es útil evaluar la calidad del mismo, para asegurarse de que no estamos creyendo en las predicciones producidas por un modelo de baja calidad. Para la evaluación, se utiliza un nuevo conjunto de datos, denominado conjunto de datos de prueba. Por supuesto, el conjunto de datos de prueba tiene que contener datos diferentes a los del conjunto de datos de entrenamiento, para permitir la evaluación de la capacidad del modelo de funcionar correctamente en datos nuevos desconocidos. Por lo tanto, a efectos de evaluación, todos los algoritmos de modelización necesitan también un conjunto de datos de prueba.

Para proporcionar un conjunto de entrenamiento y un conjunto de prueba para el algoritmo, normalmente el conjunto de datos original se divide en dos conjuntos de datos más pequeños: el conjunto de datos de aprendizaje/entrenamiento y el conjunto de datos de prueba. Para particionar, reorganizar y volver a unir los conjuntos de datos, utilizamos nodos de la categoría "Manipulation" → "Row" → "Transform".

A veces pueden surgir problemas cuando hay valores perdidos en los datos. De hecho, no todos los algoritmos de modelización pueden tratar los datos que faltan. El modelo también puede requerir que el conjunto de datos tenga una distribución normal. Para eliminar los datos que faltan en los conjuntos de datos y para normalizar los valores de una columna, podemos utilizar más nodos de la categoría "Manipulation" → "Column" → "Transform".

En este capítulo, ofrecemos una visión general de los nodos de aprendizaje automático, es decir, los nodos de aprendizaje y predicción, y de los nodos para manipular las filas y transformar los valores en las columnas. Trabajamos con el conjunto de datos de adultos, ya utilizado en los capítulos anteriores. Aquí creamos un nuevo grupo de flujo de trabajo "Chapter4" y, dentro de él, un nuevo flujo de trabajo llamado "Data Preparation". Utilizamos este flujo de trabajo para preparar los datos para posteriores operaciones de modelado de datos. El primer paso de este flujo de trabajo es leer el conjunto de datos de adultos con un nodo "CSV Reader"

## 4.2. Dividir y combinar conjuntos de datos

Dado que muchos modelos necesitan datos de entrenamiento y datos de prueba separados, estos dos conjuntos de datos tienen que ser configurados antes de modelar los datos. Para extraer dos conjuntos de datos -uno de entrenamiento y otro de prueba- del conjunto de datos original, se puede utilizar el nodo "Partitioning". Si sólo se necesita un conjunto de entrenamiento y no un conjunto de prueba o si el conjunto de datos original es demasiado grande para ser utilizado en su totalidad, podemos utilizar el nodo "Row Sampling".

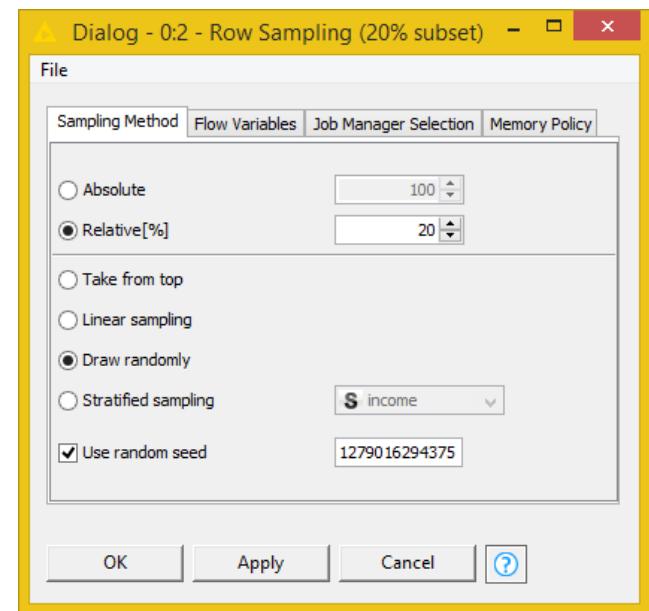
# Muestreo por filas (Row Sampling)

El nodo "Row Sampling" extrae una muestra (= un subconjunto de filas) de los datos de entrada. La ventana de configuración le permite especificar:

- El tamaño de la muestra como número absoluto de filas o como porcentaje del conjunto de datos original
- El modo de extracción
  - o "Take from the top" significa las filas superiores del conjunto de datos original
  - o "Linear Sampling" toma la primera y la última fila y muestrea entre estas filas en pasos regulares
  - o "Draw randomly" extrae filas al azar
  - o "Stratified sampling" extrae filas de forma aleatoria por lo que la distribución de los valores en la columna seleccionada se mantiene aproximadamente en la tabla de salida

Para "Draw randomly" y "Stratified sampling" se puede definir una semilla aleatoria para que la extracción aleatoria sea reproducible (nunca se sabe cuándo se necesita recrear exactamente el mismo conjunto de entrenamiento aleatorio).

## 4.1. Ventana de configuración del nodo "Row Sampling"



En la figura 4.1, seleccionamos un tamaño del 20% del conjunto de datos original para el conjunto de aprendizaje. Las filas se extrajeron al azar del conjunto de datos original. Un tamaño del 20% del conjunto de datos original es probablemente demasiado pequeño; para estar seguros de que todas las clases están realmente representadas en el conjunto de aprendizaje podríamos utilizar la opción de muestreo estratificado.

**Nota.** El nodo "Row Sampling" sólo produce un subconjunto de datos que podemos utilizar para entrenar o para probar un modelo, pero no ambos. Si queremos generar dos subconjuntos de datos, el primero según nuestras especificaciones en la ventana de configuración, y el segundo con las filas restantes, tenemos que utilizar el nodo "Partitioning".

# Partitioning

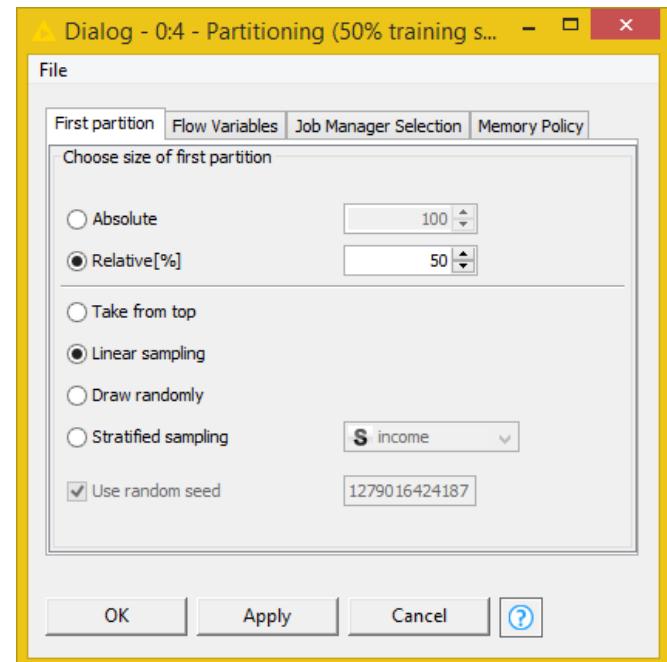
El nodo "Partitioning" realiza la misma tarea que el nodo "Row Sampling": extrae una muestra (= un subconjunto de filas) de los datos de entrada. También construye un segundo conjunto de datos con las filas restantes y lo pone a disposición en el puerto de salida inferior.

La ventana de configuración permite especificar

- El tamaño de la muestra como número absoluto de filas o como porcentaje del conjunto de datos original
- El modo de extracción
  - o "Take from the top" se refiere a las primeras filas del conjunto de datos original
  - o "Linear Sampling" toma la primera y la última fila y muestrea entre las filas a pasos regulares
  - o "Draw randomly" extrae filas al azar
  - o El "Stratified sampling" extrae filas en las que la distribución de los valores de la columna seleccionada se mantiene aproximadamente en la tabla de salida

Para "Draw randomly" y "Stratified sampling" puede definirse una semilla aleatoria para que la extracción aleatoria sea reproducible (nunca se sabe cuándo hay que volver a crear el mismo conjunto de aprendizaje).

4.2. Ventana de configuración del nodo "Partitioning"



Aquí, seleccionamos un tamaño del 50% del conjunto de datos original para el conjunto de entrenamiento más un modo de extracción lineal. El conjunto de entrenamiento se puso a disposición en el puerto de salida superior; los datos restantes se pusieron a disposición en el puerto de salida inferior. En la técnica de muestreo lineal, las filas se ajustan al orden definido en el conjunto de datos original.

A veces es necesario presentar las filas de datos en el orden original al algoritmo de entrenamiento, por ejemplo, cuando se trata de la predicción de series temporales. El orden de las filas, en este caso, es un orden temporal y es utilizado por el modelo para representar secuencias temporales. En éste caso, el muestreo lineal technique is advised. If we are dealing with time series analysis, where the past and the future have to remain separate, the "take from top" strategy is recommended.

Sin embargo, a veces no es aconsejable presentar las filas a un nodo de aprendizaje en un orden específico; de lo contrario, el modelo podría aprender el orden de las filas entre todos los demás patrones subyacentes. Por ejemplo, el orden de los clientes en la base de datos no significa nada más que la asignación de una clave de identificación secuencial a cada cliente. Para estar seguros de que las filas de datos se presentan al nodo Learner del modelo en un orden aleatorio, podemos extraerlas al azar o aplicar el nodo "Shuffle".

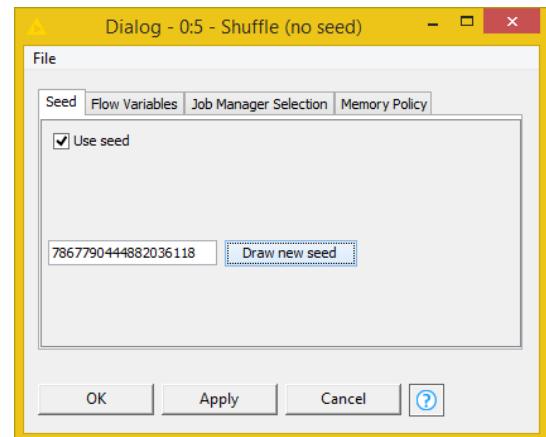
## Shuffle

El nodo "Shuffle" baraja las filas de la tabla de entrada poniéndolas en un orden aleatorio.

En general, el nodo "Shuffle" no necesita ser configurado. Si queremos ser capaces de repetir exactamente el mismo barajado aleatorio de las filas, necesitamos usar una semilla, como sigue:

- Marque la casilla "Use seed".
- Haga clic en el botón "Draw new seed" para crear una semilla para el barajado aleatorio y volver a crearla en cada ejecución

4.3. Configuration window for the "Shuffle" node



Sólo aplicamos el nodo "Shuffle" al conjunto de entrenamiento. No importa si las filas de datos del conjunto de prueba se presentan en un orden predefinido o no.

Ahora tenemos un conjunto de datos de entrenamiento y un conjunto de datos de prueba. ¿Pero qué pasa si queremos recrear el conjunto de datos original reunificando el conjunto de entrenamiento y el de prueba? KNIME tiene un nodo "Concatenate" que resulta muy útil para esta tarea.

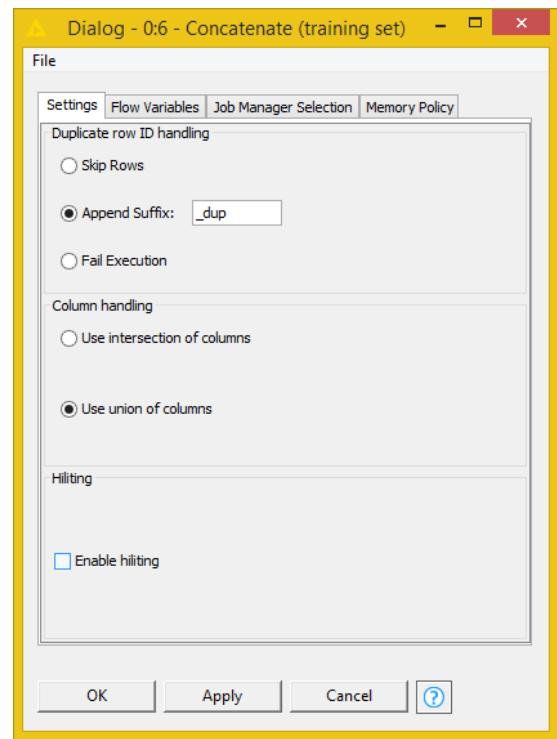
## Concaternar (Concatenate)

El nodo "Concatenate" tiene dos puertos de entrada, cada uno para un conjunto de datos. El nodo "Concatenate" añade el conjunto de datos del puerto de entrada inferior al conjunto de datos del puerto de entrada superior.

La ventana de configuración se ocupa de lo siguiente

- Qué hacer con las filas con el mismo ID
  - o omitir las filas del segundo conjunto de datos
  - o renombrar el RowID con un sufijo añadido
  - o abortar la ejecución con un error (Esta opción puede utilizarse para comprobar si los RowIDs son únicos)
- Qué columnas conservar
  - o todas las columnas del segundo y primer conjunto de datos (unión de columnas)
  - o sólo la intersección de las columnas de los dos conjuntos de datos (es decir, las columnas contenidas en ambas tablas)
- La opción "Enable hiliting" hace referencia a la propiedad hiliting disponible en los antiguos nodos "Data Views".

4.4. Ventana de configuración del nodo  
"Concatenate"



La figura 4.4 muestra un ejemplo del funcionamiento del nodo "Concatenate", cuando se activan las siguientes opciones de la ventana de configuración:

- añadir el sufijo al RowID en las filas con RowID duplicado
- utilizar la unión de columnas
- no habilitar el hilado (hiliting)

Un nodo similar al nodo "Concatenar" es el nodo "Concatenate (Optional in)". El nodo "Concatenate (Optional in)" funciona exactamente igual que el nodo "Concatenate", pero permite concatenar hasta 4 conjuntos de datos al mismo tiempo.

**4.5. Este es un ejemplo de cómo funciona el nodo "Concatenate"**

*Primera tabla de datos*

RowID	scores
Row1	22
Row3	14
Row4	10

*Segunda tabla de datos*

RowID	name	scores
Row1	The Black Rose	23
Row2	Cynthia	2
Row5	Tinkerbell	4
Row6	Mother	6
Row7	Augusta	8
Row8	The Seven Seas	3

*Tabla de datos Concatenada*

RowID	name	scores
Row1	?	22
Row3	?	14
Row4	?	10
Row1_dup	The Black Rose	23
Row2	Cynthia	2
Row5	Tinkerbell	4
Row6	Mother	6
Row7	Augusta	8
Row8	The Seven Seas	3

## 4.3. Transformando columnas (Transform Columns)

Hemos obtenido un conjunto de entrenamiento y un conjunto de prueba a partir del conjunto de datos original. El conjunto de datos original, sin embargo, contenía valores perdidos en algunas de sus columnas de datos y algunos algoritmos de aprendizaje automático no pueden tratar con valores perdidos. Las celdas de datos de KNIME, de hecho, pueden tener un estado especial de "missing value". Por defecto, los valores perdidos se muestran en la vista de la tabla con un signo de interrogación ("?").

Algunos de los nodos de aprendizaje pueden ignorar las filas de datos que contengan valores perdidos, reduciendo así la base de datos con la que trabajan; y algunos de los nodos de aprendizaje simplemente fallarán al encontrar un valor perdido. En el último caso, se requiere una estrategia para tratar los valores que faltan; pero incluso en el primer caso, una estrategia para tratar los valores que faltan es aconsejable para ampliar la base de datos para el futuro modelo.

Hay muchas estrategias para tratar los valores que faltan y se han escrito libros sobre qué estrategia es mejor utilizar en cada contexto. Cada estrategia consiste en sustituir el valor ausente en cuestión por otro valor plausible. Cuál es el valor más plausible depende del contexto y, a menudo, de los conocimientos del experto.

KNIME Analytics Platform implementa las estrategias más comunes para tratar los valores perdidos, como el uso del valor medio de la columna de datos, la media móvil, el máximo/mínimo, el valor más frecuente, la interpolación lineal y media, el valor anterior o el siguiente, un valor fijo, y probablemente por ahora más. Por supuesto, siempre está disponible la opción de eliminar la fila de datos que contiene un valor perdido.

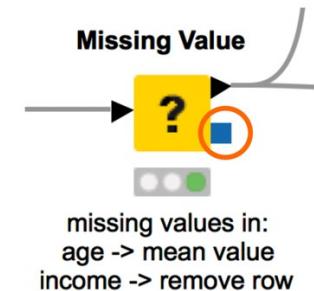
El nodo que se ocupa de los valores perdidos se llama "Missing Value". El nodo "Missing Value" toma una tabla de datos como entrada y reemplaza los valores perdidos en todas partes o sólo en las columnas seleccionadas con un valor de su elección. La nueva tabla de datos con los valores perdidos reemplazados se produce entonces en el puerto de salida superior. De hecho, este nodo tiene dos puertos de salida. El puerto de salida inferior tiene la forma de un cuadrado azul en lugar del habitual triángulo negro. Un puerto cuadrado azul significa un modelo compatible con PMML.

# PMML

PMML (Predictive Model Markup Languages) es una estructura estándar basada en XML que permite almacenar modelos predictivos y transformaciones de datos.

Al ser una estructura estándar, permite la portabilidad de modelos y transformaciones entre plataformas y aplicaciones.

La plataforma KNIME Analytics soporta PMML para modelos y transformaciones. Los cuadrados azules como puertos de entrada y salida en los nodos KNIME identifican los objetos compatibles con PMML, ya sean modelos predictivos o transformaciones ETL.



En KNIME no sólo es posible exportar modelos y transformaciones individuales como estructuras PMML, sino también concatenarlas modularmente de modo que la estructura PMML final contenga la secuencia de transformaciones y el modelo creado en el flujo de trabajo y alimentado en la estructura PMML. Dos nodos son clave para el PMML modular: "PMML Transformation Appender" y "PMML Model Appender".

**Note.** Algunas de las estrategias de valores perdidos están marcadas con un asterisco en los menús de la ventana de configuración en el nodo "Valor perdido". El asterisco indica que dichas transformaciones no son compatibles con PMML.

# Valores faltantes (Missing Value)

El nodo "Missing Value" sustituye los valores ausentes en un conjunto de datos en todas partes o sólo en las columnas seleccionadas por un valor de su elección.

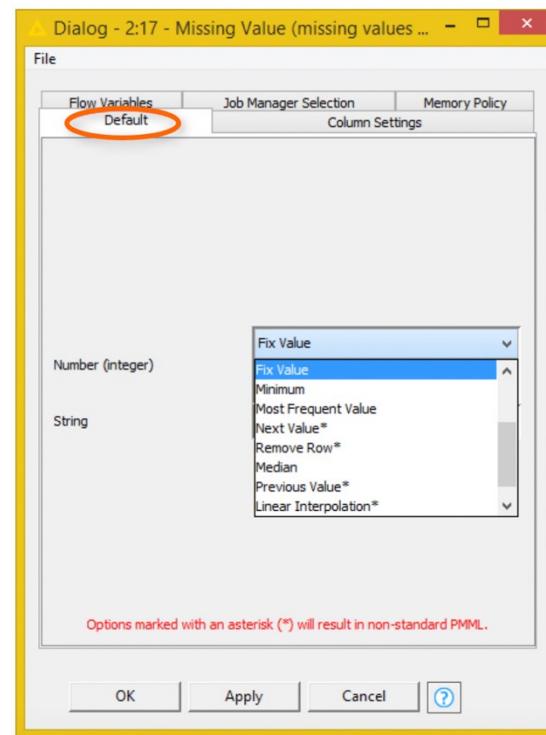
En la pestaña "**Default**", los valores de sustitución se definen por separado para las columnas de tipo numérico y de cadena y se aplican a todas las columnas de datos del mismo tipo.

En la pestaña "**Column Settings**", se define un valor de sustitución específicamente para cada columna de datos seleccionada y se aplica sólo a esa columna. Para definir el valor de sustitución de una columna:

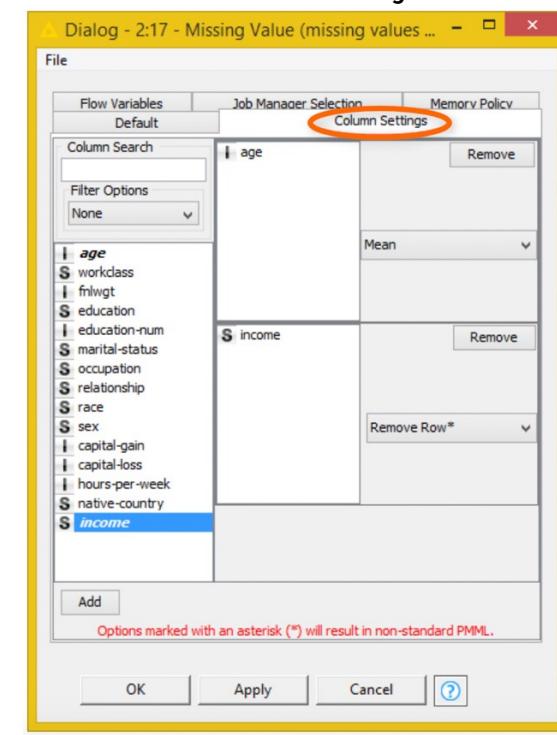
- aga doble clic en la columna de la lista de la izquierda
- ó
- Seleccione la columna en la lista de la izquierda
- Haga clic en el botón "Add" debajo de la lista

A continuación, seleccione la estrategia de tratamiento de los valores perdidos que deseé.

4.7. Configuración para la ventana "Missing Value"  
: Tab "Default"



4.8. Configuración para la ventana "Missing Value"  
: Tab "Column Settings"



Se proporciona un cuadro de "Column Search" para ayudar a encontrar columnas entre muchas. También se proporciona un botón "Remove" en el marco de la columna de datos para eliminar la estrategia individual de manejo de valores perdidos para la columna seleccionada.

Introducimos un nodo "Missing Value" antes del nodo "Partitioning" en nuestro flujo de trabajo "Data Preparation". Aquí establecemos 0 como valor fijo para reemplazar los valores perdidos en todas las columnas numéricas y "Do nothing" para los valores perdidos en las columnas de cadena. A continuación, para las columnas "age" (Entero) e "income" (Cadena), establecemos estrategias de sustitución individuales para los valores perdidos. En la columna "edad", los valores perdidos se sustituyen por el valor medio de la columna de datos; en la columna "income", las filas con valores perdidos simplemente se eliminan. Mientras que la estrategia de valores perdidos para la "age" es puramente demostrativa, la estrategia de valores perdidos para los "income" es necesaria, ya que queremos predecir el valor de los "income" dados todos los demás atributos del censo para cada persona.

Algunos modelos de datos -como las redes neuronales, el clustering u otros modelos basados en la distancia- requieren valores de atributos de entrada normalizados, para que los datos se normalicen y sigan la distribución gaussiana o simplemente caigan en el intervalo [0,1]. Para cumplir este requisito, utilizamos el nodo "Normalize".

# Normalizer (normalizador)

El nodo "Normalizer" normaliza los datos; es decir, transforma los datos para que caigan en un intervalo determinado o para que sigan una distribución estadística determinada.

El nodo "Normalizer" se encuentra en el panel "Node Repository" en la categoría "Manipulation" → "Column" → "Transform"

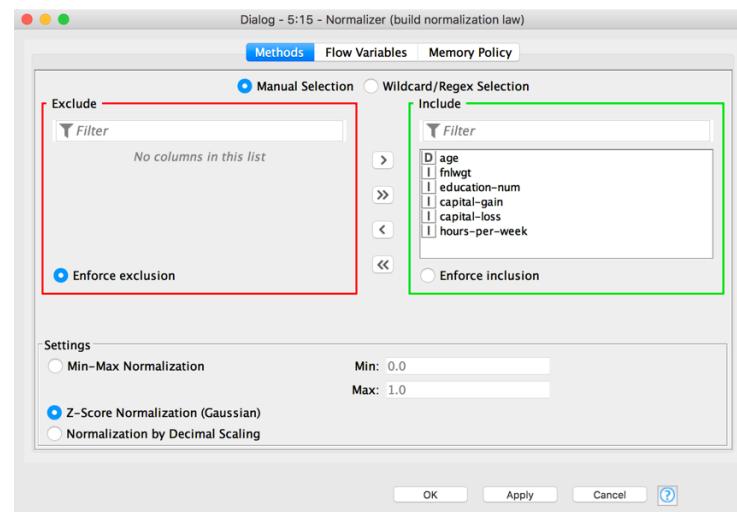
La ventana de configuración requiere:

- la lista de columnas de datos numéricos a normalizar
- el método de normalización

La selección de la columna se realiza mediante un cuadro "Exclude"/"Include", por selección manual o por selección Wildcard/RegEx. Para la selección manual:

- Las columnas que se van a normalizar aparecen en el cuadro "Normalize". Todas las demás columnas aparecen en el marco "No normalizar".
- Para pasar del marco "Normalize" al marco "Do not normalize" y viceversa, utilice los botones "add" y "remove". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".

## 4.9. Configuración del nodo Normalizer"



El nodo "Normalizer" tiene 2 puertos de salida:

- En el puerto superior se encuentran los datos normalizados
- En el puerto inferior se proporcionan los parámetros de transformación para repetir la misma normalización en otros datos (puerto cuadrado azul claro/azul oscuro)

**Nota.** Salida/lectura (output/read ) de datos de los puertos triangulares. Parámetros de output/read de puertos cuadrados: parámetros del modelo, parámetros de normalización, parámetros de transformación, parámetros gráficos, etc...

Hay dos nodos normalizadores: "Normalizer" ay "Normalizer (PMML)" . Realizan exactamente la misma tarea utilizando los mismos parámetros. La única diferencia está en la estructura de los parámetros exportados: Estructura propietaria de KNIME (cuadrado azul claro) o estructura compatible con PMML (cuadrado azul oscuro).

# Metodos de normalización (Normalization Methods)

## Normalización Mín-Máx

Se trata de una transformación lineal en la que todos los valores de los atributos de una columna caen en el intervalo [mín., máx.] y los mínimos y máximos son especificados por el usuario.

## Normalización Z-score

También se trata de una transformación lineal en la que los valores de cada columna tienen una distribución gaussiana (0,1), es decir, la media es 0,0 y la desviación estándar es 1,0.

## Normalización por escala decimal

El valor máximo de una columna se divide j veces por 10 hasta que su valor absoluto sea menor o igual a 1. A continuación, todos los valores de la columna se dividen por 10 a la potencia de j.

# Nodo Normalizer (Apply)

Este nodo "Normalizer (Apply)" normaliza los datos; es decir, transforma los datos para que caigan en un intervalo determinado o para que sigan una distribución estadística determinada. Sin embargo, no calcula los parámetros de transformación; los obtiene de un nodo "Normalizador" previamente aplicado a un conjunto de datos similar.

El nodo "Normalizer (Apply)" tiene dos puertos de entrada: uno para los datos a normalizar y otro para los parámetros de normalización.

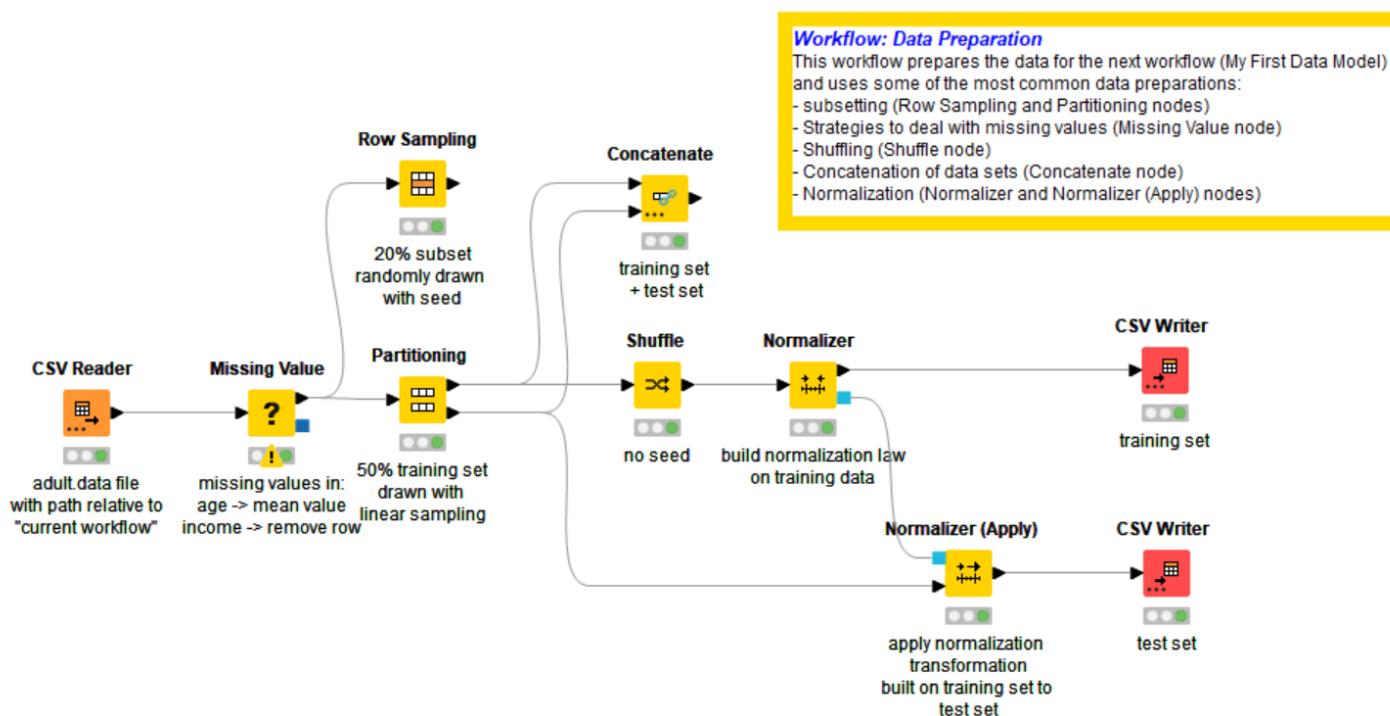
El nodo "Normalizer (Apply)" se encuentra en el panel "Node Repository" en la categoría the "Manipulation" → "Column" → "Transform"

No se requiere ninguna configuración adicional.

Aplicamos el nodo "Normalizer" al conjunto de entrenamiento desde el puerto de salida del nodo "Partitioning", para normalizar el conjunto de entrenamiento y definir los parámetros de normalización. Luego introducimos un nodo "Normalizer (Apply)" para leer los parámetros de normalización y utilizarlos para normalizar los datos restantes desde el nodo "Partitioning" (2º puerto de salida).

Ahora vamos a escribir el conjunto de datos de entrenamiento y el conjunto de datos de prueba procesados en archivos CSV, llamados "training\_set.csv" y "test\_set.csv" respectivamente. Utilizamos dos nodos "CSV Writer": uno para escribir el conjunto de entrenamiento en el archivo "training\_set.csv" y otro para escribir el conjunto de prueba en el archivo "test\_set.csv". Estos dos últimos nodos concluyen el flujo de trabajo de "Data Preparation".

#### 4.10. Flujo de trabajo "Data Preparation"



## 4.4. Modelos de Aprendizaje Automático (Machine Learning Models)

Ahora vamos a crear un nuevo flujo de trabajo y lo llamaremos "My First Model". Utilizaremos este flujo de trabajo para mostrar cómo los modelos pueden ser entrenados en un conjunto de datos y luego aplicados a nuevos datos. Para dar una visión general, pasaremos por algunos paradigmas de métodos de análisis de datos estándar. Estándar se refiere aquí a la forma en que los paradigmas se implementan en KNIME -por ejemplo, con un nodo como el Learner un nodo separado como el Predictor/Applier- y no con respecto a la calidad del algoritmo en sí.

Los dos primeros nodos de este nuevo flujo de trabajo son dos nodos "CSV Reader": uno para leer el conjunto de entrenamiento y otro para leer el conjunto de prueba que se guardó en dos archivos CSV en el flujo de trabajo "Data Preparation" al final de la última sección.

En este flujo de trabajo queremos predecir la etiqueta "income" del conjunto de datos de adultos utilizando los demás atributos y basándonos en algunos modelos diferentes. Esta sección no pretende comparar esos modelos en términos de precisión o rendimiento. De hecho, no se ha dedicado mucho trabajo a optimizar estos modelos para que sean los predictores más precisos. Por el contrario, el objetivo aquí es mostrar cómo crear y configurar esos modelos. Cómo optimizar los parámetros del modelo para garantizar que sean lo más precisos posible es un problema que puede explorarse en otro lugar [3] [4] [5].

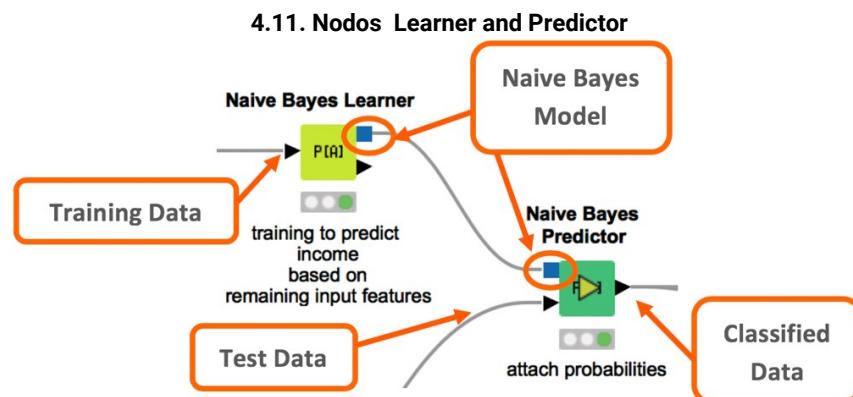
En todo problema de predicción/clasificación supervisada, necesitamos un conjunto de entrenamiento etiquetado; es decir, un conjunto de entrenamiento en el que cada fila ha sido asignada a una clase determinada. Estas clases de salida de las filas de datos están contenidas en una columna del conjunto de datos: se trata de la clase o columna objetivo.

La mayoría de los paradigmas de minería de datos y estadística constan de dos nodos: un Learner y un Predictor

El nodo Learner define los parámetros y las reglas del modelo que hacen que éste sea adecuado para realizar una determinada tarea de clasificación/predicción. El nodo Learner utiliza la tabla de datos de entrada como conjunto de entrenamiento para definir estos parámetros y reglas. La salida de este nodo es un conjunto de reglas y/o parámetros: el modelo.

El nodo Predictor utiliza el modelo construido en el paso anterior y lo aplica a un conjunto de datos desconocidos (es decir, nuevos sin clasificar) para realizar la tarea de clasificación/ predicción para la que fue construido.

El nodo Learner requiere una tabla de datos como entrada y proporciona un modelo como salida. El puerto de salida del nodo



Learner se representa como un cuadrado azul, que es el símbolo de un modelo compatible con PMML.

El nodo Predictor toma una tabla de datos y un modelo en los puertos de entrada (un triángulo negro para los datos y un cuadrado azul para el modelo) y proporciona una tabla de datos que contiene los datos clasificados en el puerto de salida.

## Modelo Naïve Bayes

Empecemos con un modelo Bayes naïve. Un modelo bayesiano define un conjunto de reglas, basadas en las distribuciones gaussianas y en las probabilidades condicionales de los datos de entrada, para asignar una fila de datos a una clase de salida [3][4][5]. En el panel "Node Repository" en la categoría "Mining" → "Bayes"

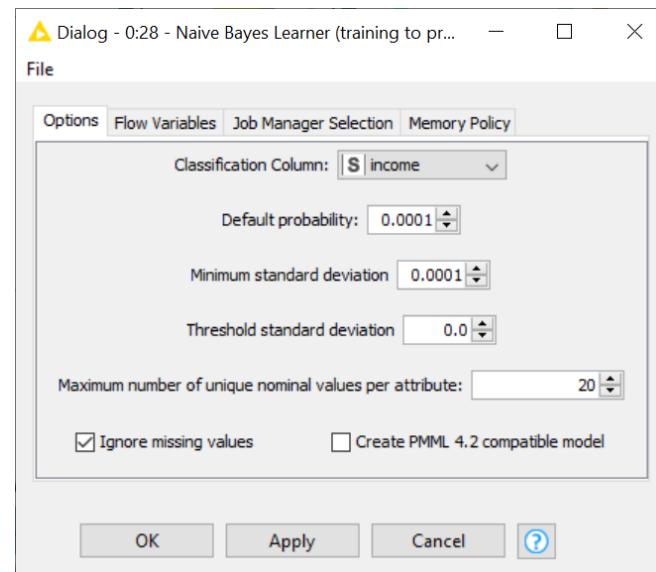
Let's start with a naïve Bayes model. A Bayesian model defines a set of rules, based on the Gaussian distributions and on the conditional probabilities of the input data, to assign a data row to an output class [3][4][5]. In the "Node Repository" panel in the "Mining" → "Bayes" category encontramos dos nodos: "Naïve Bayes Learner" y "Naïve Bayes Predictor".

### Nodo "Naïve Bayes Learner"

El nodo "Naïve Bayes Learner" crea un modelo bayesiano a partir de los datos de entrenamiento de entrada. Calcula las distribuciones y probabilidades para definir las reglas del modelo bayesiano a partir de los datos de entrenamiento. Los puertos de salida producen el modelo y los parámetros del modelo respectivamente. En la ventana de configuración hay que especificar

- La columna de clases (= la columna que contiene las clases)
- La probabilidad por defecto y la desviación estándar mínima (casi 0)
- El número máximo de valores nominales únicos permitidos por columna. Si una columna contiene más de este número máximo de valores nominales únicos, será excluida del proceso de entrenamiento.
- Cómo tratar los valores perdidos (omitar o mantener)
- Compatibilidad del modelo de salida con PMML 4.2

4.12. Ventana de configuración para el nodo "Naïve Bayes Learner"



## Nodo “Naïve Bayes Predictor”

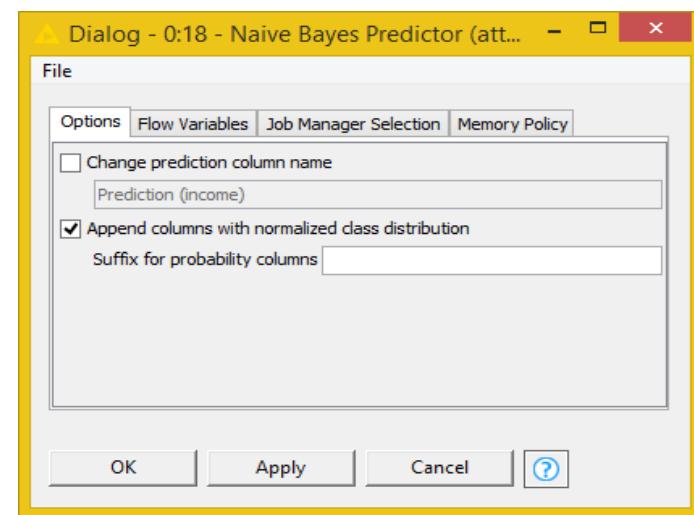
El nodo "Naïve Bayes Predictor" aplica un modelo bayesiano existente a la tabla de datos de entrada.

Todos los ajustes de configuración necesarios están disponibles en el modelo de entrada.

En la ventana de configuración sólo puede

- Añadir a la tabla de datos de entrada los valores de distribución de clase normalizados para todas las clases
- Personalizar el nombre de la columna para la clase predicha

4.13. Ventana de configuración para el nodo “Naïve Bayes Predictor”



**Note.** Todos los nodos de predicción exponen la misma ventana de configuración: una opción para añadir probabilidades de clase predichas/distribuciones normalizadas y una opción para cambiar el nombre de la columna de clase de predicción por defecto.

En el flujo de trabajo "My First Model" conectamos un nodo "Naïve Bayes Learner" al nodo "CSV Reader" que lee el conjunto de datos de entrenamiento. En la ventana de configuración del "Naïve Bayes Learner", especificamos "ingresos" como clase/columna objetivo, optamos por omitir las filas con valores perdidos en la estimación del modelo y por omitir una columna si se encontraban más de 20 valores nominales.

Después de establecer esta configuración, aparece un triángulo amarillo bajo el "Naïve Bayes Learner" para indicar que la columna "native country" del conjunto de datos de entrada tiene demasiados valores nominales (> 20 según la configuración) y será ignorada. A continuación, ejecutamos la opción "Execute" para el nodo "Naïve Bayes Learner".

El siguiente paso consiste en conectar un nodo "Naïve Bayes Predictor" al nodo "CSV Reader" para leer el conjunto de pruebas a través del puerto de datos; el nodo "Naïve Bayes Predictor" también se conecta al puerto de salida del nodo "Naïve Bayes Learner" a través del puerto del modelo.

Después de la ejecución, el "Naïve Bayes Predictor" muestra una nueva columna añadida a la tabla de salida: "Prediction (income)". Esta columna contiene las asignaciones de clase para cada fila realizadas por el modelo bayesiano. La corrección de estas asignaciones, es decir, el rendimiento del modelo, sólo puede evaluarse comparándolas con las etiquetas originales de "ingresos".

Si se activó la bandera para añadir los valores de probabilidad para cada clase de salida, en la tabla de datos final habrá tantas columnas nuevas como valores haya en la columna de clase; cada columna contiene la probabilidad para un valor de clase determinado según el modelo bayesiano entrenado.

KNIME tiene una categoría "Analytics" → "Mining" → "Scoring" con nodos que miden el rendimiento de los clasificadores. El más sencillo de estos nodos de evaluación es el nodo "Scorer". Utilizaremos el nodo "Scorer (Javascript)" porque también ofrece una visualización más agradable de los resultados..

#### 4.14. Bayes Model's Classified Data

The classified data - 0:18 - Naive Bayes Predictor (attach probabilities)							
File		Table "default" - Rows: 15363		Spec - Columns: 18		Properties Flow Variables	
Row ID	I4...	D hours-p...	S native-...	S income	D P (inco...	D P (income=>50K)	S Prediction (income)
Row0	-2.346	United-States	<=50K	0.149	0.851	>50K	
Row1	-0.085	United-States	<=50K	0.99	0.01	<=50K	
Row2	-0.085	United-States	<=50K	0.047	0.953	>50K	
Row3	0.334	United-States	>50K	0.16	0.84	>50K	
Row4	-0.085	United-States	>50K	0	1	>50K	
Row5	-0.085	India	>50K	0.055	0.945	>50K	
Row6	0.753	United-States	<=50K	0.99	0.01	<=50K	
Row7	0.334	Mexico	<=50K	0.995	0.005	<=50K	
Row8	-0.085	United-States	<=50K	0.999	0.001	<=50K	
Row9	0.334	United-States	>50K	0.674	0.326	<=50K	
Row10	-1.76	United-States	<=50K	1	0	<=50K	
Row11	-0.085	United-States	<=50K	0	1	>50K	
Row12	-0.085	United-States	>50K	0.034	0.966	>50K	
Row13	3.266	United-States	<=50K	0.941	0.059	<=50K	
Row14	0.92	United-States	<=50K	0.979	0.021	<=50K	
Row15	-0.085	United-States	<=50K	0.006	0.994	>50K	
Row16	-2.179	United-States	<=50K	0.998	0.002	<=50K	

# Scorer (Javascript)

El nodo "Scorer" compara los valores de dos columnas (columna de destino y columna de predicción) en la tabla de datos; basándose en esta comparación muestra la matriz de confusión y algunas medidas de precisión.

Este nodo produce tres tablas de datos de salida: la matriz de confusión, las estadísticas de filas correctamente identificadas para cada clase, y las medidas de precisión globales establecidas en la ventana de configuración.

Este nodo tiene una opción de Vista, donde se muestra la matriz de confusión y algunas medidas de precisión.

La ventana de configuración tiene tres pestañas: "Scorer Options", "Statistics Options", "Control Options".

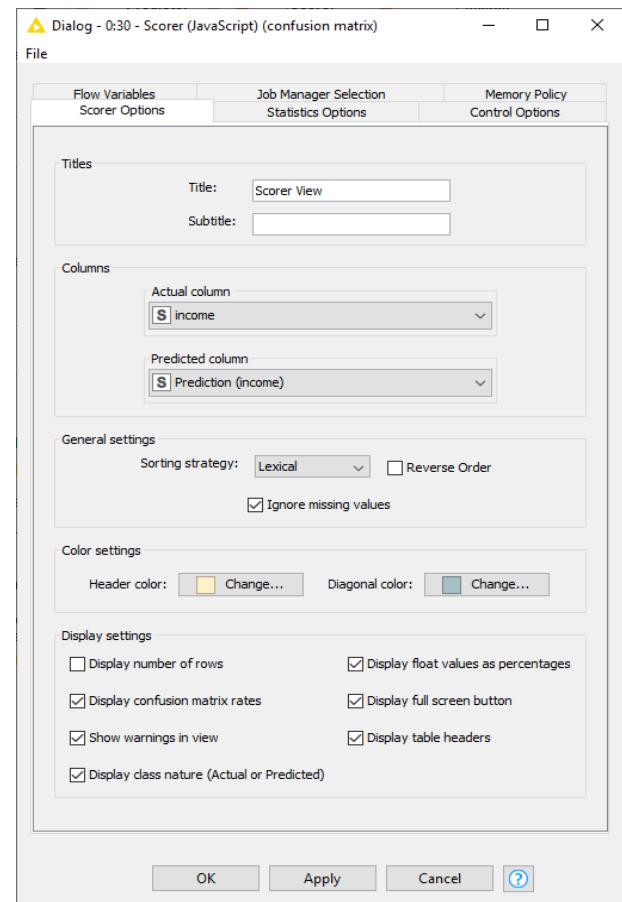
La ventana de configuración tiene tres pestañas: "Opciones del calificador", "Opciones de estadísticas", "Opciones de control".

La pestaña "Scorer Options" requiere la selección de las dos columnas a comparar ("Actual Column" y "Predicted Column") y la ordenación que se utilizará en la estrategia de evaluación. La bandera "Ignorar valores perdidos", si no está marcada, hace que el nodo falle si se encuentran valores perdidos en una de las dos columnas a comparar. Todas las demás opciones están relacionadas con la visualización de la vista del nodo.

La pestaña "Opciones de estadísticas" incluye todas las medidas de precisión y los números falsos/verdaderos positivos/negativos a calcular.

Como todos los nodos basados en JavaScript, este nodo produce una vista con cierto grado de interactividad. Las opciones de interactividad se definen en la pestaña "Opciones de control".

4.15. Ventana de configuración del nodo "Scorer (Javascript)"



Añadimos un nodo "Scorer (Javascript)" en el flujo de trabajo ""My First Mode". El nodo está conectado al puerto de salida de datos del "Naïve Bayes Predictor". La primera columna con los valores de referencia originales es "ingresos"; la segunda columna con la estimación de la clase es la columna llamada "Predicción (ingresos)" que es producida por el nodo "Naïve Bayes Predictor". Durante la ejecución, los valores se comparan fila por fila y se calcula la matriz de confusión y las consiguientes medidas de precisión.

Podemos ver la matriz de confusión y las medidas de precisión de las columnas comparadas seleccionando los tres últimos elementos o el elemento "Interactive View: Confusion Matrix" en el menú contextual del nodo "Scorer (Javascript)"..

### Matriz de confusión (Confusion Matrix)

En la Figura 4.16, se puede ver la matriz de confusión generada por el nodo "Scorer". La matriz de confusión muestra el número de coincidencias entre los valores de la columna objetivo y los valores de la columna predicha.

Los valores encontrados en la columna objetivo se reportan como IDs de fila; los valores encontrados en la columna predicha se reportan como encabezados de columna. Dado que "ingresos" sólo tiene dos valores posibles - ">50K" y "<=50K" - la lectura de la matriz de confusión es bastante sencilla.

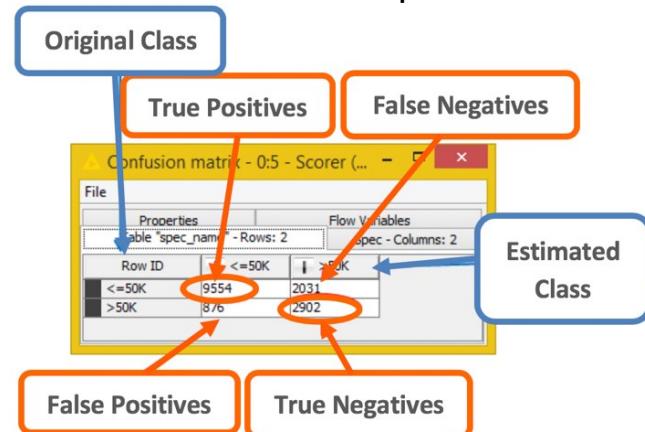
La primera celda contiene el número de filas de datos que tenían un ingreso "<=50K" y que fueron clasificadas correctamente como con un ingreso "<=50K". La última celda, la identificada como (">50K", ">50K"), contiene el número de filas de datos con una renta ">50K" y que se clasificaron correctamente como con una renta ">50K". Las otras dos celdas representan el número de filas de datos con una renta original "<=50K" y que fueron clasificadas incorrectamente como con una renta ">50K" y viceversa.

Las celdas a lo largo de la diagonal desde la esquina superior izquierda hasta la esquina inferior derecha contienen el número de eventos clasificados correctamente. La diagonal opuesta, la que va de la esquina superior derecha a la esquina inferior izquierda, contiene el número de sucesos clasificados incorrectamente, es decir, los errores que queremos minimizar.

La suma de una fila de la matriz de confusión indica el número total de filas de datos en una clase según las etiquetas del conjunto de datos original. La suma de una columna indica el número de filas de datos asignadas a una clase por el modelo. Por lo tanto, la suma de todas las columnas y la suma de todas las filas deben ser iguales, ya que representan el número total de datos.

En nuestro nodo "Scorer", seleccionamos la primera columna como columna de clasificación objetivo "ingresos" y la segunda columna como columna de salida del clasificador bayesiano. Así, esta matriz de confusión dice que 9554 filas de datos fueron clasificadas

**4.16. Verdaderos positivos, falsos negativos, verdaderos negativos y falsos positivos en la matriz de confusión para "<=50K" como clase positiva**



correctamente como de ingresos " $\leq 50K$ "; 2902 fueron clasificadas correctamente como de ingresos " $>50K$ "; y 876 y 2031 filas de datos fueron clasificadas incorrectamente.

### Precisión (Accuracy Measures)

El segundo puerto del nodo "Scorer" presenta una serie de medidas de precisión [6] [7]. En una clasificación binaria (o en cualquier clasificación), tenemos que elegir una de las clases como la clase positiva. Esta elección es completamente arbitraria y suele estar dictada por el contexto de los datos. Una vez que se ha asumido una de las clases como la positiva, pueden tener lugar las siguientes definiciones:

**Verdaderos positivos** es el número de filas de datos que pertenecen a la clase positiva en el conjunto de datos original y que se clasifican correctamente como pertenecientes a esa clase.

**Verdaderos Negativos** es el número de filas de datos que no pertenecen a la clase positiva en el conjunto de datos original y se clasifican como no pertenecientes a esa clase.

**Falsos positivos** son el número de filas de datos que no pertenecen a la clase positiva pero que se clasifican como si lo hicieran.

**Falsos negativos** son el número de filas de datos que pertenecen a la clase positiva pero que el modelo asigna a una clase diferente.

En nuestro caso, si elegimos arbitrariamente " $\leq 50K$ " como clase positiva, los Verdaderos Positivos están en la primera celda, identificada por (" $\leq 50K$ ", " $\leq 50K$ "); los Falsos Negativos están en la celda adyacente; los Falsos Positivos están debajo de ella; y los Verdaderos Negativos están en la celda diagonal restante.

Sobre la base de estos números de Verdaderos Positivos (TP), Verdaderos Negativos (TN), Falsos Positivos (FP) y Falsos Negativos (FN), se pueden definir una serie de medidas de corrección, cada una de las cuales mejora algún aspecto de la corrección de la tarea de clasificación.

Estas medidas de exactitud se proporcionan en el puerto de salida inferior del nodo "Scorer". Let's see how they are defined.

$$\text{Sensitivity} = \text{True Positives} / (\text{True Positives} + \text{False Negatives})$$

$$\text{Specificity} = \text{True Negatives} / (\text{True Negatives} + \text{False Positives})$$

"Sensitivity" mide la capacidad del modelo para reconocer correctamente una clase. Si todas las instancias de una clase determinada se reconocen correctamente, el resultado es 0 "Falsos Negativos" para esa clase; lo que significa que ningún elemento de esa clase se asigna a otra clase. La "sensitivity" es entonces de 1,0 para esa clase.

"Specificity" mide la capacidad del modelo para reconocer lo que no pertenece a una clase determinada. Si el modelo reconoce lo que no pertenece a esa clase, el resultado es 0 "Falsos Positivos"; lo que significa que no hay filas de datos extraños mal clasificados en mi clase.

En un problema de dos clases, la "Sensibilidad" y la "Especificidad" se utilizan para trazar las Curvas ROC (ver "Curva ROC" más adelante en esta sección).

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) = \text{Sensitivity}$$

$$\text{Precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives})$$

"Precision" y "Recall" son dos medidas de precisión estadística muy utilizadas. "Precision" puede considerarse una medida de exactitud o fidelidad, mientras que "Recall" es una medida de exhaustividad.

En una tarea de clasificación, la "Precision" de una clase es el número de "Verdaderos Positivos" (es decir, el número de elementos etiquetados correctamente como pertenecientes a esa clase) dividido por el número total de elementos etiquetados como pertenecientes a esa clase. "Recall" se define como el número de "True Positives" dividido por el número total de elementos que realmente pertenecen a esa clase. "Recall" tiene la misma definición que "Sensitivity".

$$\text{F-measure} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$$

La medida F puede interpretarse como una media ponderada de "Precision" y "Recall", donde la medida F alcanza su mejor valor en 1 y la peor puntuación en 0.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

siendo TP = True Positives, FP = False Positives, TN = True Negatives, and FN = False Negatives.

**Cohen's Kappa** es una medida de la concordancia entre evaluadores como  $((\text{Accuracy} - P(\text{chance})) / (1 - P(\text{chance}))$  donde  $P(\text{chance})$  es la probabilidad media de clasificar los eventos como positivos o negativos, ponderada por la respectiva probabilidad a priori de la clase positiva y negativa. La kappa de Cohen ofrece una estimación de la precisión más equilibrada en caso de fuertes diferencias en las distribuciones de las clases.

"Accuracy" es una medida global y se calcula para todas las clases. Una precisión de 1,0 significa que los valores clasificados son exactamente iguales a los valores de la clase original.

Todas estas medidas de precisión se recogen en la tabla de datos del segundo puerto en la parte inferior del nodo "Scorer (Javascript)" y nos dan información sobre la corrección y la integridad de nuestro modelo.

4.17. EsEstadísticas de precision del nodo "Scorer (Javascript)" para cada clase											
Accuracy statistics - 0.5 - Scorer (confusion matrix)											
File											
Table "default" - Rows: 3   Spec - Columns: 11   Properties   Flow Variables											
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
<=50K	9554	876	2902	2031	0.825	0.916	0.825	0.768	0.868	?	?
>50K	2902	2031	9554	876	0.768	0.588	0.768	0.825	0.666	?	?
Overall	?	?	?	?	?	?	?	?	?	0.811	0.537

## Ver Confusion Matrix

El menú contextual del nodo "Scorer" ofrece 2 posibilidades para visualizar la matriz de confusión:

Elemento "Interactive View: Confusion Matrix"

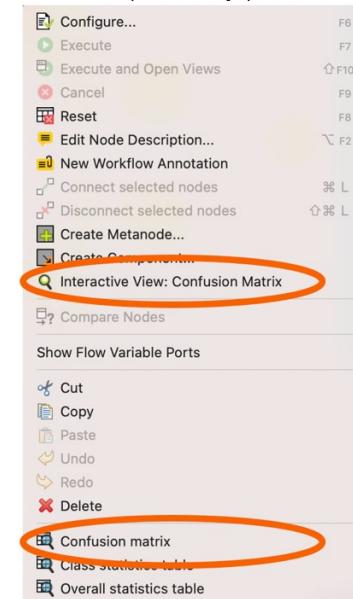
Elemento "Confusion Matrix"

Estos dos elementos conducen a una visualización ligeramente diferente de la misma matriz de confusión. El último elemento lleva a la visualización de la matriz de confusión que hemos visto anteriormente. El primer elemento "Interactive View: Confusion Matrix" incluye algunas opciones más. Veamos la ventana "Vista interactiva: Matriz de confusión".

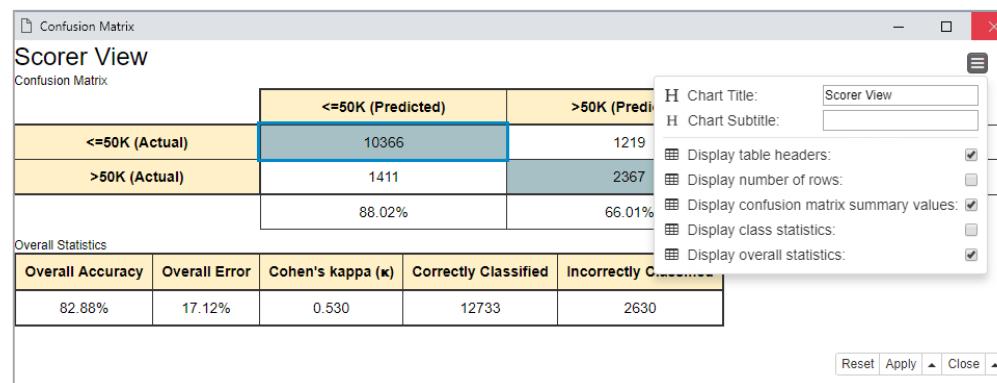
La Vista incluye un título y un subtítulo -como se definió en la ventana de configuración- en la esquina superior izquierda. A continuación, justo debajo aparecen la matriz de confusión, las estadísticas de rendimiento por clase y las medidas estadísticas globales, al menos las que se han seleccionado en la pestaña "Statistics Options" de la ventana de configuración.

Como todos los nodos basados en JavaScript en KNIME Analytics Platform, también este nodo incluye un botón de menú en la esquina superior derecha. El menú que se abre, tras hacer clic en este botón, permite cambiar la visualización de la vista, como el título y el subtítulo, pero también incluir (o no) valores de resumen, estadísticas de clase, estadísticas globales, etc. Por último, se puede seleccionar el contenido de las celdas de la matriz de confusión.

4.18. Menú de contexto del nodo "Scorer (Javascript)"



4.19. Matriz de confusión desde la vista interactiva del nodo "Scorer"



# Arboles de Decisión (Decision Tree)

Utilizando el mismo flujo de trabajo "My First Model", apliquemos ahora otro clasificador bastante popular: un árbol de decisión [8] [9].

El algoritmo del árbol de decisión es un algoritmo supervisado y, por lo tanto, consta de dos fases -entrenamiento y prueba- como el clasificador Naïve Bayes que hemos visto en la sección anterior. El árbol de decisión se implementa en KNIME con dos nodos: uno para el entrenamiento y otro para la prueba, es decir:

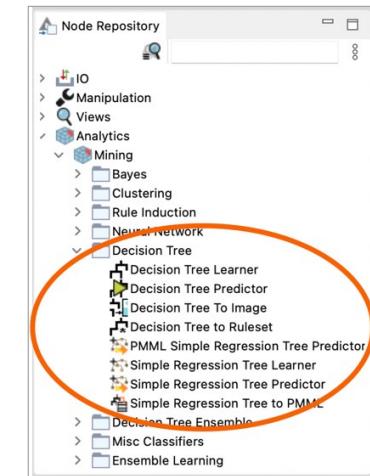
- El nodo "Decision Tree Learner" node
- El nodo "Decision Tree Predictor" node

El nodo "Decision Tree Learner" toma un conjunto de datos como entrada (triángulo negro), aprende las reglas necesarias para realizar la tarea deseada y produce el modelo final en el puerto de salida (cuadrado azul). Conectemos un nodo "Decision Tree Learner" al nodo "CSV Reader" llamado "training set".

Creemos también un nodo "Decision Tree Predictor" que siga al nodo "Decision Tree Learner". El nodo "Decision Tree Predictor" tiene dos entradas:

- Una entrada de datos (triángulo negro) con los nuevos datos a clasificar
- Una entrada de modelo (cuadrado azul) con los parámetros del modelo producidos por un nodo "Decision Tree Learner".

## 4.20. Dos nodos implementan un Árbol de Decisión: el "Decision Tree Learner" y el "Decision Tree Predictor"



# Pestaña del nodo Decision Tree Learner: “Options”

El nodo "Decision Tree Learner" construye un árbol de decisión a partir de los datos de entrenamiento de entrada. En la ventana de configuración hay que especificar:

## General

La *class column*. El atributo de destino debe ser nominal (String).

La *quality measure* para el cálculo de la división: "Índice de Gini" o "Ratio de ganancia".

El *pruning method*: "No Pruning" o una poda basada en el principio de "Minimum Description Length (MDL)" [8] [9]. La opción "Reduced Error Pruning", si está marcada, aplica una poda simple de posprocesamiento.

El *stopping criterion* : el número mínimo de registros en cada nodo del árbol de decisión. Si un nodo tiene menos registros que este número mínimo, el algoritmo detiene la división de esta rama. Cuanto mayor sea el número, menos profundo será el árbol.

El número de registros a almacenar para la vista: el número máximo de filas a almacenar para la funcionalidad de la "hilite". Un número elevado ralentiza la ejecución del algoritmo.

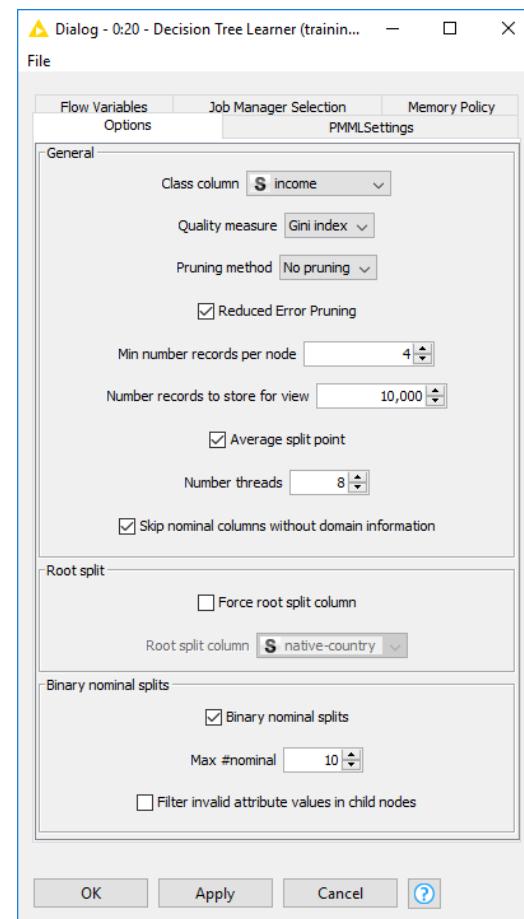
La casilla "Average Split Point". Para los atributos numéricos, el usuario debe elegir una de las dos estrategias de división:

- El punto de división se calcula como el valor medio entre los atributos de las dos particiones (casilla "Average Split Point" activada)
- El punto de división se fija en el mayor valor de la partición inferior (indicador "Average Split Point" desactivado)

El número de hilos en los que se ejecutará el nodo (número de hilos por defecto = 2 \* número de procesadores disponibles para KNIME).

**Root Split** : Si sabe que un atributo debe ser importante para la clasificación, puede forzarlo en el nodo raíz del árbol, activando "Force root split column" y seleccionando la "Root split column".

4.21. Pestaña del nodo Decision Tree Learner:  
“Options”



**Binary nominal splits (divisiones nominales binarias ):** Aquí puede definir si las divisiones nominales binarias se aplican a los atributos nominales. En este caso, puede establecer el umbral Número máximo de divisiones nominales, hasta el cual se calcula una división precisa en lugar de sólo una heurística. La heurística, aunque menos precisa, reduce la carga computacional. "Filtrar valores de atributos no válidos..." inspecciona el árbol al final del procedimiento de formación y elimina posibles duplicados e incongruencias.

## Nodo “Decision Tree Learner”: ventana “PMML Settings”

La ventana de configuración del nodo "Decision Tree Learner" ofrece dos pestañas: "Opciones" (descritas anteriormente) y "Ajustes PMML". La pestaña "Ajustes PMML" se ocupa de la configuración del modelo PMML final.

Cómo afrontar el problema de no tener hijos de verdad

A veces, el proceso de evaluación llega a un nodo del árbol para el que el atributo requerido muestra un valor fuera del dominio de entrenamiento. En este caso, para la clase predicha se puede

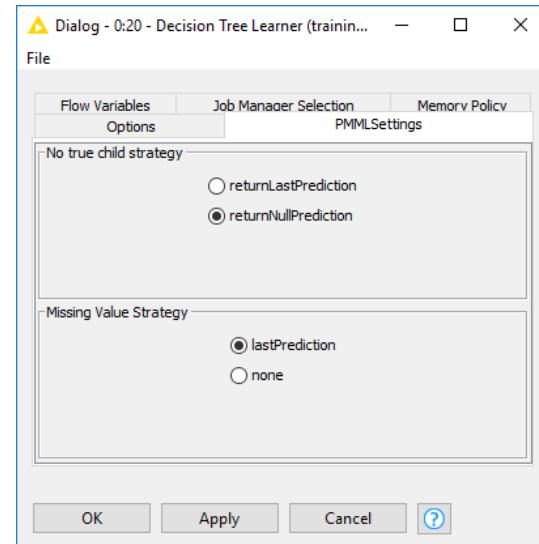
- Utilizar la clase mayoritaria del nodo anterior (opción "returnLastPrediction")
- Devolver un valor perdido (opción "returnNullPrediction")

Cómo tratar los valores perdidos

A veces, el proceso de evaluación llega a un nodo del árbol para el que el atributo requerido muestra un valor ausente. En este caso, para la clase predicha se puede

- Utilizar la clase mayoritaria del nodo anterior (opción "últimaPredicción")
- Volver a la estrategia de no tener hijos verdaderos (opción "ninguno")

4.22. Nodo “Decision Tree Learner”: pestaña “PMML Settings”



Entrenamos el nodo “Decision Tree Learner” node con:

Class column = “income”

Gini Index como medida de calidad

Pruning = No Pruning

Stopping criterion = 4 data points per node

Number of records for hiliting = 10000

Split point calculado como el punto medio entre las dos particiones

Binary splits for nominal values  
ra valores nominales

Número máximo de valores nominales distintos permitidos en una columna = 10

8 como número de hilos, ya que estamos trabajando en una máquina de 4 núcleos

Ahora podemos ejecutar el comando "Ejecutar" y, por tanto, entrenar nuestro modelo de árbol de decisión. Al final de la fase de entrenamiento, el modelo está disponible en el puerto de salida (cuadrado azul) del nodo "Decision Tree Learner".

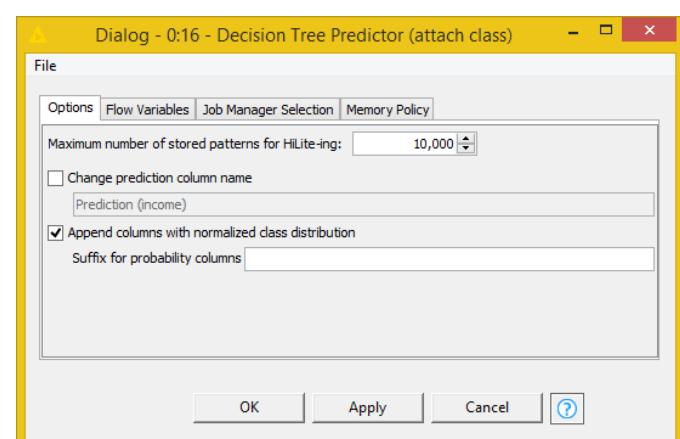
El nodo "Decision Tree Predictor" sólo tiene una tabla de salida, que consiste en el conjunto de datos original con la columna de predicción añadida y, opcionalmente, las columnas con la probabilidad de cada clase, como todos los demás nodos de predicción. El nodo "Decision Tree Predictor" se introdujo para obtener los datos de prueba del nodo ""CSV Reader"" y el modelo del nodo "Decision Tree Learner", con la opción de añadir la distribución de clases normalizada al final de la tabla de datos de predicción.

## Decision Tree Predictor

El nodo "Decision Tree Predictor" importa un modelo de árbol de decisión del puerto de entrada y lo aplica a la tabla de datos de entrada. En la ventana de configuración se puede

- Definir el número máximo de registros para el hilado (de nuevo una herencia de los antiguos nodos de visualización "Data Views")
- Definir un nombre personalizado para la columna de salida con la clase predicha
- Añadir las columnas con la distribución normalizada de cada predicción de clase al conjunto de datos de salida

4.23. Nodo "Decision Tree Predictor"



## Decision Tree Views

En el menú contextual tanto del nodo "Decision Tree Predictor" como del nodo "Decision Tree Learner", podemos ver dos opciones para visualizar las reglas del árbol de decisión:

- "View: Decision Tree View (simple)"
- "View: Decision Tree View "

4.24. Tabla de datos de salida del nodo "Decision Tree Predictor".

Sub-category "Mining" → "Decision Tree" también incluye un nodo "Decision Tree To Image" y un nodo "Decision Tree to Ruleset". El nodo "Decision Tree To Image" convierte la vista del modelo de árbol de decisión en una imagen. El nodo "Decision Tree to Ruleset" convierte las divisiones del árbol de decisión en un conjunto de reglas.

Echemos un vistazo a las vistas del árbol de decisión.

La vista más compleja ("View: Decision Tree View") muestra cada rama del árbol de decisión como un rectángulo. Los datos cubiertos por esta rama se muestran dentro del rectángulo y la regla que implementa la rama se muestra en la parte superior del rectángulo.

La vista más sencilla ("View: Decision Tree View (simple)") representa cada rama como una fracción de círculo, donde la fracción indica qué parte de los datos subyacentes está cubierta por la etiqueta asignada por la regla correspondiente. La regla se muestra al lado de la rama.

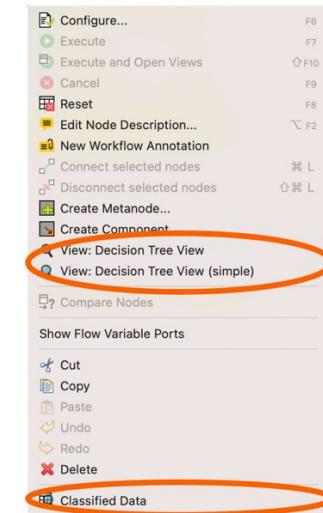
En ambas vistas de los árboles de decisión, una rama puede mostrar un pequeño signo "+" que indica la posibilidad de ampliar la rama con más nodos.

El árbol de decisión siempre comienza con una rama "Root" que contiene todos los datos de entrenamiento. La rama "Root" de nuestro árbol de decisión está etiquetada como "<=50K", porque abarca 11490 registros "<=50K" de los 153362 del conjunto de entrenamiento. Se utiliza un voto mayoritario para etiquetar cada rama del árbol de decisión. En la vista más sencilla, el factor de cobertura de la etiqueta se visualiza mediante el círculo dibujado en el lado de la rama. La rama "Root" tiene un círculo de  $\frac{3}{4}$ , lo que significa que su etiqueta cubre tres cuartas partes de los registros de entrenamiento entrantes.

La primera división se produce en la columna "relationship". A partir de la rama "Root", las filas de datos se separan en un número de sub-ramas según su valor del atributo "relationship". Cada rama se etiqueta con una clase predicha procedente de su factor de cobertura. Por ejemplo, la rama definida por la condición de división "relationship = Wife, Husband" se etiqueta como "<=50K" ya que durante la fase de entrenamiento cubrió 3788 registros "<=50K" de sus 7074 patrones de entrenamiento entrantes. Los valores fraccionados en el número total de patrones de entrenamiento pueden ocurrir cuando se encuentran valores perdidos durante el entrenamiento. En este caso, sólo se pasan fracciones de los patrones por las siguientes ramas.

The screenshot shows two views side-by-side. The left view is titled 'Normalized class distributions' and displays a table of data with columns: Row ID, Urs-p..., \$ native..., \$ income, D P (income=<=50K), D P (income=>50K), and \$ Prediction (income). The right view is titled 'Final prediction' and also displays a table of data with similar columns. Both tables show rows for various entries like 'Row0 United-States <=50K 0.333 0.667 >50K', 'Row1 United-States <=50K 0.962 0.038 <=50K', etc. Red boxes highlight the column headers '\$ Prediction (income)' in both tables.

#### 4.25. menucontextual del nodo "Decision Tree Predictor"



Dentro de cada rama se realizan más divisiones y las filas de datos se separan en diferentes ramas y así sucesivamente, cada vez más profundo en el árbol, hasta las hojas finales. Las hojas finales producen la predicción/clase final.

En las vistas del árbol de decisión se puede seleccionar una rama del árbol de decisión haciendo clic en ella. Las ramas seleccionadas se muestran con un borde rectangular negro (vista simple) o con un fondo más oscuro (vista compleja). No es posible seleccionar varias ramas.

La ventana "Decision Tree View" tiene un menú superior con tres elementos.

"**File**" tiene las opciones:

- "Always on top" garantiza que esta ventana esté siempre visible
- "Export as PNG" exporta esta ventana como imagen para utilizarla en un informe, por ejemplo
- "Close" cierra LA VENTANA

"**Hilite**" contiene los comandos de hilera para trabajar conjuntamente con los nodos "Data Views" ..

"**Tree**" ofrece los comandos para expandir y colapsar las ramas del árbol:

- "Expand Selected Branch" abre las sub-ramas, si las hay, de una rama seleccionada del árbol
- "Collapse Selected Branch" cierra las sub-ramas, si las hay, de una rama seleccionada del árbol

En el lado derecho, hay una visión general del árbol de decisión. Esto es especialmente útil si el árbol de decisión es grande y muy tupido. En el mismo panel, en la parte inferior, hay una función de zoom para explorar el árbol con la resolución más adecuada. Los nodos de la vista del árbol de decisión pueden colorearse utilizando un nodo "Gestor de colores". En la figura 4.26 se muestra el árbol de decisión final con las distribuciones de color para los hombres (azul) y las mujeres (rojo).

El nodo "Scorer", al final, mide también el rendimiento del árbol de decisión, que alcanza un 83% de precisión y un 53% de kappa de Cohen. El clasificador Naive Bayes obtuvo un 81% de precisión y un 54% de kappa de Cohen. Esto significa que los resultados de los dos modelos son comparables, aunque el árbol de decisión funciona ligeramente mejor en una de las dos clases, probablemente la más poblada..

Otra posible visualización para un árbol de decisión consiste en la vista interactiva producida por el nodo "Decision Tree View". Este es otro de los nodos de visualización basados en Javascript y está dedicado a visualizar las divisiones en un modelo de árbol de decisión (Fig. 4.27). Otra posible evaluación del rendimiento del modelo podría lograrse a través de una curva ROC. En realidad, ambos modelos, el Bayes ingenuo y el árbol de decisión, podrían ser evaluados y comparados por medio de una curva ROC. Por supuesto, existe un nodo basado en Javascript que produce una visualización interactiva de una serie de curvas ROC. Para dibujar una curva ROC, la columna de clasificación objetivo tiene que contener sólo dos etiquetas de clase. Una de ellas se identifica como la clase positiva. A continuación, se

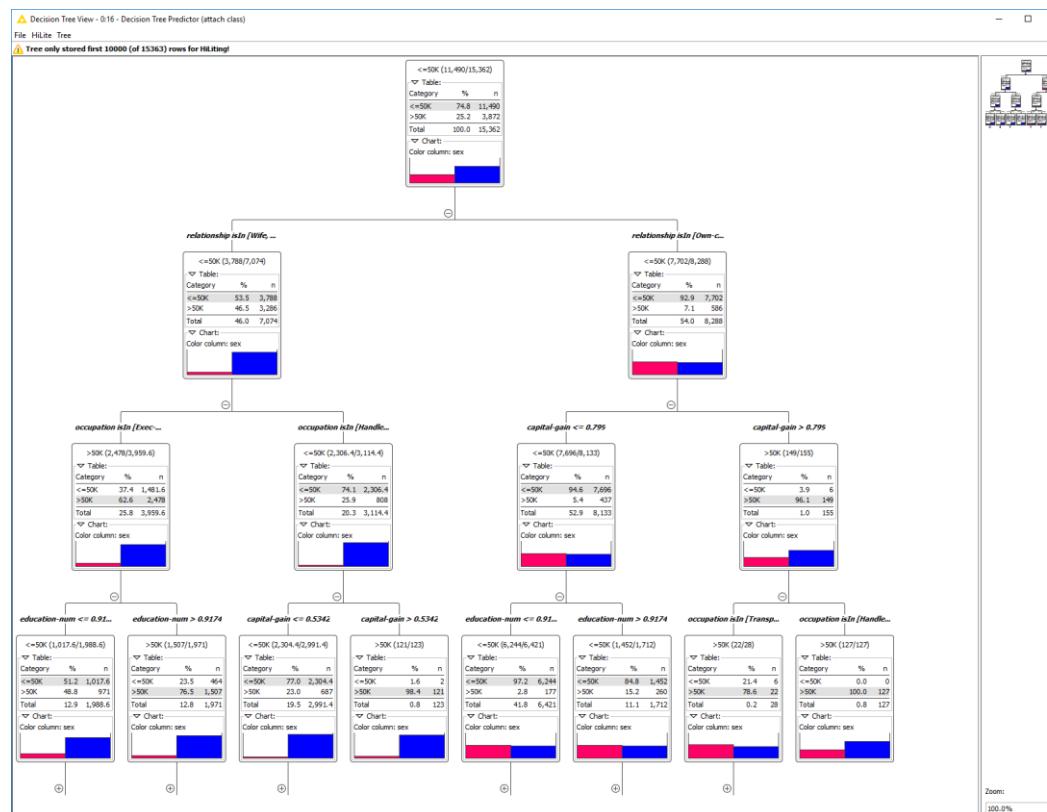
aplica un umbral de forma incremental a la columna que contiene las probabilidades de la clase positiva, definiendo así la tasa de verdaderos positivos y la tasa de falsos positivos para cada valor de umbral [5].

En cada paso, la clasificación se realiza como:

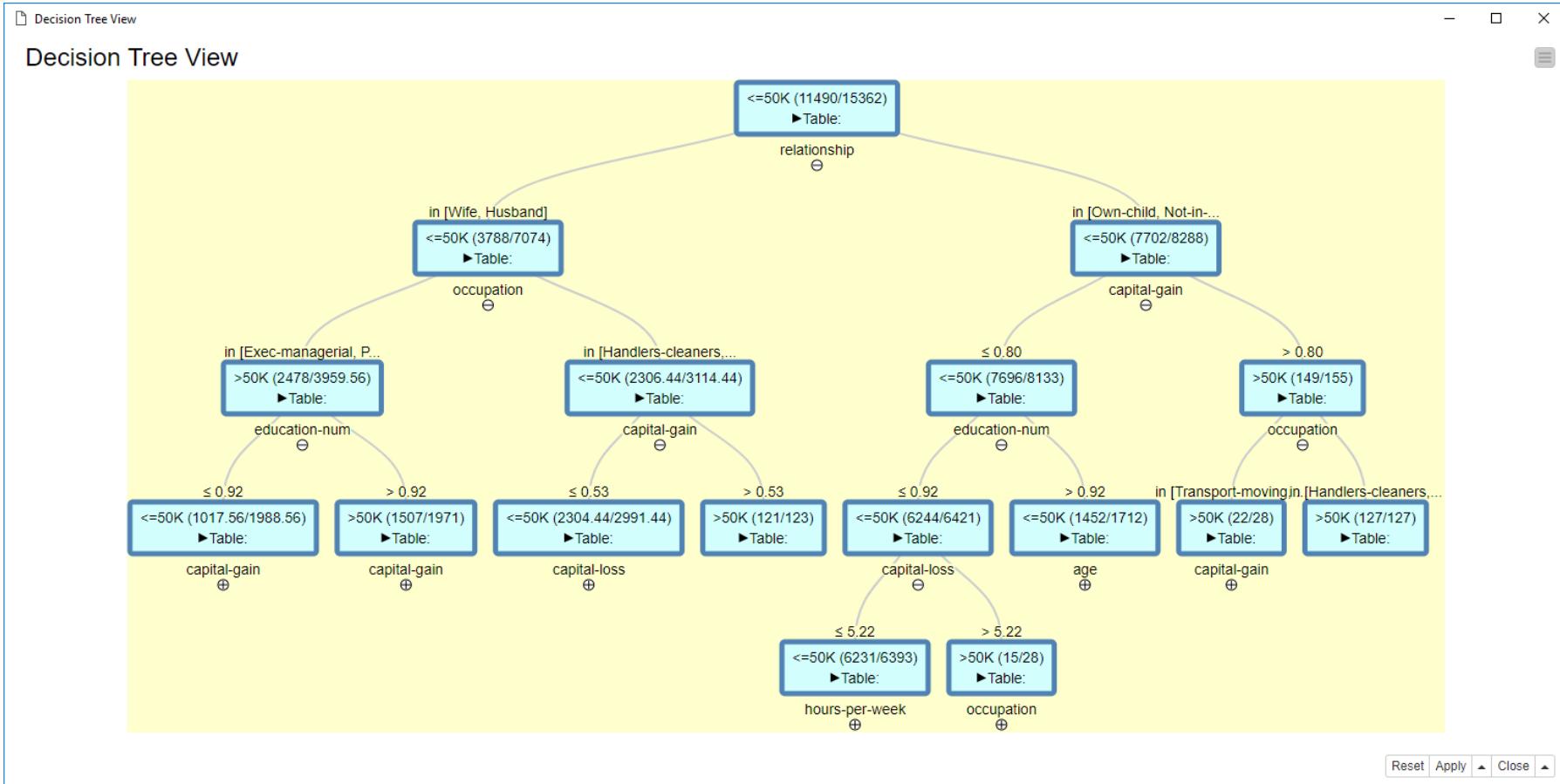
IF Probability of positive class > threshold	=> positive class
ELSE	=> negative class

La curva ROC, en particular el área bajo la curva ROC, da una indicación del rendimiento del predictor. Es posible mostrar múltiples curvas para diferentes columnas en la vista de la curva ROC, si queremos comparar el rendimiento de más de un clasificador. En la figura 4.29 podemos ver que los dos modelos de clasificación tienen rendimientos muy similares (Área bajo la curva = 89% para Naïve Bayes, Área bajo la curva = 85% para el árbol de decisión).

**4.26. Vista del modelo de árbol de decisión creado por el nodo "Decision Tree Learner"**



4.27. Vista del modelo de árbol de decisión producido por el nodo "Decision Tree View"



## Decision Tree View

Como para todos los nodos de visualización basados en Javascript, la ventana de configuración de este nodo contiene tres pestañas: "Decision Tree Plot Options", "General Plot Options", and "View Controls".

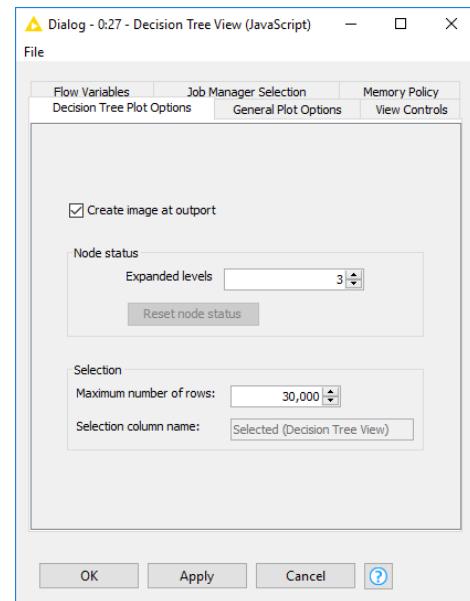
"**Decision Tree Plot Options**" define el contenido a graficar, como el número de filas y el número de niveles a expandir ya en la apertura de la vista. Por supuesto, cuanto mayor sea el número de filas a visualizar, más lenta será la ejecución del nodo. También contiene la bandera para crear una imagen a partir de la vista producida.

"**Decision Tree Plot Options**" define las propiedades generales del gráfico, como el color de fondo, el color de fondo del área del árbol, el color del nodo, el título y el subtítulo, y el formato.

"**View Controls**" establece la interactividad de la trama, como el zoom y la edición de títulos y subtítulos.

La Figura 4.27. muestra una posible vista final del árbol de decisión generado con un nodo "Decision Tree View".

4.28. Ventana de configuración del nodo "Decision Tree View": Pestaña "Decision Tree Plot Options".



En este ejemplo, las predicciones del árbol de decisión y de los algoritmos Naïve Bayes se combinan a través de un nodo "Joiner". Omitiremos los detalles sobre el nodo "Joiner" en este capítulo y explicaremos este nodo con más detalle en el próximo capítulo. Por ahora es suficiente saber que el nodo "Joiner" une las columnas de dos tablas de entrada haciendo coincidir los valores clave de las columnas seleccionadas.

# Curva ROC

El nodo "ROC Curve" dibuja una serie de curvas ROC para un problema de clasificación de dos clases. La ventana de configuración abarca cuatro pestañas: "ROC Curve Settings", "General Plot Options", "Axis Configuration", and "View Controls".

## Configuración de la curva ROC

- La columna que contiene la clase de referencia
- El valor positivo de la clase (asumido arbitrariamente como positivo)
- La(s) columna(s) con las probabilidades de la clase positiva
- El límite del número de puntos a trazar. Recuerde que con menos puntos la curva es menos precisa, con más puntos la ejecución es más lenta.

La selección de las columnas con las probabilidades para la clase positiva se realiza mediante un marco "Excluir"/"Incluir".

## Opciones generales de trazado

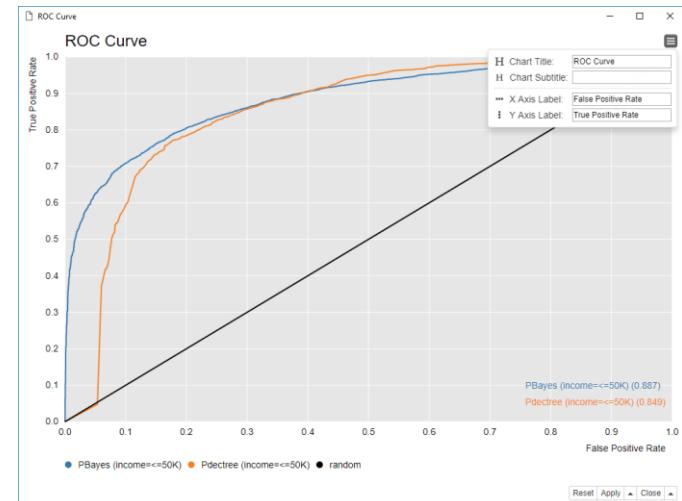
Aquí se requieren todos los ajustes de la parcela: tamaño de la imagen, el formato, los colores de fondo, etc ...

**Axis Control** contiene todos los ajustes para los

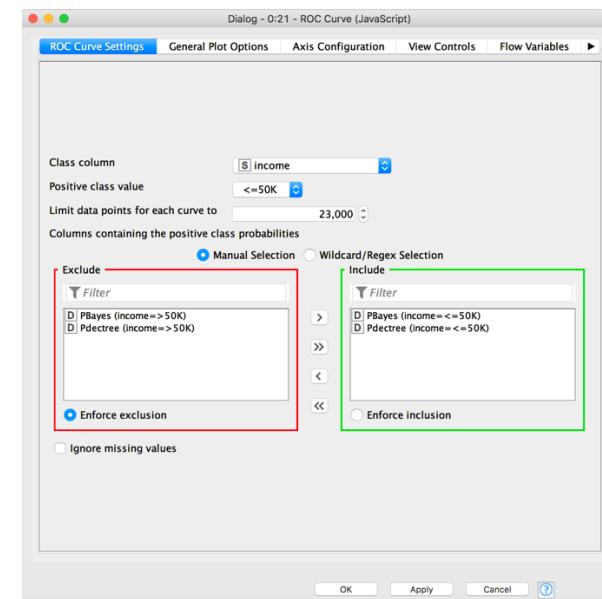
**View Controls** define el nivel de interactividad de la vista de la curva, como la edición de etiquetas o títulos.

El nodo emite la imagen (opcionalmente) de la curva ROC producida y el Área bajo la Curva (AuC) para las columnas de probabilidad.

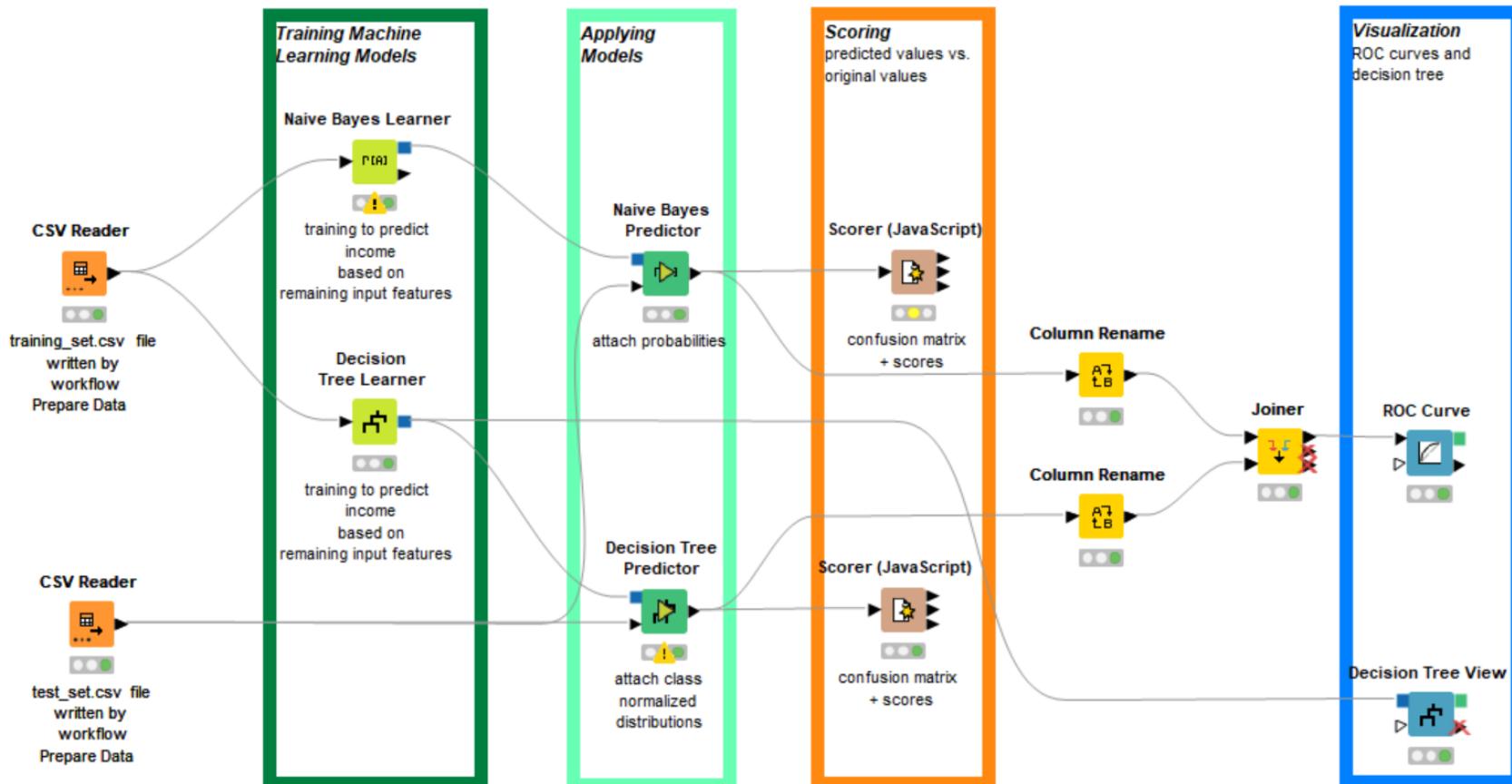
4.29. Vista del nodo "ROC Curve"



4.30. Configuración del nodo "ROC Curve"



#### 4.31. Flujo de trabajo (Workflow) "My First Model"



##### Workflow: My First Model

This workflow reads the data sets written by the Prepare Data workflow.

It then trains two machine learning algorithms:

- Naive Bayes
  - Decision Tree (C4.5)
- to predict Income (>50k / <=50K) based on the remaining input features.

Finally, it scores both models on the test set.

The last section is about visualization: ROC Curves and decision tree visualization.

# Red neuronal artificial (Artificial Neural Network)

Pasamos ahora a una red neuronal y, en concreto, a una arquitectura Perceptrón Multicapa (MLP), con una capa oculta, y al algoritmo de aprendizaje Back Propagation. El paradigma de la red neuronal está disponible en la categoría "Minería" y consiste en:

- Un nodo de aprendizaje ("RProp MLP Learner")
- Un nodo de predicción ("Multilayer Perceptron Predictor")

El nodo aprendiz aprende las reglas para separar los patrones de entrada del conjunto de entrenamiento, los empaqueta en un modelo y los asigna al puerto de salida.

El nodo predictor aplica las reglas del modelo construido por el nodo aprendiz a un conjunto de datos con nuevos registros.

## RProp MLP Learner

El nodo "RProp MLP Learner" construye y entrena un Perceptrón Multicapa con el algoritmo BackPropagation. En la ventana de configuración hay que especificar:

El número máximo de iteraciones para el algoritmo de Back Propagation  
El número de capas ocultas de la arquitectura neuronal  
El número de neuronas por cada capa oculta

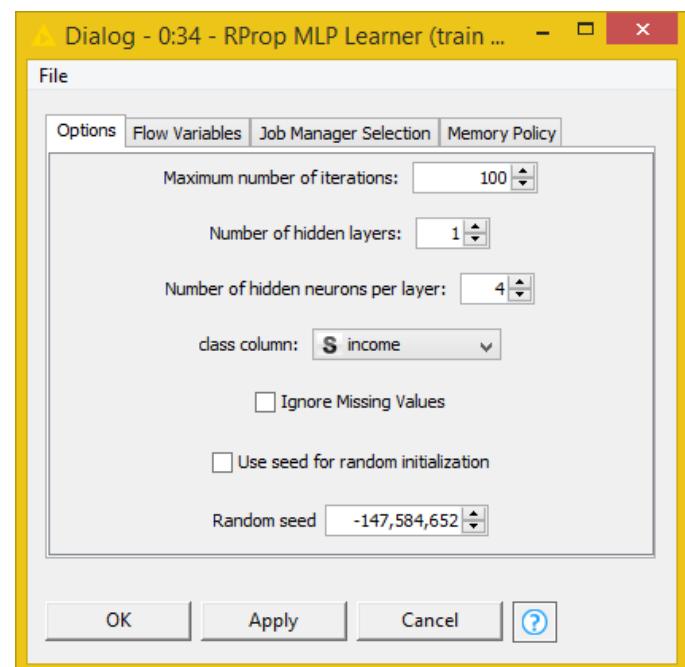
La columna de clases, es decir, la columna que contiene las clases objetivo. La columna de clases tiene que ser de tipo String (valores nominales)

También hay que especificar qué hacer con los valores perdidos. El algoritmo no funciona si hay valores perdidos. Si tiene valores perdidos, debe transformarlos antes en su flujo de trabajo o ignorarlos durante el entrenamiento. Para ignorar los valores que faltan sólo tiene que marcar la casilla correspondiente en la ventana de configuración.

Por último, es necesario especificar una semilla para que la inicialización aleatoria del peso se pueda repetir.

El "RProp MLP Learner" sólo acepta entradas numéricas. Las columnas de datos de cadena no se procesarán como atributos de entrada.

4.32. Configuration window of the „RProp MLP Learner“ node



Creamos un nuevo flujo de trabajo en el grupo de flujos de trabajo "Chapter4" y lo llamamos "My First ANN". También utilizamos los conjuntos de datos "conjunto de entrenamiento" y "conjunto de prueba" derivados del conjunto de datos adult.data en el flujo de trabajo "Data Preparation". Establecimos la tarea de clasificación/predicción para predecir el tipo de ingresos que tiene cada persona/registro. "Income" es una columna de cadena con sólo dos valores: ">50K" y "<=50K".

En primer lugar, insertamos dos nodos "CSV Reader": uno para leer el conjunto de entrenamiento y otro para leer el conjunto de prueba preparado por el flujo de trabajo "Data Preparation" anterior en este capítulo.

De todos los atributos de cadena en el conjunto de datos de adultos, decidimos mantener sólo el atributo "sex", ya que pensamos que el sexo es una variable discriminativa importante para predecir los ingresos de una persona. Por supuesto, también mantuvimos la columna "Income" para que fuera la clase de referencia. Eliminamos todos los demás atributos de cadena.

El atributo "sex", al ser de tipo String, no ha podido ser utilizado tal cual y se ha convertido en una variable binaria "sex\_01", según la siguiente regla:

IF \$sex\$ = "Male"	=>	\$sex_01\$ = "-1"
IF \$sex\$ = "Female"	=>	\$sex_01\$ = "+1"

In order to implement this rule, we used a "Rule Engine" node. "sex\_01" is the newly created Integer column containing the binary values for sex. We then used a "Column Filter" node to exclude all remaining string columns besides "Income".

El Perceptrón Multicapa requiere datos numéricos en el rango [0,1]. Para cumplir con eso, se colocó un nodo "Normalizer" después del nodo "Column Filter" para normalizar todas las columnas de datos numéricos para que caigan en el rango [0,1]. Esta secuencia de transformaciones ("Rule Engine" en "sexo", "Filtro de columna" para mantener sólo los atributos numéricos y la columna de datos "Income", y la normalización [0,1]) se aplicó tanto en el conjunto de entrenamiento como en el de prueba.

A continuación, aplicamos el nodo "RProp MLP Learner" para construir una red neuronal MLP con 6 variables de entrada ((age, education-num, fnlwgt, capital-gain, capital-loss, hours-per-week, sex\_01), 1 capa oculta con 4 neuronas y 2 neuronas de salida, es decir, una neurona de salida para cada clase de "Income". Lo entrenamos con los datos del conjunto de entrenamiento con un número máximo de iteraciones de 100.

Tras el entrenamiento, aplicamos el modelo MLP a los datos del conjunto de prueba utilizando un nodo "Multilayer Perceptron Predictor". El nodo predictor de la red neuronal aplica las reglas del modelo construido por el nodo aprendiz a un conjunto de datos con nuevos registros. El nodo predictor tiene dos puertos de entrada:

- Una entrada de datos (triángulo negro) con los nuevos datos a clasificar
- Una entrada de modelo (cuadrado azul) con los parámetros del modelo producidos por un nodo "RProp MLP Learner".

El nodo predictor tiene un puerto de salida, donde se produce el conjunto de datos originales más las clases predichas y, opcionalmente, las distribuciones de las clases..

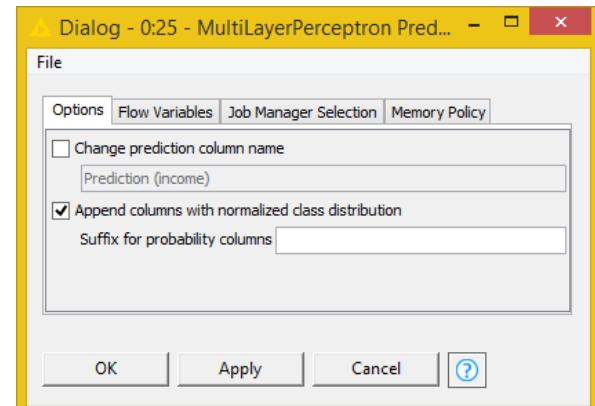
## Multilayer Perceptron Predictor

El nodo "Predictor Perceptrón Multicapa" toma un modelo MLP, generado por un nodo "RProp MLP Learner", en el puerto de entrada del modelo (cuadrado azul) y lo aplica a la tabla de datos de entrada en el puerto de datos de entrada (triángulo negro).

El nodo "Multilayer Perceptron Predictor" se encuentra en el "Node Repository" in the "Analytics" → "Mining" → "Neural Network" → "MLP" category.

Los únicos ajustes necesarios para su configuración, al igual que para todos los demás nodos de predicción, son una casilla de verificación para añadir las distribuciones de clase normalizadas a la tabla de datos de entrada y un posible nombre personalizado para la columna de clase de salida.

4.33. Ventana de configuración del nodo "MultiLayer Perceptron Predictor".



### Datos clasificados(Classified Data)

Visualicemos los resultados de la MLP Prediction:

- Click derecho en el nodo "Multilayer Perceptron Predictor"
- Seleccionar "Classified data"

La tabla de datos "Datos clasificados" contiene las clases finales predichas en la columna "Prediction (income)" y los valores de las dos neuronas de salida en las columnas "P (Income >50K)" y "P(Income<=50K)".

El valor de disparo de las dos neuronas de salida se representa con una barra roja-verde en lugar de un número doble. El rojo significa un número bajo (< 0,5), el verde un número alto (> 0,5). La neurona de salida más alta decide la clase de predicción de la fila de datos.

Para cambiar la representación de los valores de disparo de las neuronas, haga clic con el botón derecho del ratón en la

4.34. Tabla de datos de salida del nodo "Predictor Perceptrón Multicapa"

Row ID	he	sex_01	P (income=<50K)	P (income=>50K)	Prediction (income)
Row0	-0.705				<=50K
Row1	-0.705				<=50K
Row2	1.419				<=50K
Row3	-0.705				<=50K
Row4	-0.705				>50K
Row5	-0.705				<=50K
Row6	-0.705				<=50K
Row7	-0.705				<=50K
Row8	-0.705				<=50K
Row9	1.419				<=50K
Row10	1.419				<=50K
Row11	-0.705				<=50K
Row12	-0.705				>50K
Row13	-0.705				<=50K
Row14	-0.705				<=50K
Row15	-0.705				<=50K
Row16	-0.705				<=50K

cabecera de la columna y seleccione una nueva representación en "Available Renderers", como por ejemplo "Standard Double".

Hemos utilizado el paradigma de las redes neuronales en un problema de clasificación de dos clases ("Income > 50K" or "Income <=50K"). Ahora podemos aplicar un nodo "ROC Curve" a los resultados del nodo "Multilayer Perceptron Predictor". Identificamos:

La columna de clase como columna "Income"

El valor positivo "<=50K" en la columna de clase "Income"

La columna "P(Income <=50K)" como la columna que contiene la probabilidad/puntuación para la clase positiva

La curva ROC resultante muestra un área bajo la curva de alrededor de 0,85.

## Write/Read Models to/from file

Una vez que hayamos entrenado un modelo y comprobado que funciona lo suficientemente bien para nuestras expectativas, sería bueno que pudiéramos reutilizar el mismo modelo en otras aplicaciones similares con nuevos datos. Esto significa que también deberíamos poder reciclar el modelo en otros flujos de trabajo.

### Model Writer

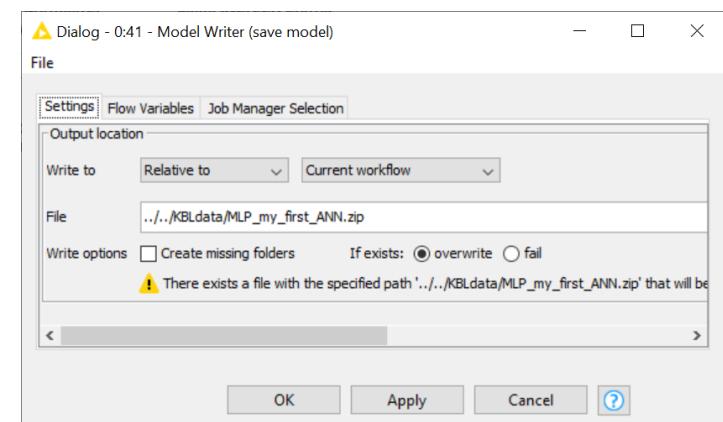
El nodo "Model Writer" toma un modelo en el puerto de entrada (cuadrado gris) y lo escribe en un archivo utilizando el formato interno de KNIME.

El nodo "Model Writer" se encuentra en la categoría "IO" → "Write" en el panel "Node Repository".

La ventana de configuración sólo requiere:

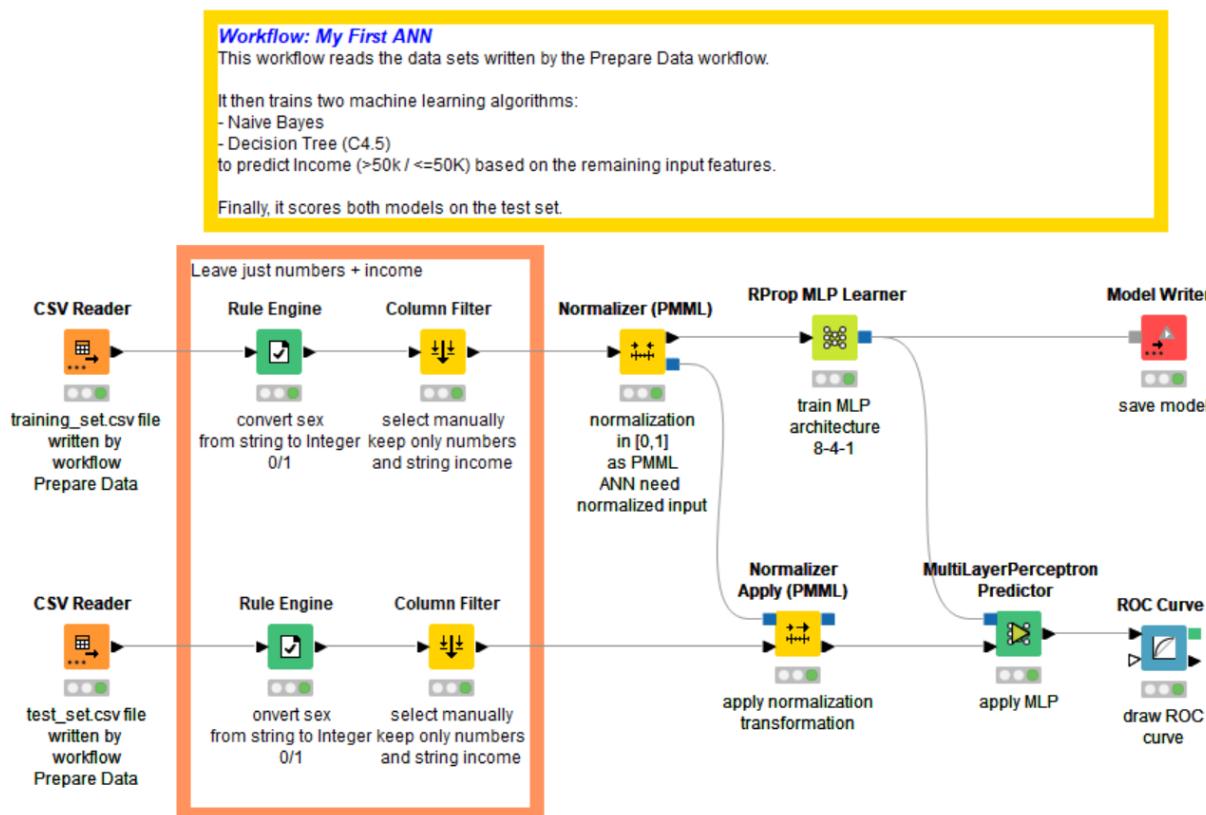
- La ruta del archivo de salida (\*.zip) (knime:// protocol también se acepta)
- La casilla para anular el archivo, si el archivo existe

4.35. Ventana de configuración del nodo "Model Writer"



El flujo de trabajo final "My First ANN" se muestra a continuación.

#### 4.36. My First ANN



Al mismo tiempo, KNIME también proporciona un nodo para leer un modelo desde un archivo, el nodo "Model Reader", situado en la categoría "IO" "Read" en el panel "Node Repository".

## Nodo "Model Reader"

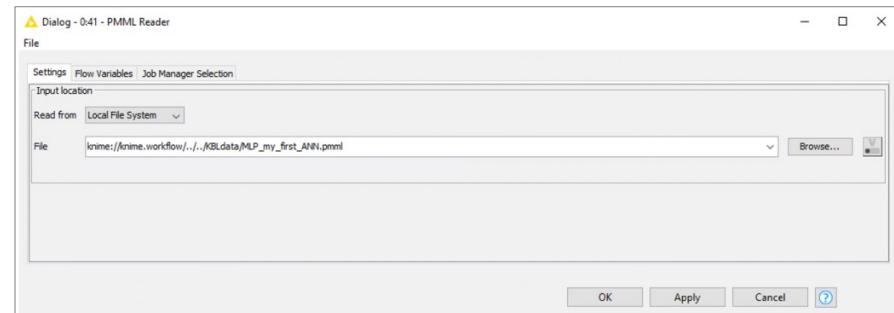
El nodo "Model Reader" lee un modelo de un archivo utilizando el formato interno de KNIME y lo pone a disposición en el puerto de salida (cuadrado gris).

La ventana de configuración sólo necesita

- La ruta del archivo de entrada (\*.pmml) (también se acepta el protocolo knime://)

Al arrastrar y soltar un archivo de modelo desde una carpeta de datos, se crea automáticamente un nodo "Model Reader" con los ajustes de configuración adecuados.

4.37. Ventana de configuración del nodo "Model Reader"



En esta última parte del capítulo, nos gustaría mostrar algunos nodos más que se utilizan habitualmente en el análisis de datos. Construiremos un nuevo flujo de trabajo, llamado "Clustering y Regresión", en el grupo de flujos de trabajo "Capítulo4" para explicar estos nodos.

Utilizaremos los mismos datos que usamos para los dos flujos de trabajo anteriores, "conjunto de entrenamiento" y "conjunto de prueba", creados en el flujo de trabajo "Data Preparation". Los dos primeros nodos del flujo de trabajo serán entonces dos nodos "CSV Reader", uno para leer el conjunto de entrenamiento y otro para leer los datos del conjunto de prueba, como en los flujos de trabajo anteriores.

# Nodo “Statistics”

El nodo “Statistics” calcula variables estadísticas sobre los datos de entrada, como por ejemplo:

Para las columnas numéricas (disponibles en una tabla en el puerto de salida 0):	Para las columnas nominales (disponibles en una tabla en el puerto de salida 1 y 2):
mínimo máximo media Desviacion estandar asimetría  Histograma	varianza mediana Sumatoria total kurtosis number of NaN/calores perdidos

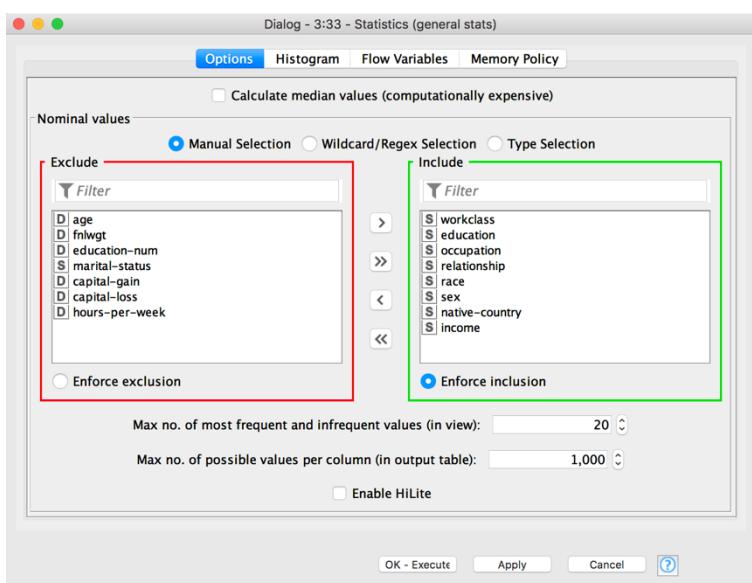
El nodo “Statistics” esta ubicado en el panel “Node Repository” en “Analytics” → “Statistics” category.

The configuration window requires:

- La selección de las columnas nominales sobre las que calcular las medidas estadísticas (las medidas estadísticas para las variables numéricas se calculan sobre todas las columnas numéricas por defecto).
- El número máximo de valores más frecuentes e infrecuentes a mostrar en la vista.
- El número máximo de valores posibles por columna. Esto es para evitar largas listas de valores nominales.
- Si se debe calcular el valor de la mediana

Todas las medidas estadísticas, descritas en la tabla anterior, están disponibles en los puertos de salida del nodo, así como en la Vista del nodo.

La selección de las columnas de datos de entrada se realiza mediante el marco de selección de columnas: por selección manual con paneles “Incluir/Excluir”; por selección de tipo, por selección de expresiones Wildcard/Regex.



El nodo "Statistics" tiene dos opciones de visualización: la "Statistics View" y las tablas de datos en los puertos de salida. Ambas opciones de visualización son accesibles a través del menú contextual.

El nodo tiene tres puertos de salida y la "Statistics View" tiene tres pestañas especulares. El puerto de salida "Statistics Table" corresponde a la pestaña "Numeric" de la vista; el puerto de salida "nominal Statistical Values" corresponde a la pestaña "Nominal" de la vista; y el puerto de salida "TOccurrences Table" corresponde a la pestaña "Top/Bottom" de la vista.

La pestaña "Numeric" contiene una serie de medidas estadísticas calculadas sobre todas las columnas numéricas, con un histograma aproximado. Cada fila con todas las medidas estadísticas y el histograma aproximado ofrece una idea de las propiedades estadísticas y la distribución de los valores en una columna de datos numéricos.

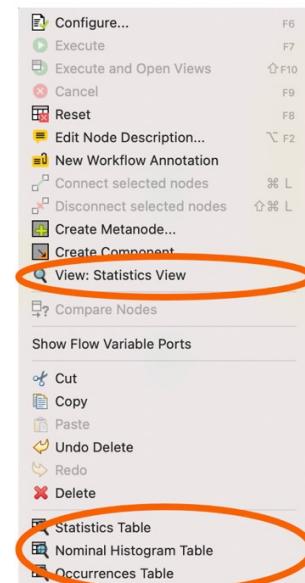
Del mismo modo, las pestañas "Nominal" y "Top/Bottom" dan una idea de las propiedades estadísticas de los valores de una columna de datos nominales.

**Note.** La estadística de las columnas nominales se calcula sólo para las columnas nominales incluidas en la ventana de configuración.

En el nodo de nuestro flujo de trabajo excluimos la columna "marital-status" de las columnas nominales y el par de columnas correspondiente "marital-status" y "marital status\_Count" no está en la "Occurrences Table".

En "Statistics View" → tab "Nominal Columns" encontramos la misma información, pero las listas de valores nominales están ordenadas por frecuencia. Para cada columna encontramos dos celdas de la tabla: una en la parte superior para los valores nominales más frecuentes (los 20 primeros) de la columna y otra en la parte inferior para los valores nominales menos frecuentes (los 20 últimos) de la columna.

#### 4.39. Context menu of the "Statistics" node



#### 4.40. La pestaña "Numeric" de la vista del nodo "Statistics" muestra las medidas estadísticas y el histograma calculado en todas las columnas numéricas.



**4.41. La "Occurrences Table" contiene el número de ocurrencias de los valores nominales calculados sólo en las columnas nominales seleccionadas**

The screenshot shows a table titled "Occurrences Table - 0:29 - Statistics". The table has 16 columns and 41 rows. The columns are labeled: Row ID, workclass, workclass\_Count, education, education\_Count, occupa..., occupa..., and relation... . The first few rows are: Row0: Private, 1203; Row1: Self-emp-no..., 1271; Row2: Local-gov, 1030; Row3: ?, 940; Row4: State-gov, 657; Row5: Self-emp-inc, 568; Row6: Federal-gov, 499; Row7: Without-pay, 6; Row8: Never-worked, 4; Row9: ?, ?; Row10: ?, ?; Row11: ?, ?. The "workclass" and "workclass\_Count" columns are circled in red.

**4.42. "Statistics View" → Tab "Top/Bottom" con el número de ocurrencias de los valores nominales calculados sólo en las columnas nominales seleccionadas y ordenados de forma descendente**

The screenshot shows the "Statistics View - 0:33 - Statistics" window with the "Top/bottom" tab selected. It displays two sets of data: "Top 20:" and "Bottom 20:". The "Top 20:" section lists various categories with their counts, such as Private (11305), HS-grad (5278), Husband (6639), White (13918), Male (10879), United-States (14560), and so on. The "Bottom 20:" section lists categories from least frequent to most frequent, such as Haiti (24), Taiwan (24), Portugal (21), Nicaragua (19), France (19), Iran (17), Ecuador (17), Peru (13), Greece (12), Hong (12), Ireland (11), Laos (10), Outlying-US(Guam-USVI-etc) (8), Trinidad&Tobago (8), Cambodia (8), Thailand (8), Honduras (6), Hungary (6), and Yugoslavia (6).

## Regresión

Otra tarea muy común en el análisis de datos es el cálculo de la regresión lineal [3] [4] [5]. En el panel "Node Repository", en "Analytics" → "Statistics" → "Regression", hay dos nodos de aprendizaje para aprender los parámetros de regresión: un nodo realiza una regresión lineal multivariante, el otro nodo una regresión polinomial multivariante. Ambos nodos de aprendizaje de regresión comparten el nodo predictor. Los nodos de aprendizaje de regresión tienen dos puertos de entrada y dos puertos de salida. En la entrada, el nodo se alimenta con los datos de entrenamiento y, opcionalmente, con un modelo preexistente. Tras la ejecución, el nodo produce el modelo de regresión y las propiedades estadísticas del modelo en una tabla de datos. El nodo predictor toma el modelo de regresión, lineal o polinómico, como entrada y lo aplica a nuevas filas de datos de entrada para predecir su respuesta. En este libro, sólo mostraremos cómo implementar la regresión lineal.

Los modelos que hemos visto hasta ahora eran clasificadores; es decir, intentaban predecir valores nominales (clases) para cada fila de datos. La regresión lineal es un modelo ajustador; es decir, un modelo que intenta predecir valores numéricos. En este caso, la columna de datos objetivo debe ser una columna numérica con valores numéricos que se aproximarán mediante el ajuste de la regresión lineal.

## Nodo “Linear Regression Learner”

El nodo “Linear Regression Learner” realiza una regresión lineal multivariante sobre una columna objetivo, es decir, la respuesta.

El nodo “Linear Regression (Learner)” se encuentra en: “Node Repository” “Analytics” → “Mining” → “Regression”.

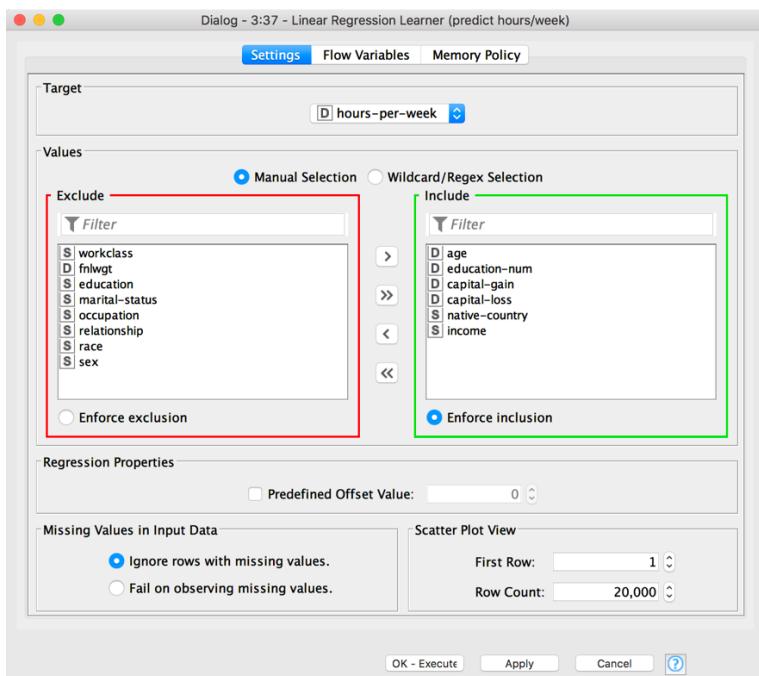
En la ventana de configuración hay que especificar:

- La columna para la que se calcula la regresión
- Las columnas que se utilizarán como variables independientes en la regresión lineal
- El número de la fila inicial y el número de filas que se visualizarán en la vista del gráfico de dispersión del nodo
- La estrategia de gestión de los valores perdidos
- Un valor de desplazamiento por defecto a utilizar (si lo hay)

La selección de las columnas de datos de entrada se realiza mediante el marco de selección de columnas: por selección manual con “Include/Exclude”; por selección de tipo, por selección de expresión Wildcard/Regex.

El nodo produce el modelo de regresión, así como los coeficientes y estadísticas del modelo.

4.43. Configuration window for the “Linear Regression Learner” node



**Nota.** El nodo “Linear Regression Learner” sólo puede tratar con valores numéricos. Las columnas nominales se discretizan automáticamente utilizando una codificación ficticia disponible para las variables categóricas en la regresión  
[http://en.wikipedia.org/wiki/Categorical\\_variable#Categorical\\_variables\\_in\\_regression](http://en.wikipedia.org/wiki/Categorical_variable#Categorical_variables_in_regression).

Conectamos un nodo "Linear Regression Learner" al "CSV Reader" con los datos de entrenamiento. Queremos predecir las columnas "hours-per-week" utilizando las columnas "age", "education-num", "capital-gain", "capital-loss", "native-country", and "income" as independent variables for the linear regression. The "Linear Regression Learner" node produces the regression model at the node's output port. The regression model is subsequently fed into a "Regression Predictor" node and used to predict new values for a different data set.

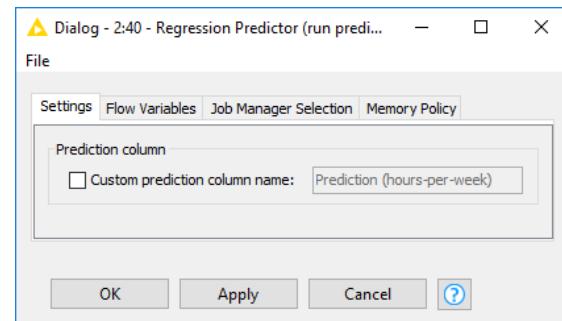
## Nodo "Regression Predictor"

El nodo "Predictor de regresión" obtiene un modelo de regresión de uno de sus puertos de entrada (cuadrado azul) y datos del otro puerto de entrada (triángulo negro). Utiliza el modelo y los datos para hacer una predicción basada en los datos.

Como toda la información ya está disponible en el modelo, este nodo sólo necesita los ajustes mínimos del predictor: un nombre alternativo personalizado para la columna de clasificación de salida.

El nodo "Regression Predictor" se encuentra en la sección "Analytics" → "Mining" → "Regression" del panel "Node Repository"

4.44. Configuration window for the "Regression Predictor"



## Clustering

El último tema que queremos tratar en este capítulo es el clustering. Existen muchas técnicas de clustering y KNIME ha implementado varias de ellas.

Como en los modelos de datos que ya vimos, tenemos un nodo entrenador y un nodo predictor para los modelos de clustering. Los nodos aprendices implementan un algoritmo de clustering; es decir, construyen un número de clusters agrupando patrones similares y calculan sus prototipos representativos. A continuación, el predictor asigna un nuevo vector de datos al clúster con el prototipo más cercano. Un predictor de este tipo no es específico de una sola técnica de clustering, sino que funciona para cualquier algoritmo de clustering que requiera una asignación de cluster sobre la base de una función de distancia en la fase de predicción. Esto conduce a muchos nodos aprendices de clustering específicos (que implementan diferentes procedimientos de clustering) pero a un solo nodo predictor de clustering.

Un nodo aprendiz podría implementar el algoritmo k-Means, por ejemplo. El procedimiento k-Means construye k clusters en los datos de entrenamiento, donde k es un número predefinido [3] [4] [5]. El algoritmo itera varias veces sobre los datos y termina cuando las asignaciones de los clusters ya no cambian. Tenga en cuenta que los k clusters sólo se construyen en base a un criterio de similitud (distancia). k-Means no tiene en cuenta la clase real de cada fila de datos: es un algoritmo de clasificación no supervisado. El predictor

realiza una clasificación crispada que asigna un vector de datos a uno solo de los k clusters que se construyeron con los datos de entrenamiento; en concreto, asigna el vector de datos al cluster con el prototipo más cercano.

Nos centraremos en el algoritmo k-Means para darle un ejemplo de cómo se puede implementar el clustering con KNIME (ver el flujo de trabajo "Clustering y Regresión").

## k-Means

El nodo "k-Means" agrupa los patrones de entrada en k clusters sobre la base de un criterio de distancia y calcula sus prototipos. Los prototipos se construyen como el valor medio de los patrones de los clusters. Este nodo toma los datos de entrenamiento en el puerto de entrada y presenta el modelo en el puerto de salida cuadrado azul y los datos de entrenamiento con la asignación de clusters en el puerto de salida de datos (triángulo negro).

El nodo "k-Means" se encuentra en el "Node Repository" en "Analytics" → "Mining" → "Clustering".

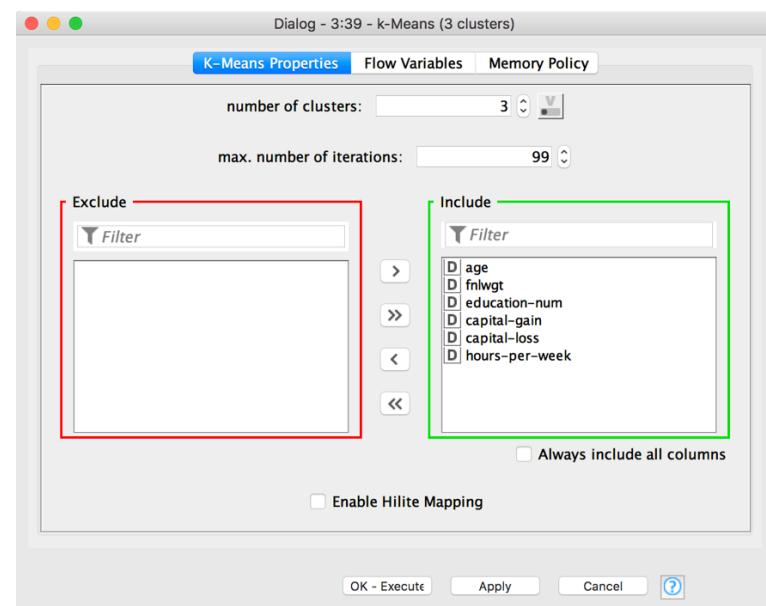
In the configuration window you need to specify:

- El número final de clusters k
- El número máximo de iteraciones para garantizar que la operación de aprendizaje converge en un tiempo razonable
- Las columnas que se utilizarán para calcular la distancia y los prototipos
- casilla "Always include all columns" es una alternativa al marco de selección de columnas.

La selección de columnas se realiza mediante un cuadro "Exclude"/"Include"

- Las columnas que se utilizarán para el cálculo de la distancia se enumeran en el marco "Include". Todas las demás columnas se enumeran en el marco "Exclude"
- Para pasar del marco "Include" al marco "Exclude" y viceversa, utilice los botones "add" y "remove". Para mover todas las

4.45. Ventana de configuración del nodo "k-Means"



columnas a un marco u otro, utilice los botones "add all" y "remove all".

El nodo "k-Means" tiene una opción "Cluster View" en el menú contextual: "View: Cluster View". La vista de clústeres muestra los prototipos de los k clústeres.

**Note.** Dado que los algoritmos de clustering se basan en la distancia, suele ser necesaria una normalización para que todos los rangos de características sean comparables. En el flujo de trabajo "Clustering y Regresión", normalizamos todas las características de entrada en [0,1] utilizando un nodo "Normalizador".

Sin embargo, el algoritmo k-Means sólo define los clusters en el espacio de entrada sobre la base de un subconjunto representativo del mismo espacio de entrada. Una vez definido el conjunto de clusters, las nuevas filas de datos necesitan ser puntuadas con respecto a él para encontrar el cluster al que pertenecen. Para ello, utilizamos el nodo "Cluster Assigner".

## Nodo "Cluster Assigner"

El nodo "Cluster Assigner" asigna los datos de la prueba a un conjunto existente de prototipos que han sido calculados por un nodo de clustering como el nodo "k-Means". Cada fila de datos se asigna a su prototipo más cercano.

El nodo toma un modelo de clustering y un conjunto de datos como entradas y produce una copia del conjunto de datos con una columna adicional que contiene las asignaciones de cluster.

El nodo "Cluster Assigner" se encuentra en la categoría "Analytics" → "Mining" → "Clustering" en el panel "Node Repository".

No necesita ningún ajuste de configuración específico para su tarea de asignación de clústeres.

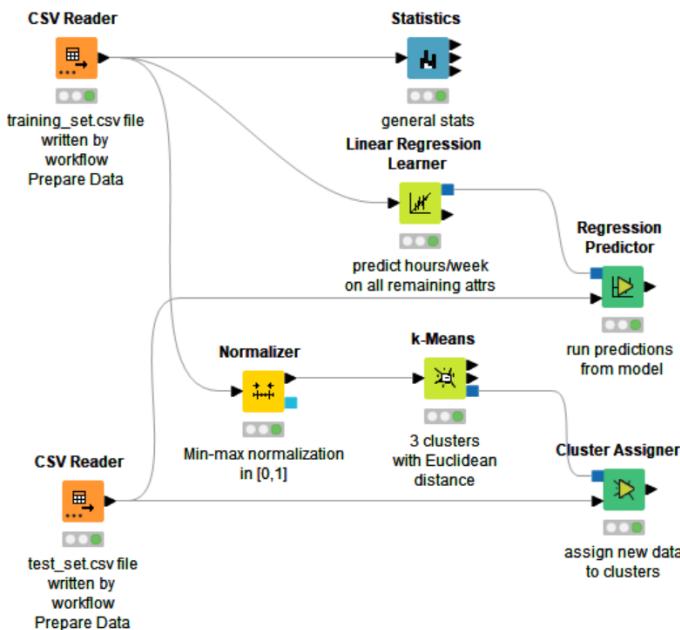
**Nota.** El nodo "Asignador de Clusters" no es específico para el nodo "k-Means". Realiza la tarea de asignación de clusters a partir de un conjunto de clusters basado en cualquiera de los algoritmos de clustering disponibles.

# Test de hipótesis

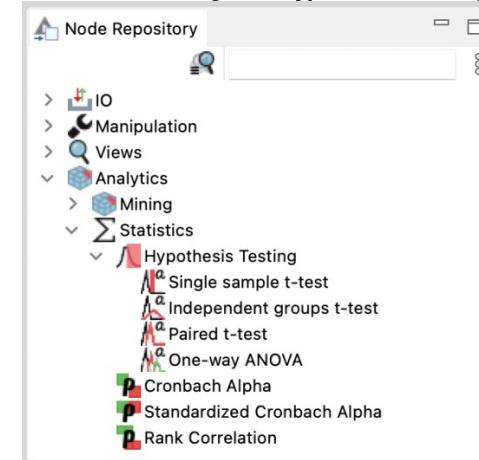
KNIME dispone de algunos nodos para realizar pruebas de hipótesis estadísticas clásicas. La mayoría de ellos se encuentran en "Analytics" → "Statistics" → "Hypothesis Testing": la prueba t de una muestra, la prueba t emparejada, el ANOVA de una vía y la prueba t de grupos independientes. Sólo el nodo que realiza la prueba de chi-cuadrado se encuentra fuera de la subcategoría "Hypothesis Testing" en el nodo "Crosstab".

Otros nodos más nuevos para la comprobación estadística de hipótesis están disponibles en "KNIME Labs" → "Statistics" en el panel "Node Repository".

4.46. Workflow "Clustering and Regression"



4.47. La subcategoría "Hypothesis Testing"



## 4.5. Ejercicios

### Ejercicio 1

Utilizando el archivo wine.data (conjunto de entrenamiento = 80% y conjunto de prueba = 20%) entrene un árbol de decisión para reconocer la clase a la que pertenece cada vino.

Ejecute el árbol de decisión en el conjunto de prueba de vinos y mida el rendimiento del árbol de decisión. En concreto, nos interesa saber cuántos falsos negativos hay para la clase 2.

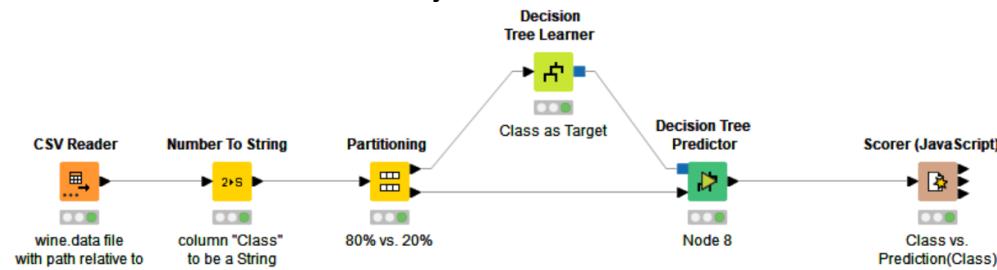
#### Solución al Ejercicio 1

En el nodo "Decision Tree Learner" utilizamos la columna "class" como columna de clase. Por defecto, el nodo "Lector CSV" lee la clase de datos del vino como Entero, ya que las clases son "1", "2" y "3".

Si utiliza un árbol de decisión, como hicimos nosotros, para la clasificación final, necesita que la columna "Clase" sea de valores nominales, es decir, que sea de tipo String. Tiene dos opciones para ello:

- Se lee "Class" como string (Cadena). En la ventana de configuración del "CSV Reader", haz clic con el botón derecho del ratón en la columna "Class" y cambia el tipo de "Integer" a "String"
- Dejas la configuración por defecto en el "CSV Reader" y luego utilizas un nodo "Number To String" para la conversión

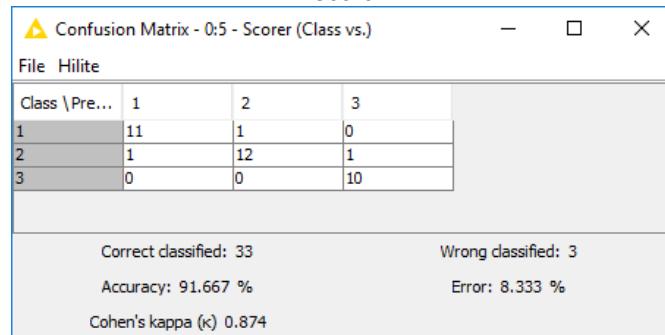
#### 4.48. Ejercicio 1: Workflow



##### Workflow: Chapter 4/Exercise 1

This workflow:  
- reads the wine data from the KBLdata folder  
- partitions the data: 80% to training set, 20% to test set  
- trains a decision tree to predict the wine class based on all other attributes (i.e. data columns)  
- applies the model to the test set  
- scores the right guess of Prediction(Class) vs. the original Class values.

#### 4.49. Ejercicio 1 "Ver: Ventana "Matriz de confusión" del nodo "Scorer"



A continuación, utilizamos un nodo "Scorer" para ver cuántos Falsos Negativos se produjeron en las estadísticas de precisión y/o en la matriz de confusión.

Abrimos la opción "Ver: Matriz de confusión" en el menú contextual y buscamos el número de registros pertenecientes a la clase 2 (eje Y) que se clasifican erróneamente como clase 3 (eje X).

Sólo hay un registro que ha sido clasificado erróneamente como clase 3 desde la clase 2.

**Note.** El nodo "Decision Tree Learner" necesita al menos un valor nominal para ser utilizado como columna de clasificación.

## Ejercicio 2

Construya un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) a partir de los datos del vino. Entrene un Perceptrón Multicapa (MLP) en el conjunto de entrenamiento para clasificar los datos según los valores de la columna "Clase".

A continuación, aplique el MLP al conjunto de prueba y mida el rendimiento del modelo.

### Solución al ejercicio 2

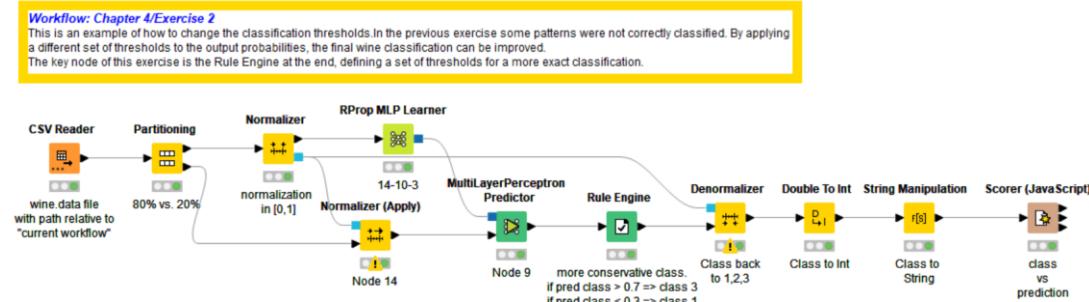
Utilizamos un nodo "Normalize" para escalar los datos antes de introducirlos en el MLP.

Como el conjunto de datos sobre el vino es muy pequeño, utilizamos todo el conjunto de datos para definir los parámetros de normalización.

El siguiente paso consistió en utilizar una red neuronal con una sola neurona de salida para modelar los tres valores de clase: "1", "2" y "3".

Como una neurona tiene un valor de salida continuo, su salida tiene que ser post-procesada para asignar una clase en forma de "1", "2" y "3" a cada fila de datos. Para ello utilizamos un nodo "Rule Engine" que implementa la siguiente regla:

#### 4.50. Ejercicio 2: Workflow



```
IF      $neuron output$ <= 0.3      => class 1  
ELSE IF $neuron output$ > 0.3 AND $neuron output$ < 0.6      => class 2  
ELSE IF $neuron output$ >= 0.6 => class 3
```

El rendimiento del modelo se mide con un nodo "Scorer". Como alternativa, puede explorar el nodo "Numerical Scorer" para medir los rendimientos con distancias numéricas.

## Ejercicio 3

Lea los datos "sitio web 1.txt" con un nodo CSV Reader. Este conjunto de datos describe el número de visitantes a un sitio web para el año 2013.

Calcule algunos parámetros estadísticos sobre el número de visitantes, como la media y la desviación estándar. Entrene un modelo Naïve Bayes sobre el número de visitantes para descubrir si una fila de datos específica se refiere a un fin de semana o a un día laborable.

Por último, dibuje la curva ROC para visualizar el rendimiento del Naïve Bayesian Classifier.

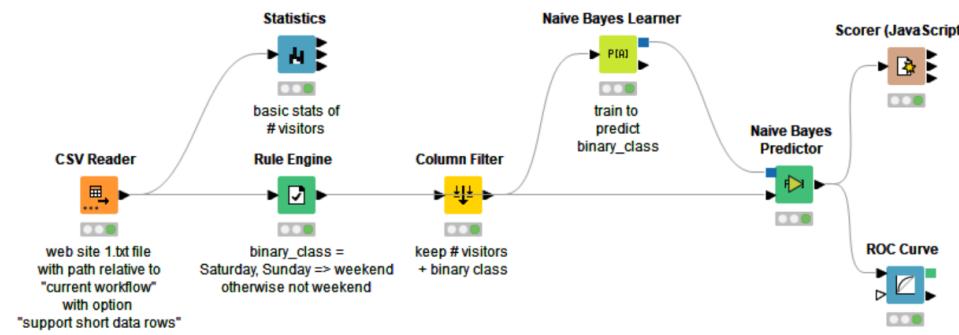
### Solución al ejercicio 3

Utilizamos un nodo "Rule Engine" para traducir la columna "Day of Week" en una clase binaria "fweekend/not weekend".

Filtramos la columna "Day of Week" para que la tarea de clasificación no fuera demasiado fácil para el Naïve Bayesian Classifier

Entrenamos la Bayesian Network con la clase binaria "weekend/not weekend" y construimos la curva ROC con la probabilidad de la clase "weekend".

#### 4.51. Ejercicio 3: workflow

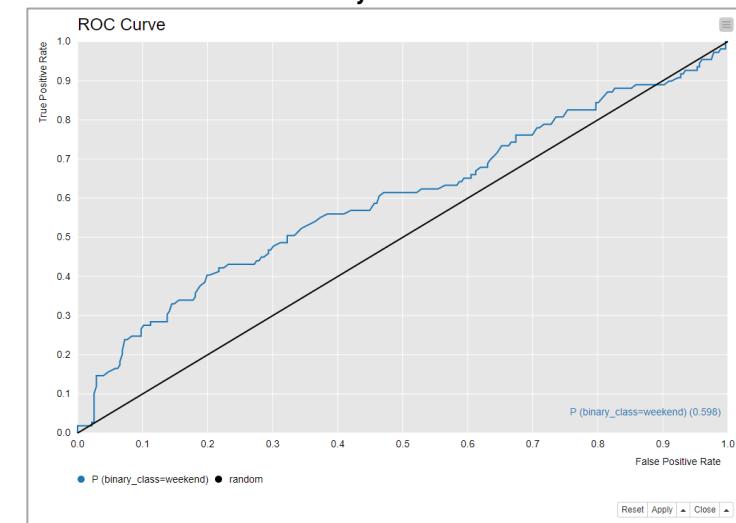


**Workflow: Chapter 4/Exercise 3**

One more exercise about classification.  
Calculate basic stats of # visitors

Train a model to classify weekend vs. no weekend by the # of visitors to a web site.  
Score model performances with accuracy and Area under the Curve (AuC).

#### 4.52. Ejercicio 3: Curva ROC sobre los resultados del clasificador bayesiano



# Capítulo 5. Preparación de los datos para la elaboración de informes

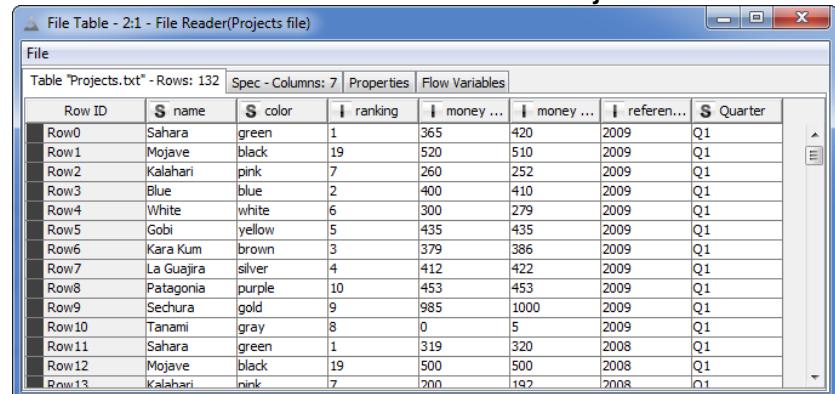
## 5.1. Introducción

Una parte de un proyecto de Ciencia de Datos es la elaboración de informes. Por ejemplo, se puede utilizar para mostrar las puntuaciones del modelo al consejo de administración o para cuantificar los resultados para su jefe. En este caso, es conveniente guardar los datos intermedios en unos archivos de historial, para poder replicar fácilmente los informes o proceder a un análisis de datos más adelante.

Mientras que KNIME Analytics Platform tiene algunas capacidades de reporte - a través de la integración con otras herramientas de reporte (BIRT, Tableau, Spotfire, Qlikview, y más) o a través de la plataforma web de KNIME Server para construir cuadros de mando: el WebPortal - nos centraremos aquí en algunas características de resumen para preparar los datos para el reporte o para el almacenamiento en algunas tablas o archivos intermedios de Data Warehousing. Hay una serie de nodos KNIME disponibles para ayudarnos en esta tarea de manipulación de datos.

Antes de continuar, vamos a crear un nuevo grupo de flujo de trabajo "Chapter5" y abrir un nuevo flujo de trabajo con el nombre de "Projects.txt".

### 5.1. Estructura de datos del archivo "Projects.txt"



The screenshot shows a KNIME node titled "File Table - 2:1 - File Reader(Projects file)". The table displays 132 rows of data with 7 columns. The columns are: Row ID, name, color, ranking, money..., money..., referen..., and Quarter. The data includes various project names like Sahara, Mojave, Kalahari, Blue, White, Gobi, Kara Kum, La Guajira, Patagonia, Sechura, Tanami, and several entries for Sahara, Mojave, and Kalahari. The "Quarter" column shows values like Q1, 2009, and 2008.

Row ID	name	color	ranking	money...	money...	referen...	Quarter
Row0	Sahara	green	1	365	420	2009	Q1
Row1	Mojave	black	19	520	510	2009	Q1
Row2	Kalahari	pink	7	260	252	2009	Q1
Row3	Blue	blue	2	400	410	2009	Q1
Row4	White	white	6	300	279	2009	Q1
Row5	Gobi	yellow	5	435	435	2009	Q1
Row6	Kara Kum	brown	3	379	386	2009	Q1
Row7	La Guajira	silver	4	412	422	2009	Q1
Row8	Patagonia	purple	10	453	453	2009	Q1
Row9	Sechura	gold	9	985	1000	2009	Q1
Row10	Tanami	gray	8	0	5	2009	Q1
Row11	Sahara	green	1	319	320	2008	Q1
Row12	Mojave	black	19	500	500	2008	Q1
Row13	Kalahari	pink	7	200	192	2008	Q1

Utilizaremos los datos "Projects.txt" disponibles en la carpeta de datos del libro "KBLData", que describen la evolución, en términos de dinero, de 11 proyectos ficticios a lo largo de 2007, 2008 y 2009. Cada proyecto se identifica con el nombre de un desierto diferente. Cada proyecto tiene también un color único y una clasificación de prioridad única, que permanecen iguales a lo largo de los años. Por último, el archivo describe la cantidad de dinero asignada a cada proyecto y la cantidad de dinero realmente utilizada por cada proyecto para cada trimestre de cada año. Tras leer el fichero con un nodo "CSV Reader", obtuvimos la estructura de datos de la figura 5.1. También escribimos "Projects File" bajo el nodo "CSV Reader", para entender rápidamente a qué datos se aplica este nodo.

## 5.2. Transformación de filas (Transform Rows)

Normalmente, los datos deben llegar al informe en una forma predefinida. En esta sección exploramos algunos nodos de KNIME que pueden ayudarnos a alcanzar la estructura deseada del conjunto de datos.

Los datos para el informe provienen del archivo "Projects.txt", que contiene una lista de proyectos y detalla cuánto dinero ha sido asignado o utilizado por cada proyecto durante los años 2007, 2008 y 2009. En el informe, queremos mostrar 3 tablas, con estructura como la mostrada en la figura 5.2, y 2 gráficos, a partir de una tabla de datos como la reportada en la figura 5.3.

La primera tabla debe mostrar los nombres de los proyectos en los encabezados de las filas, los años en los encabezados de las columnas, y cuánto dinero se ha asignado en total a cada proyecto para cada año en las celdas de la tabla. La segunda tabla tiene la misma estructura, pero muestra cuánto dinero ha sido utilizado en total por cada proyecto para cada año en las celdas de la tabla. La tercera tabla tiene la misma estructura que las dos tablas descritas anteriormente y muestra la cantidad de dinero restante (= dinero asignado - dinero utilizado) para cada proyecto para cada año.

El primer gráfico debe mostrar la cantidad total de dinero asignada a cada proyecto (eje Y) a lo largo de los tres años (eje X). El elemento de informe del gráfico BIRT se alimenta con un conjunto de datos en el que los valores del eje x y los valores del eje y aparecen en dos columnas diferentes; es decir, un conjunto de datos en el que el año y la correspondiente suma de dinero pertenecen a la misma fila.

El segundo gráfico tiene la misma estructura que el primero, pero muestra la cantidad total de dinero utilizada en lugar de la cantidad total de dinero asignada. Es decir, el gráfico debe mostrar la cantidad total de dinero utilizada por cada proyecto (eje Y) a lo largo de los tres años (eje X). Por ello, necesita un conjunto de datos con el año y el dinero total utilizado por cada proyecto separados en diferentes columnas.

### 5.2. Estructura de datos necesaria para las tablas del informe (nodo "Pivoting")

Project Name \ year	2007	2008	2009
Project 1	Sum (money) for project 1 in year 2007	Sum (money) for project 1 in year 2008	Sum (money) for project 1 in year 2009
Project 2	Sum(money)for project 2 in year 2007	...	
Project 3	..		

### 5.3. Estructura de datos necesaria para los gráficos del informe (node "GroupBy")

Project Name	year	Sum(money)
Project 1	2009	Sum (money) for project 1 in year 2009
Project 2	2009	Sum (money) for project 2 in year 2009
Project 3	2009	Sum (money) for project 3 in year 2009
...	...	...

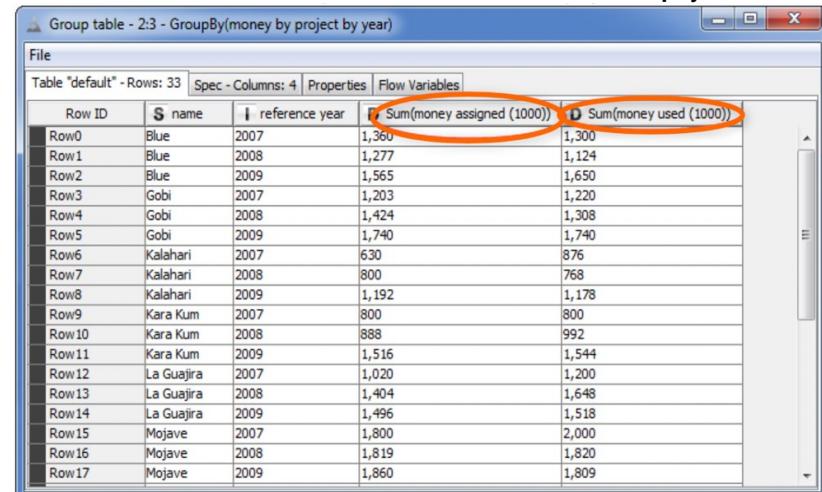
Tanto en la estructura de la tabla como en la del gráfico, necesitamos calcular la suma de dinero (asignada, utilizada o restante), pero necesitamos informarla en un diseño de datos diferente. Por ejemplo, en un caso queremos que los años sean las cabeceras de las columnas y en el otro caso queremos que los años sean los valores de las columnas.

La primera tabla de datos (Fig. 5.2) podría obtenerse con un nodo "Pivoting", mientras que la segunda tabla de datos (Fig. 5.3) con un nodo "GroupBy". A continuación, introducimos un nodo "GroupBy" y dos nodos "Pivoting" en el flujo de trabajo "Projects".

En el nodo "GroupBy", calculamos la suma (= método de agregación) de los valores en la columna "dinero asignado (1000)" y en la columna "dinero utilizado (1000)" (= columnas de agregación múltiple) para cada grupo de filas definido por la combinación de valores distintos en las columnas "año de referencia" y "nombre" (= columnas de grupo).

En la tabla de datos resultante, las dos primeras columnas contenían todas las combinaciones de valores distintos en las columnas "nombre" y "año de referencia". A continuación, se ejecutaron las agregaciones sobre los grupos de filas de datos definidos por cada par ("nombre", "año de referencia"). Los valores agregados se muestran en dos nuevas columnas "Suma(dinero asignado (1000))" y "Suma(dinero utilizado (1000))".

### 5.4. Tabla de datos de salida del nodo GroupBy



Row ID	name	reference year	Sum(money assigned (1000))	Sum(money used (1000))
Row0	Blue	2007	1,360	1,300
Row1	Blue	2008	1,277	1,124
Row2	Blue	2009	1,565	1,650
Row3	Gobi	2007	1,203	1,220
Row4	Gobi	2008	1,424	1,308
Row5	Gobi	2009	1,740	1,740
Row6	Kalahari	2007	630	876
Row7	Kalahari	2008	800	768
Row8	Kalahari	2009	1,192	1,178
Row9	Kara Kum	2007	800	800
Row10	Kara Kum	2008	888	992
Row11	Kara Kum	2009	1,516	1,544
Row12	La Guajira	2007	1,020	1,200
Row13	La Guajira	2008	1,404	1,648
Row14	La Guajira	2009	1,496	1,518
Row15	Mojave	2007	1,800	2,000
Row16	Mojave	2008	1,819	1,820
Row17	Mojave	2009	1,860	1,809

Llamamos al nuevo nodo "GroupBy" "dinero por proyecto por año".

En un nodo de "Pivoting" calculamos la suma (= método de agregación) de los valores de la columna "dinero asignado(1000)" (= columna de agregación) para cada combinación de valores en las columnas "reference year" (= pivot column) y "name" (= group column).

En el otro nodo "Pivoting" calculamos de nuevo la suma (= aggregation method) de los valores de la columna "money used(1000)" (= aggregation column) para cada combinación de valores en las columnas "año de referencia" (= pivot column) y "name" (= pivot column).

En ambos nodos de "Pivoting", elegimos mantener los nombres originales en la casilla "Column naming".

Los valores agregados se muestran en una tabla pivotante con <year + aggregation variable> como encabezado de columna, los nombres de los proyectos en la primera columna y la suma de "money assigned(1000)" o "money used(1000)" para cada proyecto y para cada año como contenido de la celda. Llamamos a los nuevos nodos pivotantes "dinero asignado al proyecto cada año" y "dinero utilizado por el proyecto cada año".

Para facilitar la lectura de la tabla pivotante, trasladamos los valores de la columna del nombre del proyecto para convertirlos en los RowIDs de la tabla de datos y renombramos las cabeceras de las columnas pivotantes con sólo el valor del año de referencia. Para ello, utilizamos un nodo "RowID" y un nodo "Column Rename" respectivamente.

### 5.5. Tabla pivotante de salida del nodo "Pivoting"

The screenshot shows a software interface titled "Pivot table - 2:2 - Pivoting(money used by project each year...)". The table has four columns: Row ID, name, 2007+money used (1000), 2008+money used (1000), and 2009+money used (1000). The rows are labeled Row0 through Row10, corresponding to project names: Blue, Gobi, Kalahari, Kara Kum, La Guajira, Mojave, Patagonia, Sahara, Sedcura, Tanami, and White. The values in the columns represent the sum of money used for each project in each year. Three callout boxes point to specific parts of the table:

- Distinct values in group column "name": Points to the "name" column header.
- Distinct values in pivot column "reference year" + aggregation variable name: Points to the column headers for 2007, 2008, and 2009.
- Sum(used money) for project Kara Kum in 2008: Points to the value 992 in the 2008 column for Kara Kum.

Row ID	name	2007+money used (1000)	2008+money used (1000)	2009+money used (1000)
Row0	Blue	1,300	1,124	1,650
Row1	Gobi	1,220	1,308	1,740
Row2	Kalahari	876	768	1,178
Row3	Kara Kum	800	992	1,544
Row4	La Guajira	1,200	1,648	1,518
Row5	Mojave	2,000	1,820	1,809
Row6	Patagonia	1,332	2,139	1,364
Row7	Sahara	905	1,460	1,670
Row8	Sedcura	3,600	3,113	4,000
Row9	Tanami	591	0	468
Row10	White	860	948	1,347

# RowID

El nodo RowID se encuentra en "Node Repository", "Manipulation" → "Row" → "Other" category.

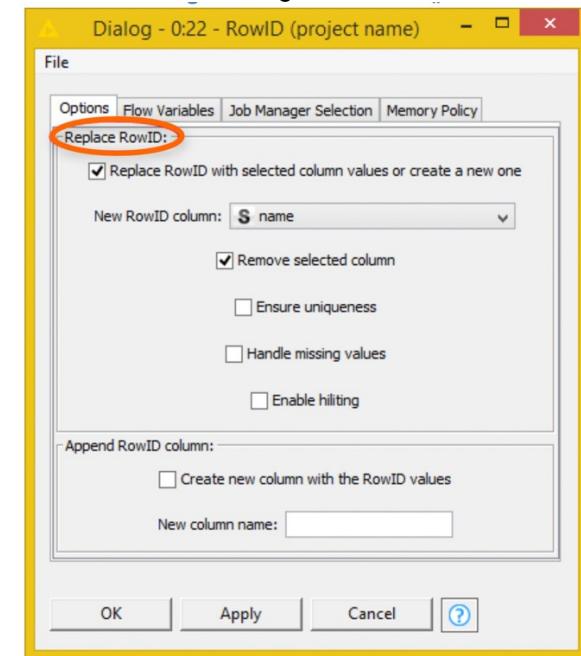
El nodo RowID node permite:

- Sustituir los RowIDs actuales por los valores de otra columna (mitad superior de la ventana de configuración)
- Copiar los RowIDs actuales en una nueva columna (mitad inferior de la ventana de configuración)

Cuando se sustituyen los RowIDs actuales se admiten algunas opciones adicionales.

- "Remove selected column" elimina la columna que se ha utilizado para reemplazar el RowIDs.
- "Ensure uniqueness" añade una extensión "(1)" a los RowIDs duplicados. La extensión se convierte en "(2)" o "(3)" etc... dependiendo de cuántos valores duplicados se encuentren para este RowID.
- "Handle missing values" sustituye los valores que faltan en los RowIDs por valores por defecto.
- "Enable hiliting" mantiene un mapa entre los RowIDs antiguos y los nuevos para que el hiliting siga funcionando en otros nodos.

5.6. Ventana de configuración del nodo "RowID"



En el panel "Repositorio de nodos", cerca del nodo "Pivoting" se encuentra el nodo "Unpivoting".

Aunque no vamos a utilizar el nodo "Unpivoting" en nuestro flujo de trabajo de ejemplo, vale la pena echarle un vistazo.

# Unpivoting

El nodo "Unpivoting" gira el contenido de la tabla de datos de entrada. Este tipo de rotación se muestra en las figuras 5.8 y 5.9. Básicamente, produce una tabla de datos de salida de estilo "GroupBy" a partir de la tabla de datos de entrada "pivoted".

El nodo "Unpivoting" está ubicado en "Manipulation" → "Row" → "Transform".

En los ajustes hay que definir

- Qué columnas deben utilizarse para la redistribución de celdas
- Qué columnas deben conservarse del conjunto de datos original

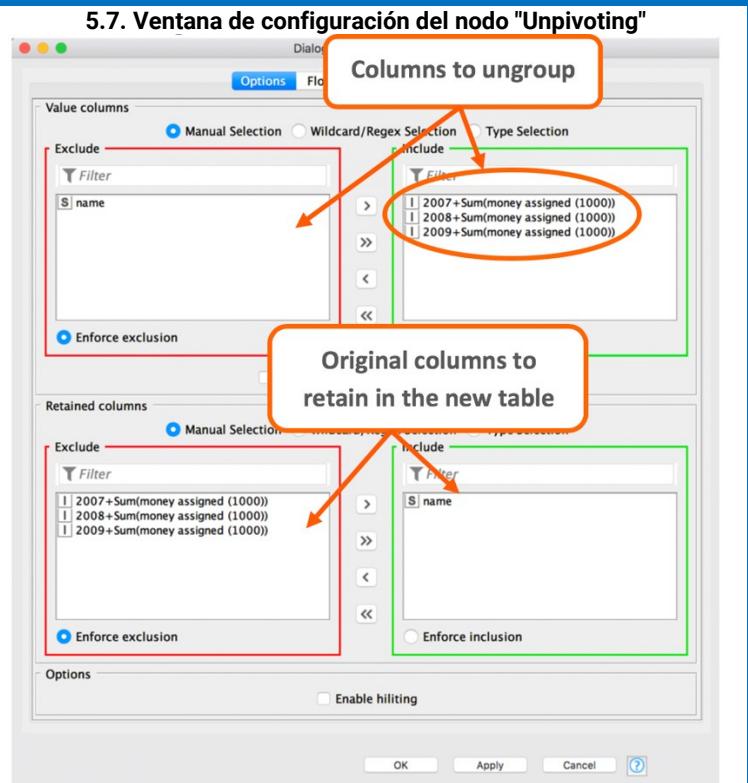
El proceso de "unpivoting" produce 3 nuevas columnas en la tabla de datos:

- Una columna llamada "RowIDs", que contiene los RowIDs de la tabla de datos de entrada
- Una columna llamada "ColumnNames", que contiene las cabeceras de las columnas de la tabla de datos de entrada
- Una columna llamada "ColumnValues", que reconecta los valores de las celdas originales con su RowID y su cabecera de columna

La selección de columnas sigue el marco ya visto de "Exclude"/"Include":

- Las columnas aún disponibles para la agrupación se enumeran en el marco "Available column(s)". Las columnas seleccionadas aparecen en el marco "Group column(s)".
- Para pasar del marco "Available column(s)" al marco "Group column(s)" y viceversa, utilice los botones "add" y "remove". Para mover todas las columnas a un marco u otro, utilice los botones "add all" y "remove all".

"Enforce Inclusion/Exclusion" mantiene las columnas incluidas/excluidas como fijas y añade posibles nuevas columnas al otro conjunto de columnas.



**5.8. Tabla de datos de entras**

	Col1	Col2	Col3
ID 1	1	3	5
ID 2	2	4	6

**5.9. Tabla de datos no pivoteada**

	RowIDs	ColumnNames	ColumnValues
Row 1	ID 1	Col1	1
Row 2	ID 1	Col2	3
Row 3	ID 1	Col3	5
Row 4	ID 2	Col1	2
Row 5	ID 2	Col2	4
Row 6	ID 2	Col3	6

**Nota.** Pivoting + Unpivoting = GroupBy (pivotear + despivotear = Groupby)

Las tablas de datos de salida del nodo "GroupBy" y del nodo "Pivoting" se ordenan por los valores de las columnas del grupo.

El nodo "Sorter", al igual que el nodo "Pivoting" y el nodo "GroupBy", es otro nodo que se utiliza frecuentemente para la elaboración de informes. Para fines demostrativos mostramos brevemente aquí el nodo "Sorter".

## Nodo "Sorter"

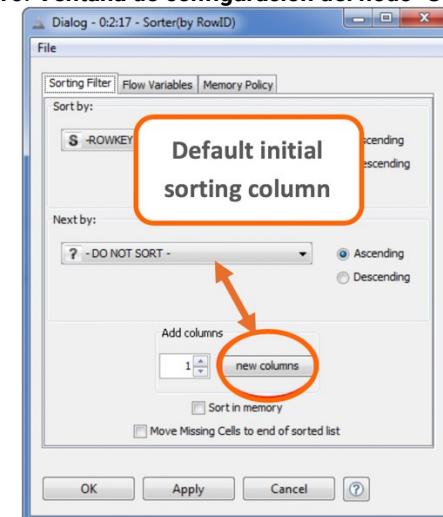
El nodo "Sorter" ordena las filas de una tabla de datos ordenando los valores de una de sus columnas. En la configuración hay que seleccionar

- La(s) columna(s) a ordenar (también se acepta la columna RowIDs)
- Si se va a ordenar en orden ascendente o descendente (is possible to sort the table by multiple columns).

Para añadir una nueva columna de ordenación:

- Haga clic en el botón "nuevas columnas".
- Seleccione la nueva columna

5.10. Ventana de configuración del nodo "Sorter"



La primera columna (en la parte superior de la ventana de configuración) da la ordenación primaria; la segunda columna da la ordenación secundaria, y así sucesivamente.

## 5.3. Uniendo Columnas

Después de aplicar los dos nodos "Sorter", ahora tenemos dos tablas de datos con la misma estructura de columnas:

- RowID que contiene los nombres de los proyectos
- Columna "2009" con el dinero utilizado/asignado para el año 2009
- Columna "2008" con el dinero utilizado/asignado para el año 2008
- Columna "2007" con el dinero utilizado/asignado para el año 2007

Ahora, sería útil Have all values for used and assigned money over the years together for each project in the same row, for example:

RowID	assigned 2009	assigned 2008	assigned 2007	used 2009	used 2008	used 2007
<project name>	...	...	...	...	...	...

- - Calcule el dinero restante para cada año para cada proyecto, como:  $remain <year> = assigned <year> - used <year>$

Básicamente, queremos unir las dos tablas de datos, la tabla con los valores del dinero asignado y la tabla con los valores del dinero usado, en una sola tabla. Después, queremos calcular los valores del dinero restante.

En primer lugar, para poder realizar la unión de tablas sin confusión, necesitamos que las distintas columnas lleven nombres diferentes. Veremos que hay que tomar medidas en caso de una unión de tablas con columnas con el mismo nombre. Conectemos un nodo "Column Rename" a cada nodo "RowID".

En la table resultante del nodo "money used by project each year", cambiemos el nombre de la columna llamada "2009 + money used(1000)" a "used 2009", la columna "2008 + money used(1000)" a "used 2008", y la columna a "2007 + money used(1000)" to "used 2007".

En la table resultante del nodo "money assigned to project each year", cambiemos el nombre de la columna llamada "2009 + money assigned(1000)" a "assigned 2009", la columna "2008 + money assigned(1000)" a "assigned 2008", y la columna "2007 + money assigned(1000)" a "assigned 2007".

Las tablas de datos que queremos unir tienen ahora la estructura que se indica a continuación.

5.11. Dinero asignado a cada proyecto cada año			
Row ID	D assigned 2007	D assigned 2008	D assigned 2009
Blue	1,360	1,277	1,565
Gobi	1,203	1,424	1,740
Kalahari	630	800	1,192
Kara Kum	800	888	1,516
La Guajira	1,020	1,404	1,496
Mojave	1,800	1,819	1,860
Patagonia	864	2,098	1,359
Sahara	806	1,457	1,495
Sechura	3,200	2,966	3,940
Tanami	453	0	453
White	860	1,087	1,420

5.12. Dinero usado por cada proyecto cada año			
Row ID	D used 2007	D used 2008	D used 2009
Blue	1,300	1,124	1,650
Gobi	1,220	1,308	1,740
Kalahari	876	768	1,178
Kara Kum	800	992	1,544
La Guajira	1,200	1,648	1,518
Mojave	2,000	1,820	1,809
Patagonia	1,332	2,139	1,364
Sahara	905	1,460	1,670
Sechura	3,600	3,113	4,000
Tanami	591	0	468
White	860	948	1,347

Ahora que tenemos la estructura de datos correcta, necesitamos realizar una unión de tablas. Queremos unir las celdas para que estén en la misma fila basándonos en el nombre del proyecto; es decir, en este caso se trata del RowID. De hecho, queremos que la fila de dinero utilizado para el proyecto "Blue" se añada al final de la fila correspondiente de la tabla con el dinero asignado. KNIME tiene un nodo muy potente que puede utilizarse para unir tablas, conocido como nodo "Joiner".

# Joiner

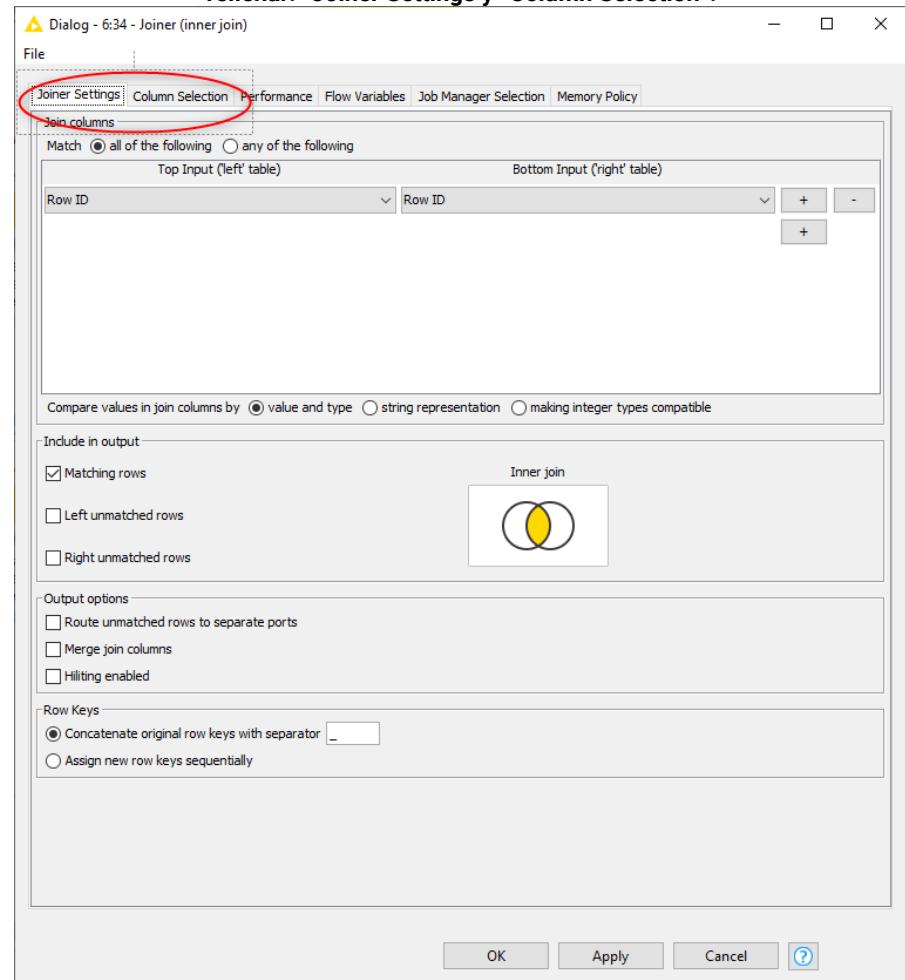
El nodo "Joiner" esta ubicado en "Node Repository" "Manipulation" → "Column" → "Split & Combine"

El nodo "Joiner" toma dos tablas de datos en los puertos de entrada y hace coincidir una columna de la tabla del puerto superior (tabla izquierda) con una columna de la tabla del puerto inferior (tabla derecha). Estas columnas también pueden ser las columnas RowID. Este nodo tiene tres puertos de salida: uno para las filas emparejadas, otro para las filas no emparejadas de la tabla superior (izquierda) (si las hay) y otro para las filas no emparejadas de la tabla inferior (derecha) (si las hay).

En la ventana de configuración del nodo "Joiner" hay dos pestañas que hay que llenar.

- La pestaña "Joiner Settings" contiene todos los ajustes relativos a:
  - El modo de unión
  - Las columnas que deben coincidir
  - Otros parámetros secundarios
- La pestaña "Column Selection" contiene todos los ajustes relativos a:
  - Las columnas de las dos tablas que se incluirán en la tabla unida
  - Cómo tratar las columnas duplicadas (es decir, las columnas con el mismo nombre)
  - Si se debe filtrar la columna de unión de la tabla de datos de la izquierda y/o de la derecha o ninguna

5.13. Ventana de configuración del nodo "Joiner", que incluye dos pestañas a llenar: "Joiner Settings" y "Column Selection".



# Nodo Joiner: pestaña "Joiner Settings"

La pestaña "Joiner Settings" establece las propiedades básicas de unión, como el "join mode", the "joining columns", the "matching criterion", etc ("modo de unión", las "columnas de unión", el "criterio de coincidencia", etc.)

La primera configuración se refiere a las columnas con valores que deben coincidir (**columns with values to match**) de la tabla superior (izquierda) y de la tabla inferior (derecha). Se admite la unión de varias columnas. Para añadir un nuevo par de columnas de unión

- Haga clic en el botón "+";

Los valores de las columnas clave pueden coincidir por valor y tipo, igual que las cadenas (los tipos pueden diferir), o igual que los números.

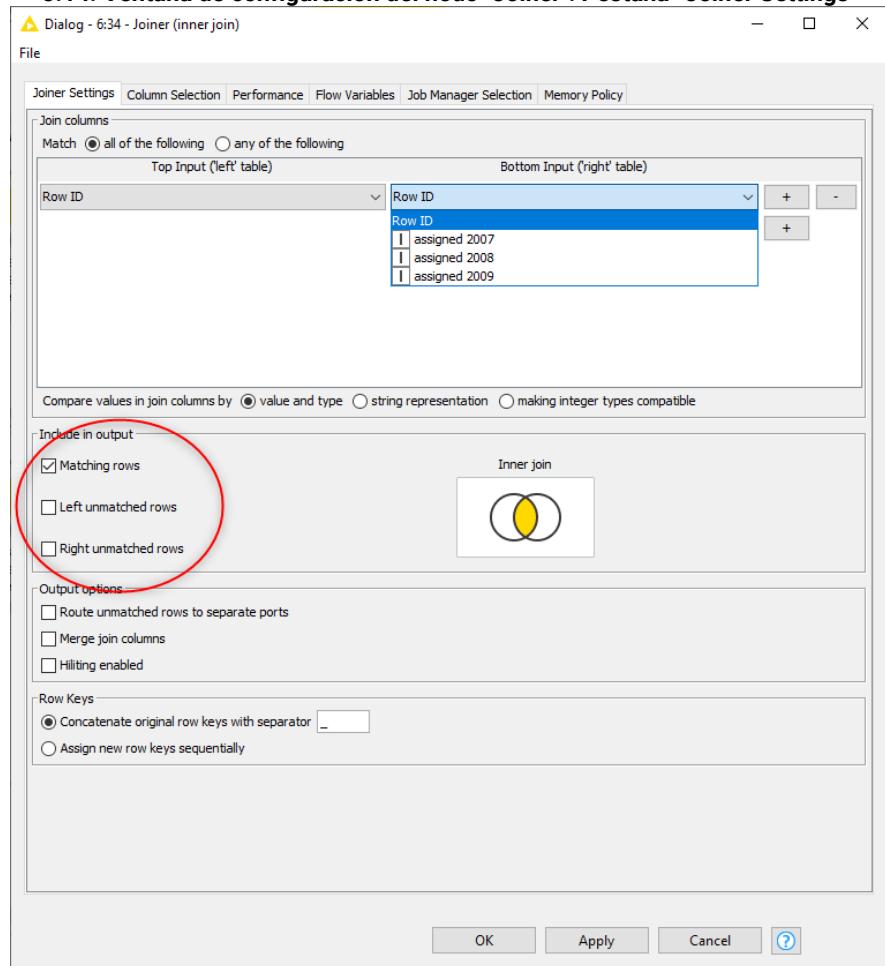
El segundo ajuste es el modo de unión (joint mode).

- **Matching rows** es el equivalente a una unión interna, es decir, mantiene sólo las filas en las que los valores de las dos columnas de unión coinciden;
- **Left unmatched rows** :Las filas no coincidentes conservan además todas las filas de la tabla izquierda (superior), aunque no coincidan. Activar la primera y esta casilla equivale a una unión a la izquierda.
- **Right unmatched rows** Las filas no coincidentes de la derecha mantienen además todas las filas de la tabla de la derecha (inferior), incluso si no coinciden. Activar la primera y esta casilla equivale a una unión a la derecha.
- Activar las tres casillas de verificación equivale a una unión externa completa.

Otros ajustes:

- Output options establece la(s) tabla(s) a exportar en los puertos de salida.

5.14. Ventana de configuración del nodo "Joiner": Pestaña "Joiner Settings"



- Row Keys establece el formato de las nuevas claves de fila.

## Nodo Joiner: pestaña “Column Selection”

La pestaña "Column Selection" define cómo manejar las columnas que no participan en el proceso de coincidencia.

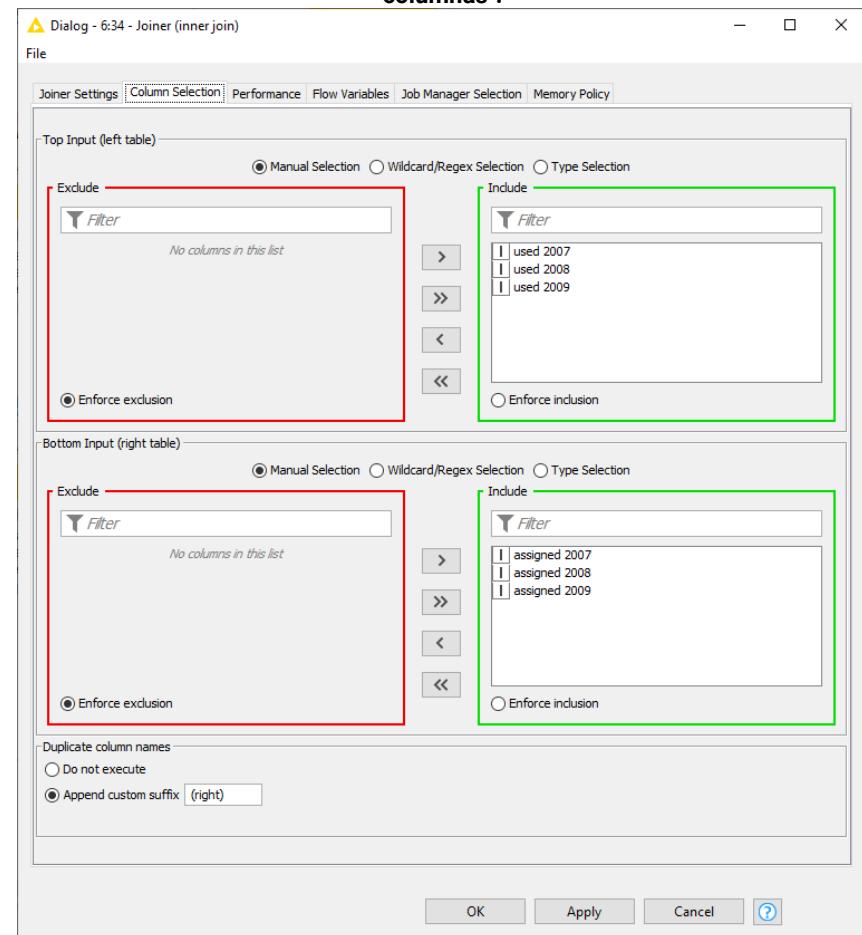
Una vez que los dos valores clave coinciden, las demás columnas de la tabla izquierda y derecha pueden conservarse o eliminarse. Un marco clásico "Exclude"/"Include" establece las columnas que se deben conservar o eliminar de ambas tablas de entrada.

- Las columnas que deben mantenerse en la nueva tabla unida se enumeran en el marco "Incluir". Todas las demás columnas se enumeran en el marco "Excluir".
- Para pasar del marco "Incluir" al marco "Excluir" y viceversa, utilice los botones "añadir" y "eliminar". Para mover todas las columnas a un marco u otro, utilice los botones "añadir todo" y "eliminar todo".

El panel "Duplicate column names" ofrece algunas opciones para tratar el problema de las columnas con el mismo encabezado header (= duplicate columns) en las dos tablas.

- "Do not execute" produce un error
- "Append suffix" añade un sufijo, por defecto o personalizado, al nombre de las columnas duplicadas en la tabla derecha

**5.15. Ventana de configuración del nodo "Joiner": Pestaña "Selección de columnas".**



## 5.16. Modos de unión con la opción "Filtrar duplicados" activada.

**Left Table**

Row ID	ID	age	income	class
Row0	1	23	<=50K	F1
Row1	3	25	<=50K	F3
Row2	6	22	>50K	A4
Row3	8	21	<=50K	C3

**Right Table**

Row ID	ID	age	income	sex
Row0	1	23	<=50K	M
Row1	2	25	<=50K	F
Row2	4	23	>50K	M
Row3	5	21	<=50K	F
Row4	6	25	>50K	M
Row5	7	24	<=50K	M

**Join by ID**

**Inner Join**

**Joined table - 3:51 - Joiner**

Row ID	ID	age	income	class	age (#1)	income (#1)	sex
Row0_Row0	1	23	<=50K	F1	23	<=50K	M
Row2_Row4	6	22	>50K	A4	25	>50K	M

**Left Outer Join**

**Missing values in the right table.**

**Joined table - 3:51 - Joiner**

Row ID	ID	age	income	class	age (#1)	income (#1)	sex
Row0_Row0	1	23	<=50K	F1	23	<=50K	M
Row2_Row4	6	22	>50K	A4	25	>50K	M
Row1_?	3	25	<=50K	F3	?	?	?
Row3_?	8	21	<=50K	C3	?	?	?

**Right Outer Join**

**Missing values in the left table.**

**Joined table - 3:51 - Joiner**

Row ID	ID	age	income	class	age (#1)	income (#1)	sex
Row0	1	23	<=50K	F1	23	<=50K	M
Row4	6	22	>50K	A4	25	>50K	M
?_Row1	?	?	?	?	?	?	?
?_Row2	?	?	?	?	?	?	?
?_Row3	?	?	?	?	?	?	?
?_Row5	?	?	?	?	?	?	?

**Joined table - 3:51 - Joiner**

**Missing values in the left table**

**Missing values in the right table**

Row ID	ID	age	income	class	age (#1)	income (#1)	sex
Row0_Row0	1	23	<=50K	F1	23	<=50K	M
Row2_Row4	6	22	>50K	A4	25	>50K	M
Row1_?	3	25	<=50K	F3	?	?	?
Row3_?	8	21	<=50K	C3	?	?	?
?_Row1	?	?	?	?	?	?	?
?_Row3	?	?	?	?	?	?	?
?_Row5	?	?	?	?	?	?	?

Unimos las dos tablas (money assigned y money used) utilizando los RowIDs como columna de unión para ambas; elegimos añadir un sufijo "(derecho)" (append a suffix "(right)") para las columnas de la tabla derecha con el mismo nombre que las columnas de la tabla izquierda; y elegimos el inner join como modo de unión.

Los datos resultantes se muestran en la figura 5.17. Puede ver que ahora los valores de "assigned money" y de "used money" están en la misma fila para cada proyecto.

Por supuesto, es posible hacer la unión en columnas diferentes a las de los RowIDs. Sin embargo, la unión en RowID permite al usuario mantener los valores originales de RowID, que podrían ser importantes para algún análisis o manipulación de datos posterior. En este caso necesitamos que los RowIDs contengan las claves de unión. Para manipular los valores RowID, KNIME tiene un nodo "RowID" ..

5.17. Tabla de datos de salida del nodo "Joiner" con "Inner Join" como modo de unión

Row ID	D used 2...	D used 2...	D used 2...	D assigne...	D assigne...	D assigne...
Blue	1,300	1,124	1,650	1,360	1,277	1,565
Gobi	1,220	1,308	1,740	1,203	1,424	1,740
Kalahari	876	768	1,178	630	800	1,192
Kara Kum	800	992	1,544	800	888	1,516
La Guajira	1,200	1,648	1,518	1,020	1,404	1,496
Mojave	2,000	1,820	1,809	1,800	1,819	1,860
Patagonia	1,332	2,139	1,364	864	2,098	1,359
Sahara	905	1,460	1,670	806	1,457	1,495
Sechura	3,600	3,113	4,000	3,200	2,966	3,940
Tanami	591	0	468	453	0	453
White	860	948	1,347	860	1,087	1,420

## 5.4. Nodos Miscelaneos (Misc Nodes)

En nuestro informe queremos incluir el dinero restante de cada año, calculado como: <remaining value> = <assigned value> - <used value>. Hay dos maneras de calcular este valor: el nodo "Math Formula" node and thy el nodo "Java Snippet". Todos estos nodos se encuentran en la categoría "Misc".

Los nodos "Java Snippet" permiten al usuario ejecutar trozos de código Java. A continuación, podemos utilizar un nodo "Java Snippet" para calcular la cantidad <valor restante>. En realidad, utilizaremos tres nodos "Java Snippet": uno para calcular el <remaining value 2009>, un segundo para calcular el <remaining value 2008>, y un tercero para calcular el <remaining value 2007>. Llamamos a los tres nodos "Java Snippet" "remain 2009", "remain 2008" y "remain 2007".

Hay dos tipos de nodos "Java Snippet": el nodo "Java Snippet" y el nodo "Java Snippet (simple)". La funcionalidad es la misma: ejecutar un trozo de código Java. Sin embargo, el nodo "Java Snippet" tiene una interfaz gráfica más compleja y flexible, mientras que el nodo "Java Snippet (simple)" ofrece una interfaz gráfica más simplificada. Es decir, el nodo "Java Snippet" es para usuarios más expertos y piezas de

código más complejas, mientras que el nodo "Java Snippet (simple)" es para usuarios medianamente expertos y piezas de código Java más sencillas.

## Java Snippet (simple)

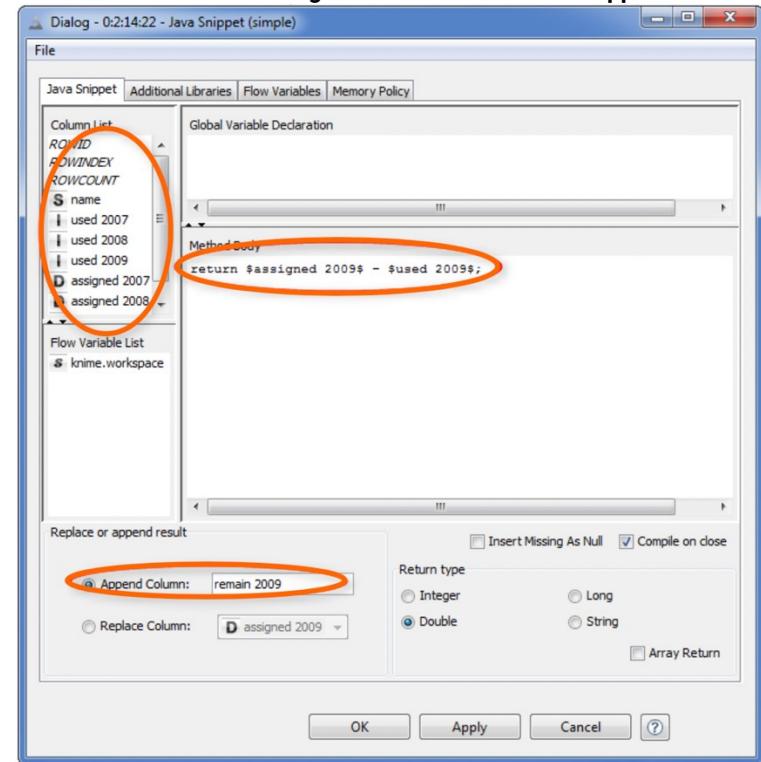
Un nodo "Java Snippet (simple)" permite la ejecución de un fragmento de código Java y coloca el valor resultante en una columna de datos nueva o existente. El código debe terminar con la palabra clave "return" seguida del nombre de una variable o expresión.

El nodo "Java Snippet (simple)" se ubica en "Node Repository" "Misc" → "Java Snippet" category.

Al abrir el diálogo de configuración del nodo, aparece una ventana que incluye varios paneles.

- El Java editor es la parte central de la ventana de configuración. Aquí es donde se escribe el código. Recuerde que el código tiene que devolver algún valor y, por tanto, debe terminar con la palabra clave "return" seguida de un nombre de variable o expresión. No se permiten múltiples declaraciones "return" en el código.
- La list of column names está en la parte superior izquierda. Los nombres de las columnas pueden utilizarse como variables dentro del código Java. Tras hacer doble clic en el nombre de la columna, la variable correspondiente aparece en el editor de Java. Las variables llevan el tipo de su columna original al código java, es decir: Double, Integer, String y Arrays. Las variables de arrays provienen de columnas del tipo Collection Type.
- **Name and type of the column** para añadir o reemplazar. El tipo de columna puede ser "Integer", "Double" o "String". Si el tipo de columna no coincide con el tipo de la variable que se devuelve, el código del fragmento de Java no se compilará.
- También es posible devolver **arrays** en lugar de variables individuales (véase la casilla de verificación en la parte inferior derecha). En este caso, se añadirá un número de columnas (tantas como la longitud del array) a la tabla de datos de salida
- En el panel "**Global Variable Declaration**" se pueden crear variables globales de panel, que se utilizarán recursivamente en el código a través de las filas de datos de la tabla.
- La pestaña "**Additional Libraries**" permite la inclusión de Bibliotecas Java no estándar.
- En "**Global Variable Declaration**" se pueden crear variables globales de panel, que se utilizarán recursivamente en el código a través de las filas de datos de la tabla.

5.18. Ventana de configuración del nodo "Java Snippet" node



Para el nodo "remain 2009" utilizamos el código Java: return \$assigned 2009\$ - \$used 2009\$

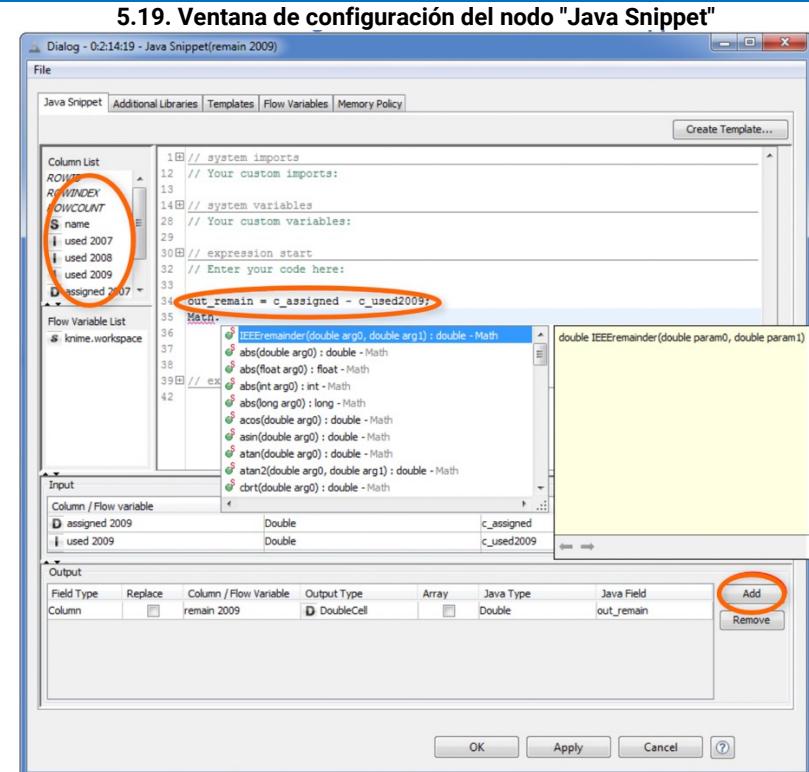
El mismo código podría utilizarse con otros dos nodos "Java snippet" para calcular "remain 2008" y "remain 2009". La misma tarea podría haberse realizado con un nodo "Java Snippet".

## Java Snippet

Al igual que el nodo "Java Snippet (simple)", el nodo "Java Snippet" permite la ejecución de un fragmento de código Java y coloca el valor resultante en una columna de datos nueva o existente. El nodo se encuentra en el panel "Node Repository" en la categoría "Misc" → "Java Snippet".

La ventana de configuración del nodo "Java Snippet" también contiene:

- El **Java editor**. Es la parte central de la ventana de configuración y es la principal diferencia con el nodo "Java Snippet (simple)". El editor tiene secciones reservadas para: declaración de variables, importaciones, código y operaciones de limpieza al final. The "**expression\_start**" section contains the code.
  - o La sección "**system variables**" contiene las variables globales, aquellas cuyo valor tiene que ser transportado fila a fila. Las variables declaradas dentro de la sección "expression\_start" restablecerán su valor en cada procesamiento de fila. The "**system imports**" section is for the library import declaration.
  - o También se habilita el autocompletado (self-completion), lo que permite una búsqueda más fácil de métodos y variables. Se pueden exportar una o varias variables de salida en una o varias columnas de datos de salida nuevas o existentes.
- La tabla llamada "**Input**". Esta tabla contiene todas las variables derivadas de las columnas de datos de entrada. Utilice los botones "Añadir" y "Eliminar" para añadir columnas de datos de entrada a la lista de variables que se utilizarán en el código Java.
- La lista de columnas (**list of column names**) on the top left-hand side. Los nombres de las columnas pueden utilizarse como variables dentro del código Java. Tras hacer doble clic en el nombre de la columna, la variable correspondiente aparece en el editor de Java y en la lista de variables de entrada de la parte inferior. Las variables llevan el tipo de su columna original al código java, es decir: Double, Integer, String y Arrays. Sin embargo, su tipo puede cambiarse (cuando sea posible) modificando el campo "Tipo Java" en la tabla denominada "Entrada".
- La **tabla llamada "Output"**. Esta tabla contiene las columnas de datos de salida que se crearán como nuevas o que se sustituirán por los nuevos valores. Para añadir una nueva columna de datos de salida, haga clic en el botón "Añadir". Utilice los botones "Añadir" y "Eliminar" para añadir y eliminar columnas de datos de salida. Active el indicador "Reemplazar" si la columna de datos va a sustituir a una columna existente. El tipo de columna de datos puede ser "Entero", "Doble" o "Cadena". Si el tipo de columna no coincide con el tipo de la variable que se devuelve,



el código del fragmento de Java no se compilará. También es posible devolver matrices en lugar de variables individuales, simplemente activando la bandera "Array". Recuerde asignar un valor a las variables de salida en la zona de código Java.

Ambos nodos "Java Snippet" son nodos muy potentes, ya que permiten al usuario desplegar la potencia de Java dentro de KNIME. Sin embargo, la mayoría de las veces, estos nodos tan potentes no son necesarios. Todas las operaciones matemáticas, por ejemplo, pueden ser realizadas por el nodo "Math Formula". El nodo "Math Formula" está optimizado para las operaciones matemáticas y, por lo tanto, tiende a ser más rápido que los nodos "Java Snippet".

## Nodo Math Formula

El nodo "Math Formula" permite implementar fórmulas matemáticas y funciona de forma similar al nodo "String Manipulation" (ver sección 3.5).

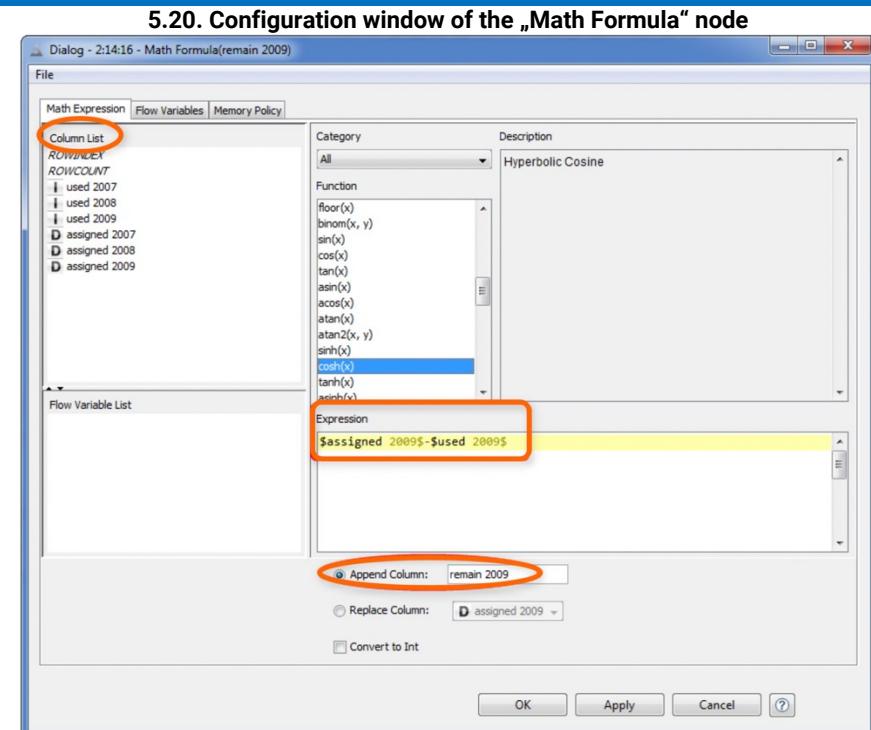
El nodo "Math Formula" no forma parte de la versión básica de KNIME. Tiene que ser descargado con el paquete de extensión (ver par. 1.5) "KNIME Math Expression Extension (JEP)". Una vez instalado el paquete de extensión, el nodo "Math Formula" se encuentra en el panel "Repositorio de nodos" en la categoría "Misc".

En la ventana de configuración hay 3 listas:

- La lista de nombres de columnas de la tabla de datos de entrada
- La lista de variables (para ver en el "KNIME Cookbook")
- Una lista de funciones matemáticas, por ejemplo,  $\log(x)$ .
- El editor de expresiones

Al hacer doble clic en las columnas de datos de la lista de la izquierda, se insertan automáticamente en el editor de expresiones. Puede completar la expresión matemática escribiendo lo que falta. Aquí, al igual que en los nodos "Java Snippet",  $\$<\text{column\_name}>\$$  indica el uso de una columna de datos.

En la lista central están disponibles varias funciones para construir una fórmula matemática.



En la parte inferior se puede insertar el nombre de la columna que se va a añadir o sustituir.

El nodo exporta datos de tipo double, pero también se pueden exportar datos de tipo entero activando la opción "Convert to Int".

En la ventana de configuración del nodo Fórmula matemática introducido en el flujo de trabajo "Projects" , implementamos el mismo cálculo de valores <remain 2009> mencionado en el nodo "Fragmento de Java" anteriormente en este capítulo. Simplemente necesitábamos

- Hacer doble clic en las dos columnas \$used 2009\$ y \$assigned 2009\$ en la lista de columnas
- Escribir un carácter "-" entre los dos nombres de las columnas de datos en el editor de expresiones.

En el flujo de trabajo "Proyectos" decidimos utilizar esta implementación de los valores restantes con los nodos "Math Formula". Es decir, utilizamos 3 nodos "Math Formula", uno tras otro, para calcular los valores restantes de 2009, los valores restantes de 2008 y los valores restantes de 2007 respectivamente. También mantuvimos la implementación con los nodos Java Snippet para fines de demostración. Sin embargo, sólo la secuencia de nodos "Math Formula" ha sido conectada al siguiente nodo. Hemos dado a los nodos "Math Formula" los mismos nombres que utilizamos para los nodos Java Snippet, para indicar que realizan exactamente la misma tarea.

Otro tipo de nodo "Math Formula" es el nodo "Math Formula (Multi Column)"

# Math Formula (Multi Column)

El nodo "Math Formula (Multi Column)" permite implementar la misma fórmula matemática sobre una lista de columnas, según se especifique en la ventana de configuración.

En la parte superior de la ventana de configuración, elegimos la lista de columnas sobre las que implementar la fórmula, a través de un cuadro de Exclude/Include

Después de esto, tenemos la misma configuración que para el nodo más simple "Fórmula matemática".

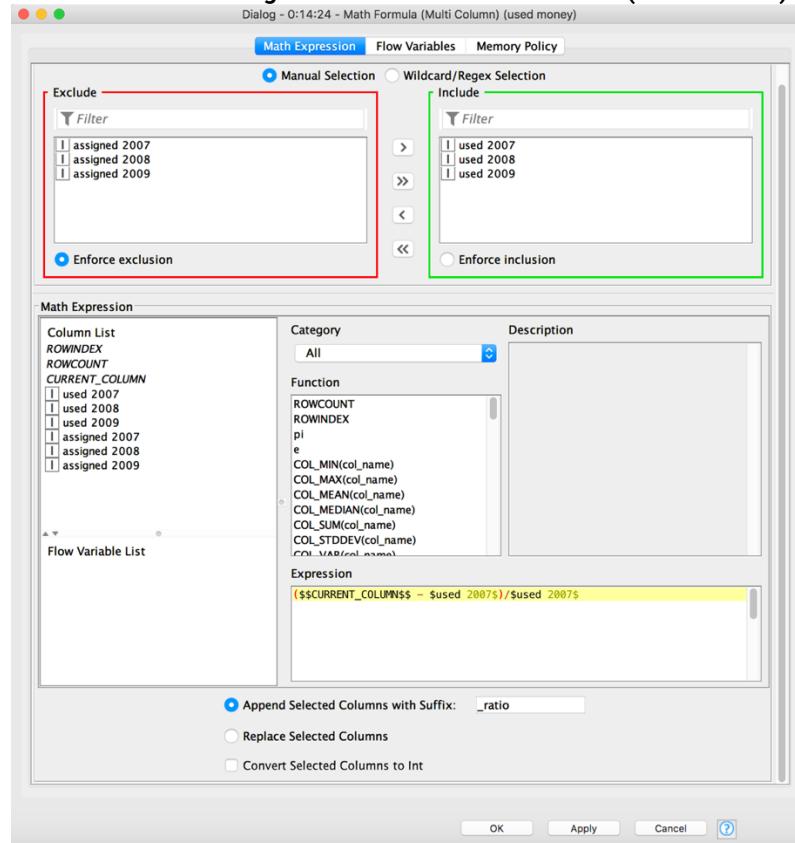
- La lista de nombres de columnas de la tabla de datos de entrada
- La lista de variables (para ver en el "KNIME Cookbook")
- Una lista de funciones matemáticas, por ejemplo  $\log(x)$  y sus descripciones
- El editor de expresiones matemáticas

Al hacer doble clic en las columnas de datos de la lista de la izquierda, se insertan automáticamente en el editor de expresiones. Puede completar la expresión matemática escribiendo lo que falta.

Las últimas tres opciones incluyen:

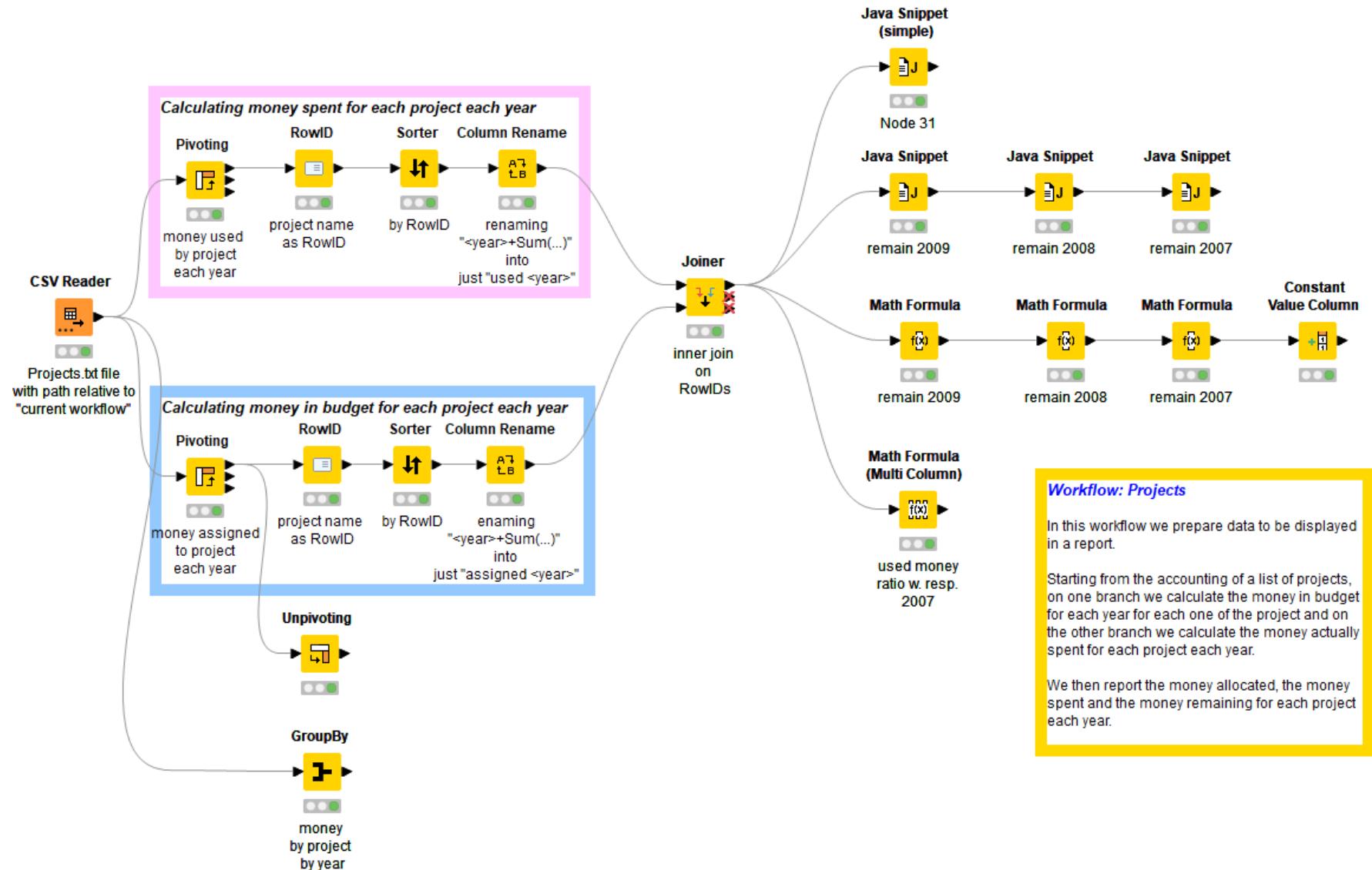
- Un sufijo para añadir a los nombres de las columnas originales para formar los nombres de las columnas de salida, si queremos crear nuevas columnas;
- La opción de sobrescribir los valores de las columnas originales
- La casilla para convertir todos los resultados en números enteros

5.21. Ventana de configuración del nodo "Math Formula (Multi Column)"



Terminemos el flujo de trabajo con un nodo "Constant Value Column" para añadir una columna temporal adicional con el objetivo de este flujo de trabajo "Preparing Data Workflow". La Columna de Valor Constante fue entonces conectada al último nodo de Math Formula.

## 5.22. Flujo de trabajo "Projects"



## 5.5. Limpieza final (Cleaning Up the Final Workflow)

El flujo de trabajo de "Projects" ya está terminado; sin embargo, podemos ver que está muy lleno de nodos, especialmente si queremos mantener todos los nodos "Java Snippet" y los nodos "Math Formula". Para hacer el flujo de trabajo más legible, podemos agrupar todos los nodos que pertenecen a la misma tarea en un "Meta-nodo". Por ejemplo, podemos crear el meta-nodo "dinero restante" que agrupa todos los nodos para los cálculos de los valores restantes.

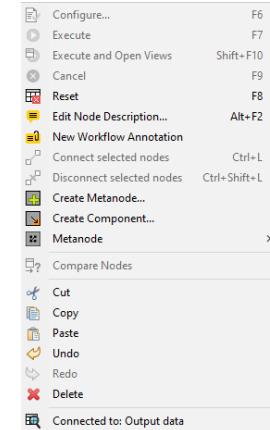
**Nota.** Un metanodo es un nodo (gris) que contiene otros nodos.

Hay dos maneras de construir un metanodo. La forma más fácil colapsa nodos preexistentes en un metanodo; la otra forma crea un metanodo desde cero.

### Colapsar nodos preexistentes en un Meta-nodo

- En el editor de flujo de trabajo, seleccione los nodos que formarán parte de los metanodos (para seleccionar varios nodos en Windows utilice las teclas Shift y Ctrl)
- Haga clic con el botón derecho en cualquiera de los nodos seleccionados
- Seleccione "Crear metanodo..."
- Se crea un nuevo meta-nodo con el sub-flujo de trabajo de los nodos seleccionados
- El número de puertos de entrada y salida se define automáticamente en función de los nodos seleccionados.

5.23. Opción "Create a Metanode..." en el menú contextual de los nodos seleccionados



En el flujo de trabajo "Proyectos" hemos seleccionado todos los nodos Java Snippet y Math Formula para que formen parte de un metanodo llamado "Remaining Money" con un puerto de entrada y otro de salida. El flujo de trabajo resultante se ha guardado como "Projects\_final".

# Crear un Meta-nodo desde cero

Un meta-nodo es un nodo que contiene un sub-flujo de trabajo de nodos. Un meta-nodo no realiza una tarea específica; es sólo un contenedor de nodos.

Para crear un "Meta-nodo":

- En la barra de herramientas, haga clic en el icono "Metanode".  
ó
- En el menú superior, haga clic en "Node" y seleccione "Open Meta Node Wizard" "Abrir asistente de metanodo".

Puede elegir entre una serie de estructuras de metanodos predefinidas (1 input - 1 output, 2 inputs - 1 output, etc.). Además, el botón "Customize""Personalizar" le permite ajustar cualquier estructura de metanodo seleccionada.

Para abrir un "Meta-node":

- Hacer doble click en "Meta-node"  
ó
- Click derecho en el "Meta-node" y seleccione "Open sub-workflow editor"

Se abre una nueva ventana del editor para que pueda editar el subflujo de trabajo asociado que contiene el "Meta-node".

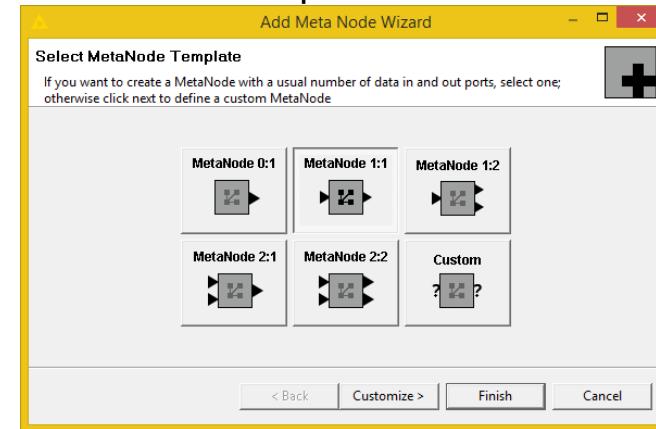
Para llenar un "Meta-node" con nodos:

- Arrastre y suelte los nodos desde el panel "Repositorio de nodos" como lo haría con un flujo de trabajo normal,  
ó
- Corte los nodos que ya existen en tu flujo de trabajo y pégalos en la ventana del editor de sub-flujos de trabajo

5.24. "Icono "Meta-node" en la barra de herramientas

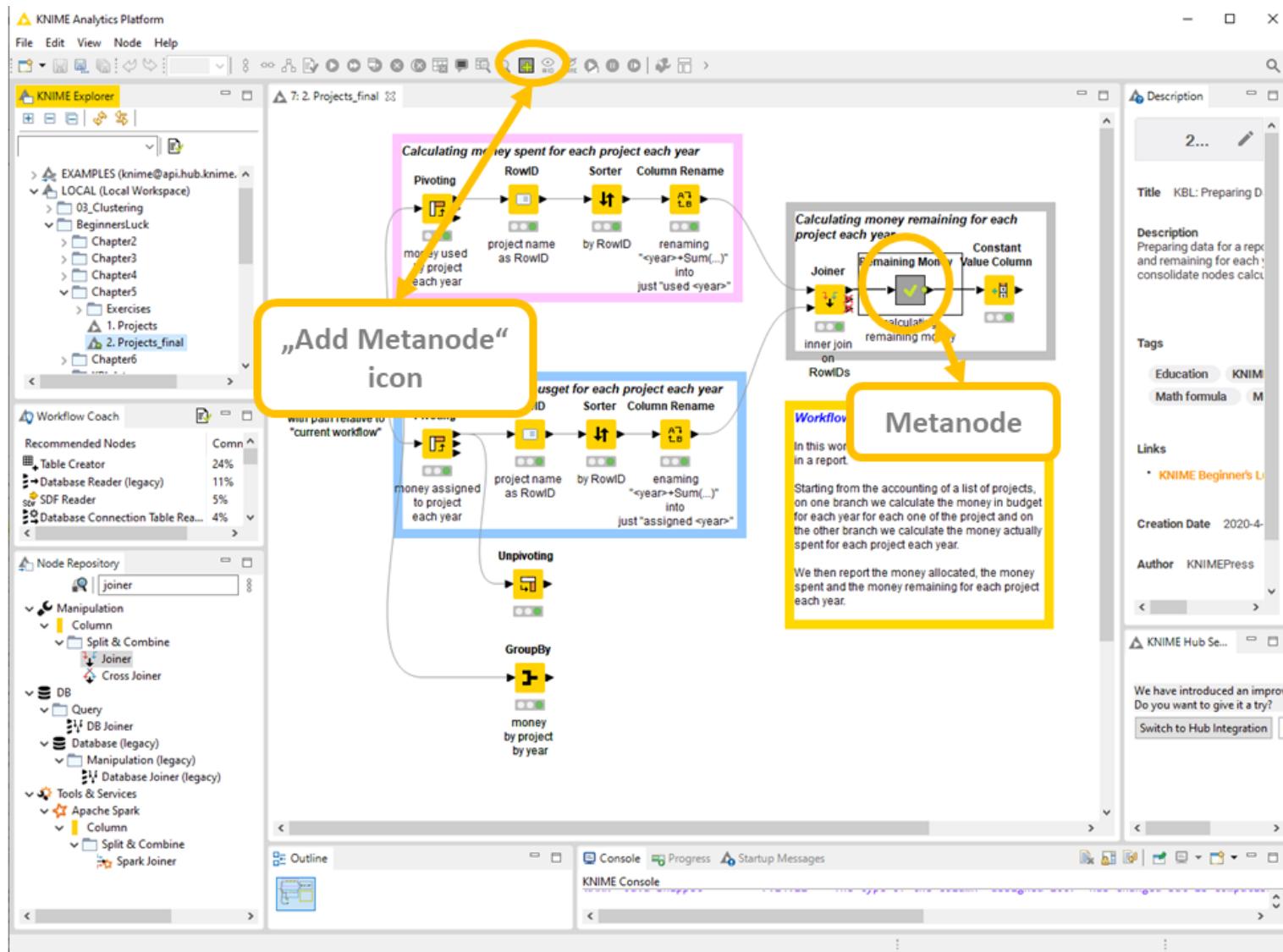


5.25. Estructuras predefinidas de metanodo



**Note.** El "Configure" está habilitado en un meta-nodo pero no es realmente utilizable, ya que no hay nada que configurar. Todos los demás comandos del nodo, como "Execute", "Reset", "Node name and description" etc..., se aplican de la forma habitual, como para cualquier otro nodo. En particular, el "Execute" and "Reset" ejecutan y reinician respectivamente todos los nodos dentro del meta-nodo.

## 5.26. Icono "add Metanode" and Metanode



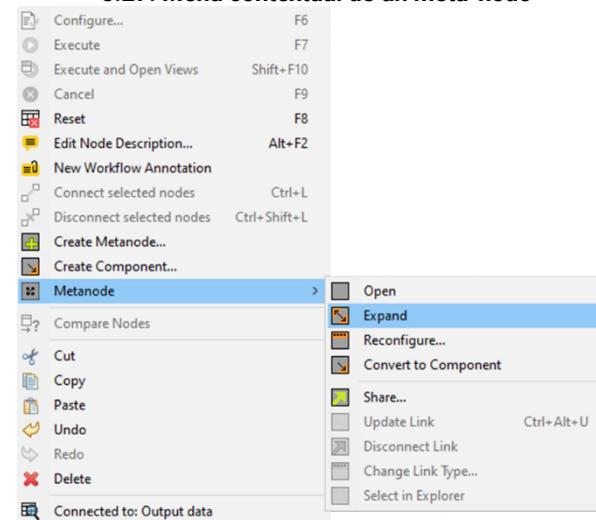
# Expandir y reconfigurar un metanodo

Una vez que el meta-nodo existe, es posible interactuar con él (reconfigure, expand, etc...) a través de su menú contextual.

El menú contextual de un meta-nodo es completamente similar al menú contextual de cualquier otro nodo, además de la opción "Meta Nodo". La opción "Meta nodo" abre un submenú con comandos aplicables sólo a un meta nodo, como por ejemplo:

- "Open", to open the meta-node content in the workflow editor
- "Expand", para reintroducir el contenido del meta-nodo en el flujo de trabajo principal y deshacerse del contenedor del meta-nodo
- "Reconfigure", para cambiar el meta-nodo en términos de puertos de entrada/salida y nombre
- "Wrap", para transformar el meta-nodo en un sub-nodo con opciones de la ventana de configuración
- "Save as Template", guardar el metanodo actual como plantilla en un repositorio central y permitir que otros usuarios lo utilicen con los permisos adecuados (esta opción sólo está disponible con una licencia comercial)

5.27. Menú contextual de un meta-nodo



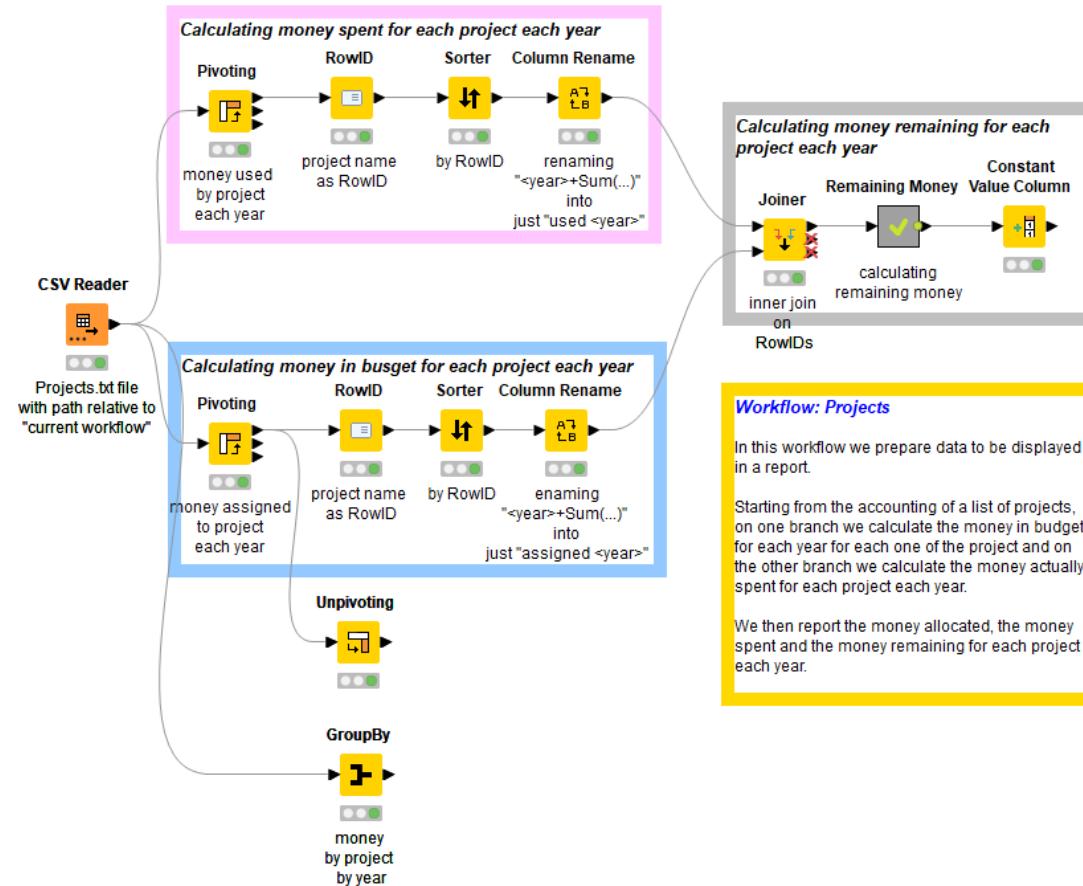
Es posible que haya notado la presencia de elementos relacionados con "Components" en el menú contextual. Un componente es un tipo especial de meta-nodo autónomo. Los componentes se describen en la continuación de este libro, en el "KNIME Advanced Luck".

**Nota.** La principal diferencia entre los metanodos y los componentes es el uso de vistas compuestas en el KNIME WebPortal, que se heredan de los nodos contenidos.

Creamos un nuevo "Metanodo" a partir de todos los nodos de la Math Formula y Java Snippet y lo llamamos "Remaining Money". Conectamos su puerto de entrada al puerto de salida del nodo "Joiner" y su puerto de salida al puerto de entrada del nodo "Constant Value Column".

Hay una serie de metanodos preconfigurados en el panel "Repositorio de nodos" de KNIME en algunas subcategorías "Metanodos" bajo algunas categorías principales, como "Workflow Control", "Mining", "R", "Time Series", y otras. Las categorías "Meta Nodos" contienen útiles implementaciones de meta nodos pre-empaquetados para la categoría principal.

### 5.28. El "Projects\_final" workflow



## 5.6. Próximo paso: Crear un Reporte

En este punto los datos están listos. Necesitamos construir el informe. Hay muchas opciones para hacerlo, a través de los nodos nativos de KNIME, así como a través de integraciones con herramientas de informes externas. Aquí enumeramos algunas opciones posibles, aunque seguramente hay más opciones posibles.

- Construir cuadros de mando a través de componentes KNIME y sus vistas compuestas, para ser visualizados en un navegador web a través de KNIME WebPortal
- Exportar datos a la solución de informes BIRT, de código abierto e integrada en KNIME Analytics Platform

- Exportar datos a las soluciones de informes Tableau, PowerBI o Spotfire a través de nodos dedicados. Para estas soluciones se necesita una licencia de pago. Existen nodos dedicados para exportar los datos a las soluciones. Aunque se proporcionan flujos de trabajo de ejemplo en la carpeta del Capítulo 7, no describiremos la construcción de dichos informes en detalle en este libro debido al requisito de una licencia de pago.
- Exporte los datos a un archivo CSV, o a otro archivo compatible, e impórtelos en su herramienta de informes preferida.

## 5.7. Ejercicios

### Ejercicio 1

Utilice los datos de entrada adult.data para hacer lo siguiente:

- Calcular el número total de personas con ingresos > 50K y el número total de personas con ingresos <= 50K para cada clase de trabajo
- Ordenar las filas de la columna "clase de trabajo" por orden alfabético
- Crear una tabla de datos con la siguiente estructura

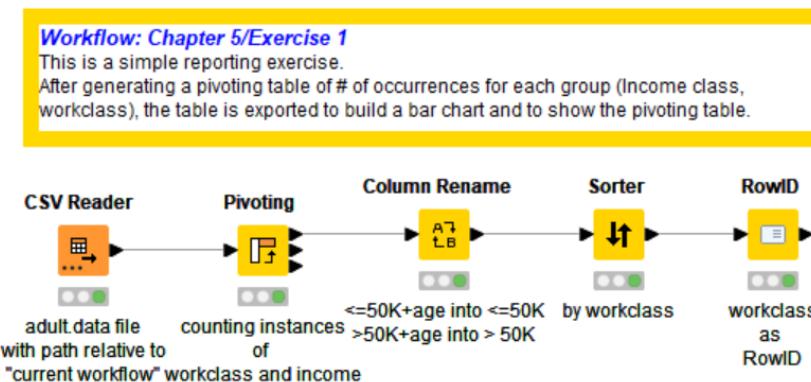
<b>Work class</b>	<b>Nr of people with Income &gt; 50K</b>	<b>Nr of people with Income &lt;= 50K</b>
<b>Work class 1</b>		
...		
<b>Work class n</b>		

- Marque este conjunto de datos para informar

#### Solución al ejercicio 1

1. Lee los datos.
2. Utilice un nodo "Pivote" para construir la tabla de datos en el formato solicitado. La columna "clase de trabajo" debe ser la columna de grupo y la columna "Ingresos" debe ser la columna pivotante.
3. Opcionalmente, cambie el nombre de los encabezados de las columnas para facilitar la lectura de la tabla.

### 5.29. Ejercicio 1: workflow



## Ejercicio 2

Ampliar el ejercicio 1

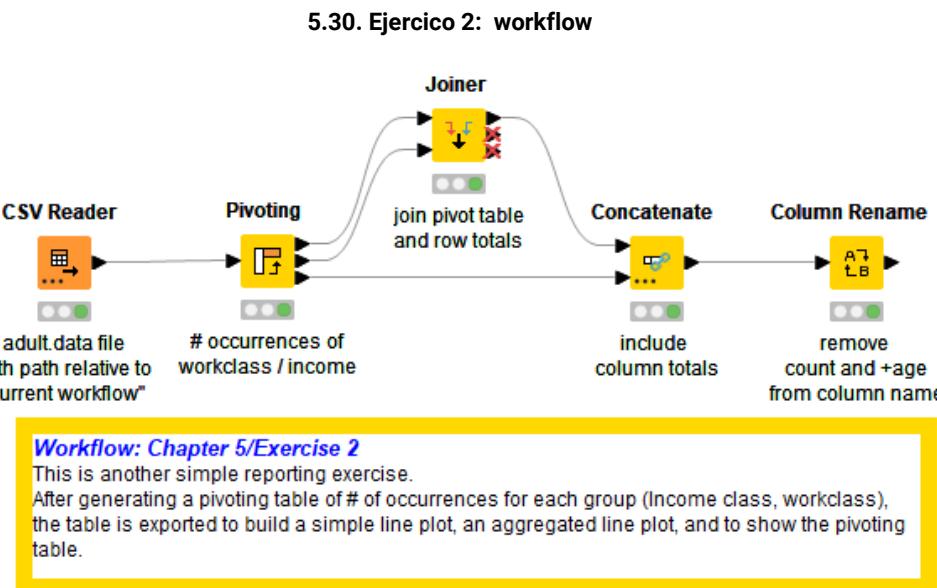
- Calcule el número total de personas con ingresos > 50K y con ingresos <= 50K
- Calcule el número total de personas para cada clase de trabajo
- Calcule el número total de personas
- Amplíe la tabla de datos elaborada para el ejercicio 1 de la siguiente manera:

Work class	Nr of people with Income > 50K	Nr of people with Income <= 50K	Nr of people
Work class 1			
...			
<b>total</b>	Sum(nr of people with Income > 50K)	Sum(nr of people with Income <= 50K)	Sum(nr of people)

### Solución al ejercicio 2

1. Para calcular el número de personas de cada "clase de trabajo" y de cada clase de renta, utilizamos el nodo "Pivotante" construido en el Ejercicio 1. El nodo "Pivoting" tiene tres salidas: la tabla pivotante, los totales por fila y los totales por columna. Recuerde activar la opción "Anexar totales globales" en la pestaña "Pivotes".
2. A continuación, unimos internamente en los valores de la "clase de trabajo" la tabla pivotante con los totales por fila utilizando un nodo "Joiner".

3. Luego concatenamos la tabla de datos resultante del nodo "Joiner" con los totales por columna del nodo "Pivote".



## Ejercicio 3

Lea el archivo csv SoccerWorldCup2006.txt de la carpeta KBLdata. Este archivo describe los resultados de los partidos de fútbol durante el mundial de fútbol 2006 ([www.fifa.com](http://www.fifa.com)). No se informa del segundo partido de semifinales para el tercer y cuarto puesto.

Para cada equipo calcular:

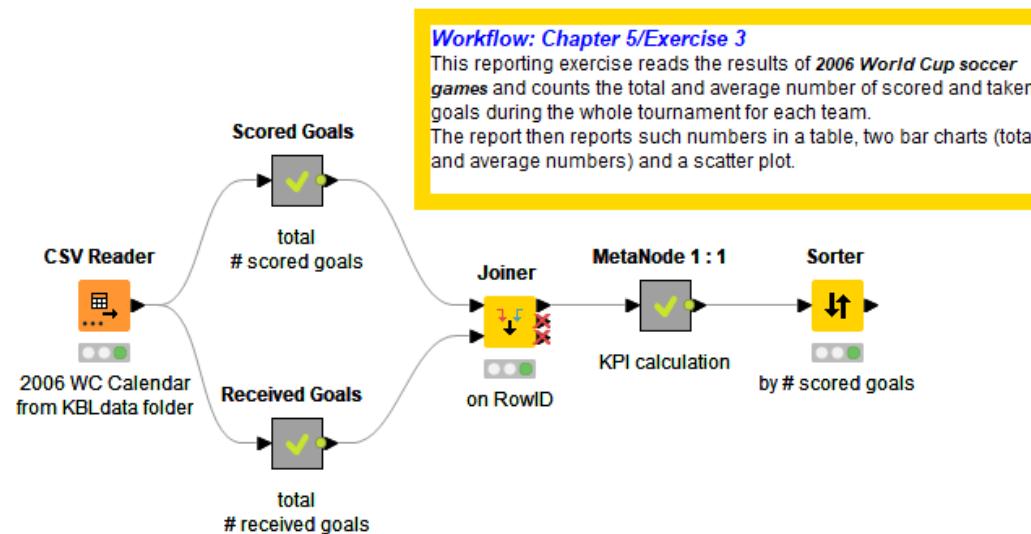
- El número total de partidos jugados
- El número total de goles marcados
- El número total de goles encajados
- La media de goles marcados por partido
- El número medio de goles encajados por partido
- Una medida de ajuste como: (número total de goles marcados - número total de goles encajados)/número de partidos jugados

Documente cada paso con el nombre y la descripción del nodo correspondiente.

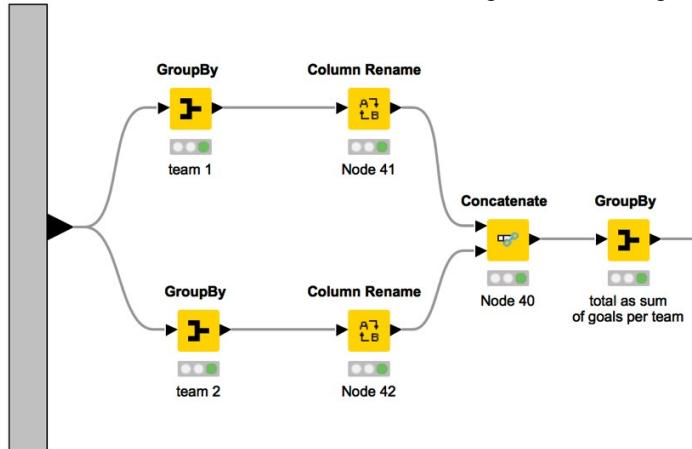
Haz que el flujo de trabajo sea legible utilizando metanodos.

### Solution to Exercise 3

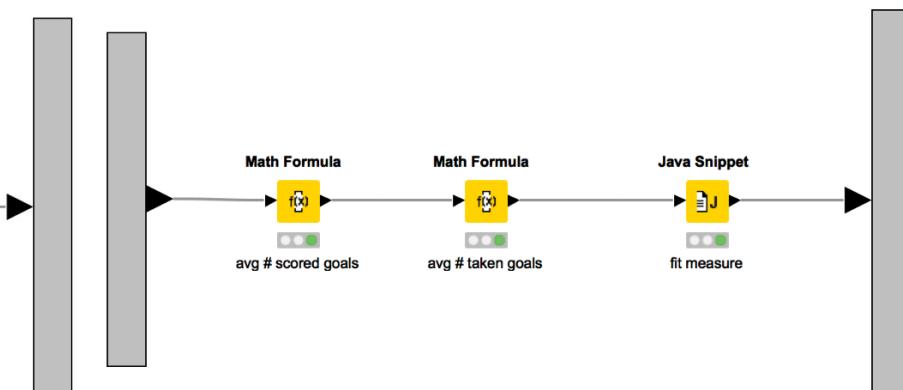
#### 5.31. Ejercicio 3: workflow



5.32. Meta-node "# scored goals"/"# taken goals"



5.33. Meta-node "KPI Calculation"



En el "# scored goals" meta-node, sumamos las puntuaciones del equipo 1 sobre todos los equipos 1, luego la suma de las puntuaciones del equipo 2 sobre todos los equipos 2, y finalmente sumamos las puntuaciones totales del equipo 1 y del equipo 2 cuando el equipo 1 = el equipo 2.

El Meta-node "# taken goals" Tiene la misma estructura que el Meta-node "# scored goals". La única diferencia radica en la variable de agregación de los dos primeros "nodos GroupBy". En el meta-nodo "# goles marcados" el primer nodo "GroupBy" suma la "puntuación del equipo 1" para todos los valores del "equipo 1" y el segundo nodo "GroupBy" suma la "puntuación del equipo 2" para todos los valores del "equipo 2". En el meta-nodo "# goles encajados", el primer nodo "GroupBy" suma la "puntuación del equipo 2" para todos los valores del "equipo 1" y el segundo nodo "GroupBy" suma la "puntuación del equipo 1" para todos los valores del "equipo 2".

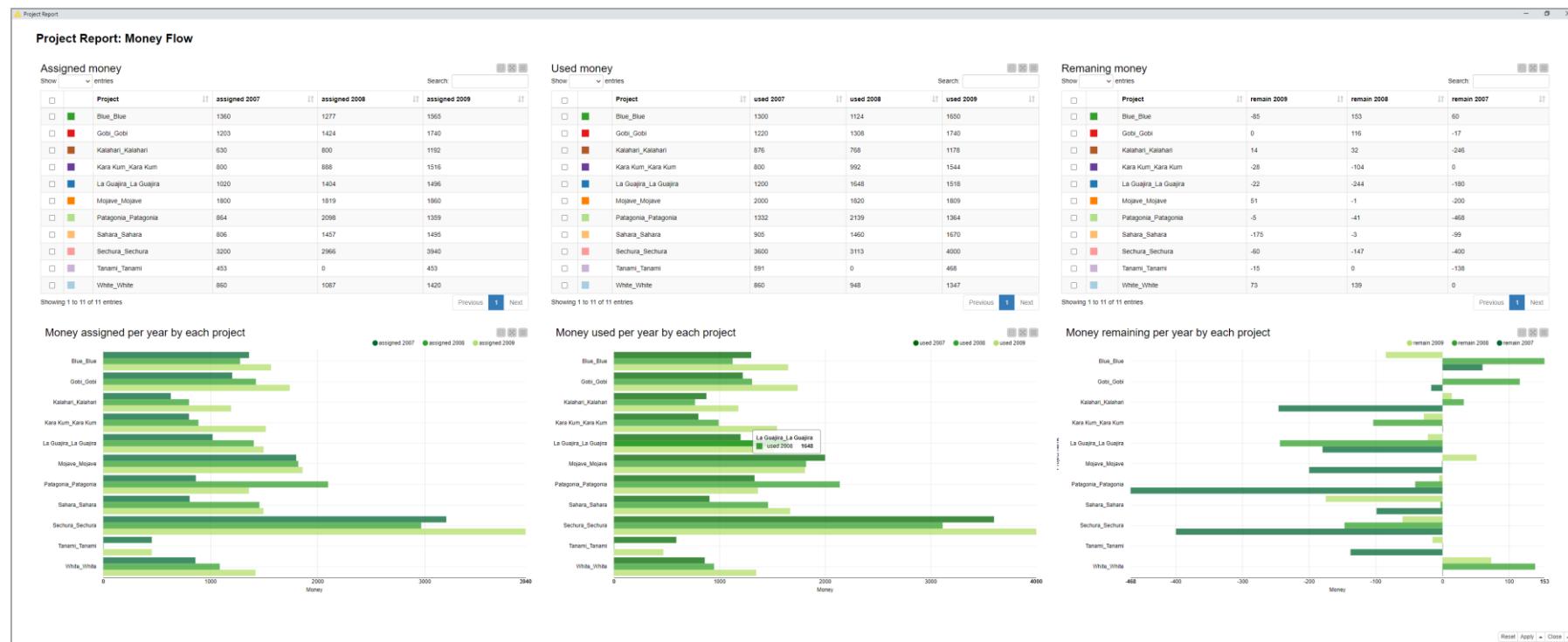
En el meta-nodo "Cálculo de KPI" hemos utilizado 2 nodos "Math Formula" y un nodo "Java Snippet". Podría haber sido cualquier otra combinación de nodos "Java Snippet" y "Math Formula".

# Capítulo 6. Tableros de mando con vistas compuestas

## 6.1. El tablero de mando (Dashboard)

A partir de las tablas de datos elaboradas en el capítulo anterior, construimos aquí un sencillo cuadro de mando, que incluye una tabla de dinero asignado, una tabla de dinero utilizado y una tabla de dinero restante en todos los proyectos a lo largo de los tres años de observación y los correspondientes tres gráficos de barras.

6.1. Panel final que visualiza las tablas y los gráficos de barras del dinero asignado, utilizado y restante en los 11 proyectos y en los tres años de observación



El cuadro de mandos, por supuesto, debe ser interactivo, es decir, debe ofrecer tantas opciones de personalización como sea posible al usuario final mientras se muestra en un navegador web. Las opciones de personalización deben incluir:

- Cambiar los títulos, subtítulos y otras etiquetas de cada tabla y gráfico de barras
- Seleccionar un proyecto y visualizar todos los datos relativos a ese proyecto en todas las tablas y gráficos
- Paginación de las tablas, si es necesario
- Acercamiento y alejamiento de elementos individuales del cuadro de mandos

## 6.2. Los Nodos

En el capítulo 3, dedicado a la visualización de datos, ya hemos visto el nodo Table View y el nodo Bar Chart. Como su nombre indica, el nodo Table View produce una visualización basada en una tabla, y el nodo GBar Chart produce una visualización en forma de gráfico de barras de los datos de entrada. Como se ha comentado anteriormente, todos estos nodos de visualización tienen unas pestañas en su ventana de configuración: tres el nodo de Table Viewer y cuatro el nodo de Bar Chart.

### Pestañas de la ventana de configuración del nodo "Table View's"

- *Options*. Esta pestaña establece las columnas de entrada que se mostrarán en la tabla junto con algunos otros ajustes menores, como el título y el subtítulo, y la visualización de colores, claves y encabezados. Decidimos incluir los nombres de los proyectos en la columna "Project" como primera columna desde la izquierda en la tabla.
- *Interactivity*. Esta pestaña incluye casillas para habilitar las opciones de interactividad, como las opciones de paginación, las opciones de selección para las filas de datos, así como las opciones de búsqueda y ordenación para las tablas con un número muy elevado de filas. Tenemos 11 proyectos, y nos gustaría poder verlos todos a la vez. Por lo tanto, aquí establecemos *Initial Page Size* 11.
- *Formatters*. Esta pestaña define cómo representar las fechas, los números y los valores perdidos en las celdas de la tabla.

### Pestañas de la ventana de configuración del nodo Chart's

- *Options*. De nuevo, esta pestaña establece las columnas de entrada que van a participar en la visualización, como sólo un recuento de ocurrencias, una suma o un promedio. Dado que nuestros números son únicos y se supone que se muestran tal cual, podemos elegir la opción Average ó Sum La columna de categoría es "Project" y contiene los nombres de los proyectos. Opcionalmente, los nombres de las categorías pueden ser ordenados alfabéticamente. Sin embargo, nuestras filas de datos ya están ordenadas por nombre de proyecto en orden alfabético. La casilla "Generate image" permite crear el gráfico como una imagen en el puerto de salida. Aunque es útil para algunas tareas, esto puede consumir mucho tiempo y recursos.
- *General Plot Options*. Esta pestaña incluye todos los ajustes del gráfico: título, subtítulo, etiquetas de los ejes, visualización de elementos y tamaño de la imagen creada opcionalmente.

- *Control Options*. Esta pestaña incluye casillas de verificación para la personalización del gráfico, como la edición del título, el subtítulo y la etiqueta del eje, el cambio de la orientación de las barras y la agrupación de barras frente al apilamiento, y otros ajustes de la vista.
- *Interactivity*. Esta pestaña completa la lista de casillas para la interactividad, cubriendo las opciones de selección de barras en el gráfico.

El último nodo que falta para completar este sencillo cuadro de mando es un título. Para ello utilizamos el nodo “Text Output Widget” .

## Text Output Widget

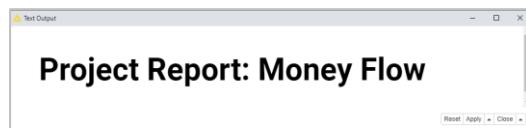
Este nodo sólo emite texto en una vista gráfica.

- “Text” contiene el texto a emitir.
- “Text format” contiene el tipo de texto y cómo interpretarlo. Son posibles tres tipos de formato de texto: texto simple, texto preformatado y texto HTML.

El texto HTML interpreta y visualiza las instrucciones HTML. Así, la línea:

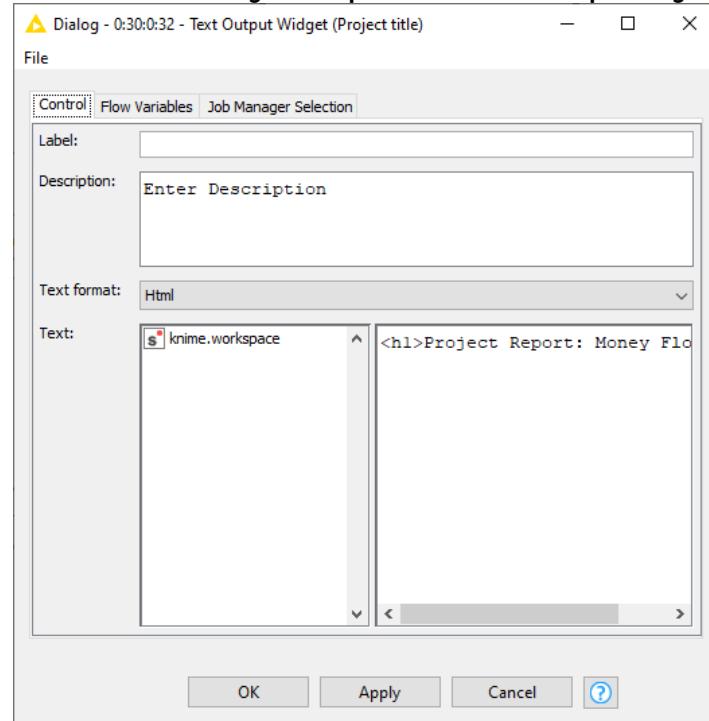
```
<h1>Project Report: Money Flow</h1>
```

Es visualizada como:



Este nodo tiene un puerto de entrada opcional de Variable de Flujo. Sin embargo, las Variables de Flujo no se describen en este libro, y no las necesitamos para este ejemplo en particular.

6.2. Ventana de configuración para el nodo “Text Output Widget”



En conclusión, para construir el cuadro de mandos mostrado arriba, se necesita:

- Tres nodos Table View
- Tres nodos Bar Charts
- Un nodo Text Output Widget

El flujo de trabajo desarrollado en el capítulo anterior, denominado Projects, se ha importado en la carpeta Chapter6 y se ha renombrado como "Dashboard". A continuación se han creado los siete nodos necesarios. Observe que hemos añadido un nodo RowID para copiar los nombres de los proyectos de los RowIDs en una columna de datos llamada "Project", y que hemos añadido un nodo Column Resorter para colocar la columna "Project" en la parte izquierda de la tabla de datos de entrada. Observe también que el nodo Text Output Widget no necesita ninguna entrada y por lo tanto se dejó flotando libremente. Ahora, vamos a ensamblar todos estos nodos juntos en un componente para crear el panel de control

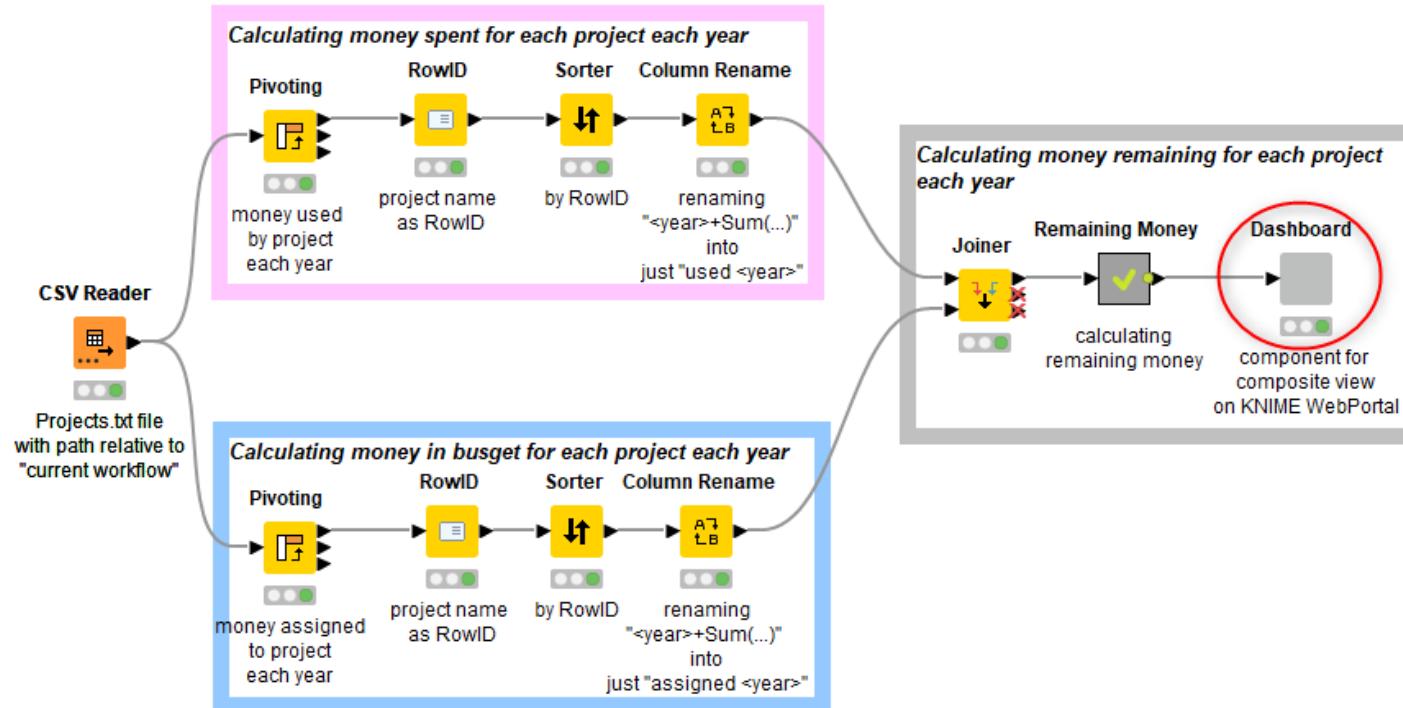
## 6.3. El Componente

¿Recuerdas el metanodo descrito en el capítulo 5? Los componentes representan la evolución natural de los metanodos. Los componentes se parecen a los metanodos en el sentido de que recogen nodos en su interior. Al igual que los metanodos, para crear un componente se debe:

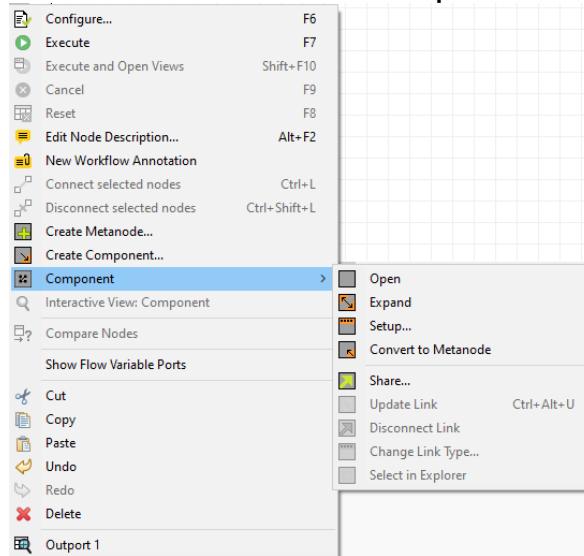
- seleccionar todos los nodos de interés haciendo clic y dibujando un rectángulo alrededor de ellos o pulsando la tecla Mayúsculas/Ctrl-clic en cada uno de ellos
- hacer clic con el botón derecho y seleccione "Create Component..."
- darle un nombre

y su nuevo componente será creado. Nosotros lo hemos llamado "Dashboard".

### 6.3. Nuevo flujo de trabajo con el componente "Dashboard" al final del flujo de datos para su visualización



#### 6.4. Menú contextual de un componente



Al igual que los metanodos, si hace clic con el botón derecho del ratón en el componente y luego selecciona "Componente", puede abrir el submenú del componente. Aquí puede encontrar las siguientes opciones:

- "Open" para abrir e inspeccionar el contenido del componente
- "Expand" para eliminar el componente y volver a colocar los nodos en el flujo de trabajo original
- "Set up" para cambiar algunos de los ajustes del componente, como por ejemplo el nombre o los puertos de entrada y salida en número y tipo. "Configurar" es equivalente a la opción "Reconfigurar..." para los metanodos.
- "Convert to Metanode" para volver a la estructura familiar de un metanodo
- "Share" para crear una plantilla en un espacio de trabajo local o remoto que pueda reutilizarse posteriormente para crear instancias enlazadas de este mismo componente.

Hasta ahora, un componente se parece mucho a un metanodo. ¿Qué puede hacer un componente que no pueda hacer un metanodo?

- **Un componente es un entorno encapsulado.** Todavía no hemos hablado de las variables de flujo. Sin embargo, podemos describir un componente como un entorno de vacío que sólo deja entrar y salir datos y nada más.
- **Un componente puede tener una ventana de configuración.** Los componentes pueden tener una ventana de configuración, los metanodos no. La inserción de uno o más nodos de la carpeta "Workflow Abstraction/Configuration" proporciona uno o más elementos para la ventana de configuración del componente. Un componente es una forma de crear un nuevo nodo sin necesidad de codificar. Todas las plantillas de nodos en "EXAMPLES/00\_Components" en el panel KNIME Explorer (en el KNIME Hub) son en realidad componentes que tienen una ventana de configuración.
- **Un componente puede recibir una vista.** Los componentes pueden obtener una vista, los metanodos no. La inserción de uno o más nodos de la carpeta "Workflow Abstraction/Widgets" proporciona uno o más elementos para la vista de su componente. Las vistas interactivas de estos nodos se pasan a la vista interactiva del componente. Las vistas con muchos elementos de muchos nodos widget correspondientes se denominan vistas compuestas. Además, las vistas de nodos widget dentro de la misma vista compuesta se suscriben a la selección y visualización de los mismos datos. Esto significa que lo que se selecciona en la vista de un gráfico, por ejemplo, también se selecciona (y se puede visualizar exclusivamente) en la vista de otro gráfico dentro de la vista del mismo componente. Esta es la parte de los componentes que nos interesa.

Todos los nodos azules del componente "Cuadro de mando" producen una vista gráfica. Así, el nuevo componente "Dashboard" tiene una vista compuesta por tres tablas, tres gráficos de barras y un texto que funciona como título. Esta es la vista compuesta del componente.

## 6.4. Agregando Colores

Lo que hemos construido hasta ahora está en colores blanco y negro. Esta última parte está dedicada a cómo introducir colores en los gráficos y diagramas de KNIME. Utilizaremos los colores para identificar los proyectos en las tablas y para identificar los años en los gráficos de barras.

Para asignar una propiedad de color a cada proyecto sólo tenemos que pasar los datos originales por el nodo Administrador de colores. Esto asigna automáticamente un color a cada fila de datos. Los colores también se pueden personalizar manualmente dentro de la ventana de configuración del nodo. Las filas de datos con colores llegan al nodo Vista de Tabla y se representan cada una con su propio color a la izquierda.

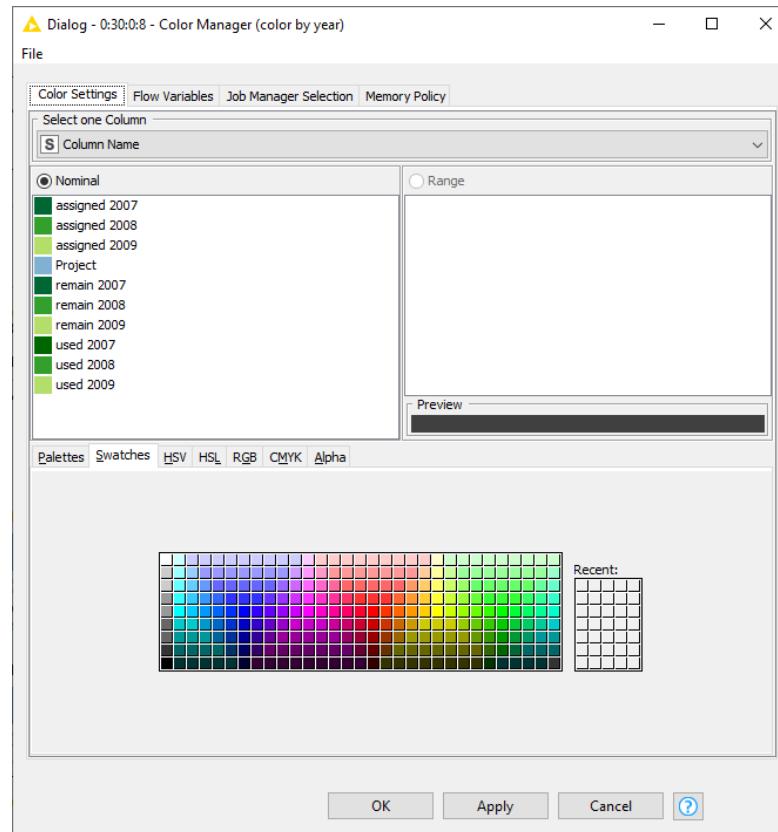
6.5. Tabla con los colores asignados por nombre de proyecto

Remaining money						
		Project	remain 2009	remain 2008	remain 2007	
<input type="checkbox"/>	<span style="background-color: green; border: 1px solid black; padding: 2px 5px;"> </span>	Blue_Blue	-85	153	60	
<input type="checkbox"/>	<span style="background-color: red; border: 1px solid black; padding: 2px 5px;"> </span>	Gobi_Gobi	0	116	-17	
<input type="checkbox"/>	<span style="background-color: brown; border: 1px solid black; padding: 2px 5px;"> </span>	Kalahari_Kalahari	14	32	-246	
<input type="checkbox"/>	<span style="background-color: purple; border: 1px solid black; padding: 2px 5px;"> </span>	Kara Kum_Kara Kum	-28	-104	0	
<input type="checkbox"/>	<span style="background-color: blue; border: 1px solid black; padding: 2px 5px;"> </span>	La Guajira_La Guajira	-22	-244	-180	
<input type="checkbox"/>	<span style="background-color: orange; border: 1px solid black; padding: 2px 5px;"> </span>	Mojave_Mojave	51	-1	-200	
<input type="checkbox"/>	<span style="background-color: lightgreen; border: 1px solid black; padding: 2px 5px;"> </span>	Patagonia_Patagonia	-5	-41	-468	
<input type="checkbox"/>	<span style="background-color: yellow; border: 1px solid black; padding: 2px 5px;"> </span>	Sahara_Sahara	-175	-3	-99	
<input type="checkbox"/>	<span style="background-color: pink; border: 1px solid black; padding: 2px 5px;"> </span>	Sechura_Sechura	-60	-147	-400	
<input type="checkbox"/>	<span style="background-color: purple; border: 1px solid black; padding: 2px 5px;"> </span>	Tanami_Tanami	-15	0	-138	
<input type="checkbox"/>	<span style="background-color: lightblue; border: 1px solid black; padding: 2px 5px;"> </span>	White_White	73	139	0	
Showing 1 to 11 of 11 entries						<span>Previous</span> <span>1</span> <span>Next</span>

Para asignar un color a cada año tenemos que hacer que las cabeceras de las columnas ("assigned 2009, "used 2009", ...) pasen por el nodo "Color Manager" como filas de datos. Para ello, utilizamos el nodo "Extract Table Specs". Este nodo extrae las propiedades de las columnas - encabezados de columna, valores máximos y mínimos, etc... y pone dicha información, incluyendo los nombres de las cabeceras de las columnas, en las filas de datos. Esta tabla puede utilizarse para asignar colores a las columnas anuales a través del

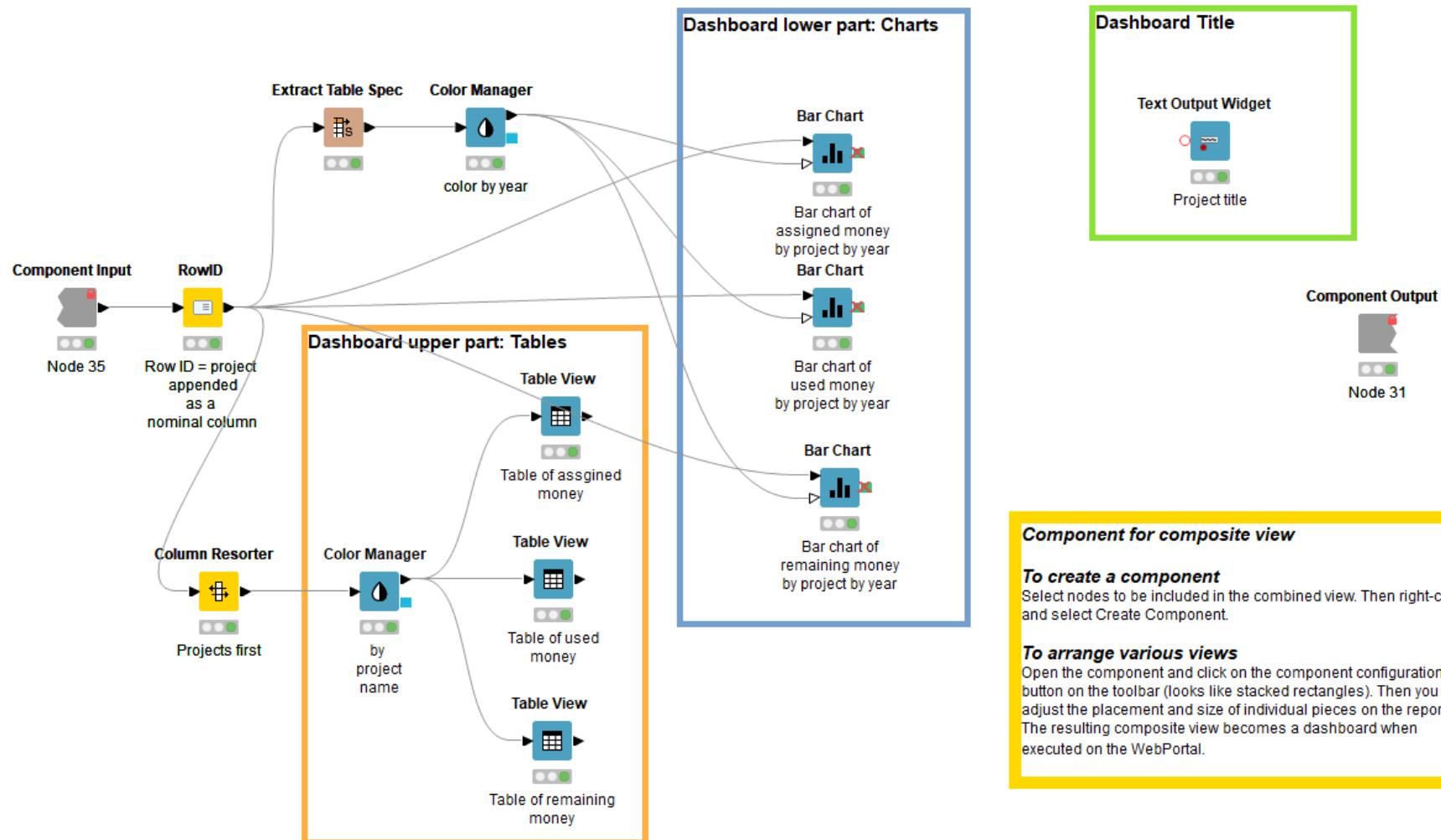
nodo Color Manager. Aquí, fuimos un poco más creativos gráficamente, y asignamos verde oscuro a todas las columnas de 2009, verde medio a todas las columnas de 2008, y verde claro a todas las columnas de 2007. Este mapa se envía al segundo puerto de entrada del nodo Bar Chart para colorear las barras del gráfico. De hecho, el segundo puerto de entrada del nodo Bar Chart, y de muchos otros nodos de visualización, requiere un mapa (<column header>, <column color>) para transferirlo al gráfico.

#### 6.6. Tonos de verde asignados a las columnas anuales en el nodo Gestor de Colores



El contenido final del componente "Dashboard" se muestra en la Fig. 6.7. Observe los dos nodos del Color Manager, uno para los nombres de los proyectos en las tablas y otro para las cantidades anuales en los gráficos de barras. Observe también el nodo flotante Text Output Widget.

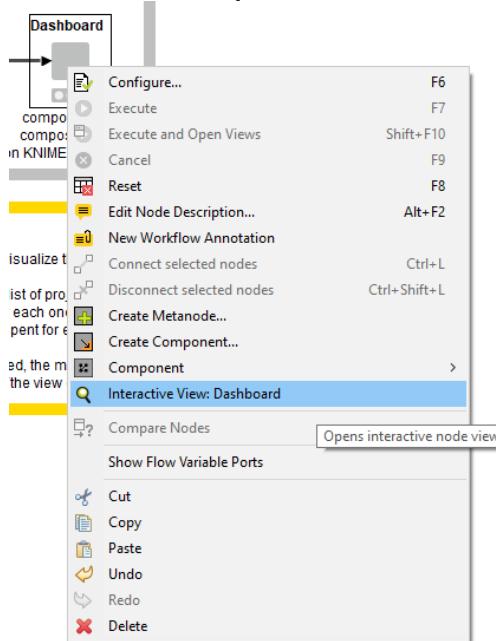
## 6.7. Contenido del componenete "Dashboard"



## 6.5. La Vista compuesta (Composite View)

Inspeccionemos ahora la vista interactiva compuesta del componente final

### 6.8. Como abrir la "composite view" de un componente



Para abrir la vista interactiva de un componente, haga clic con el botón derecho del ratón en el componente y seleccione “Interactive View: <name of component>”.

Si abre ahora la vista interactiva del componente “Dashboard”, verá sólo una larga lista de tablas y gráficos de barras sin un diseño organizado. Un buen diseño podría ser

- Un título
- Una fila con las tres tablas
- Una fila con los tres gráficos de barras

El diseño de una vista compuesta se decide a través del botón Diseño de la barra de herramientas en la parte superior del banco de trabajo KNIME. Despues de abrir el contenido del componente, haga clic en el botón de diseño para organizar los elementos de la vista compuesta.

Se abre el editor de diseño.

### 6.9. The Layout button in the tool bar in the KNIME workbench



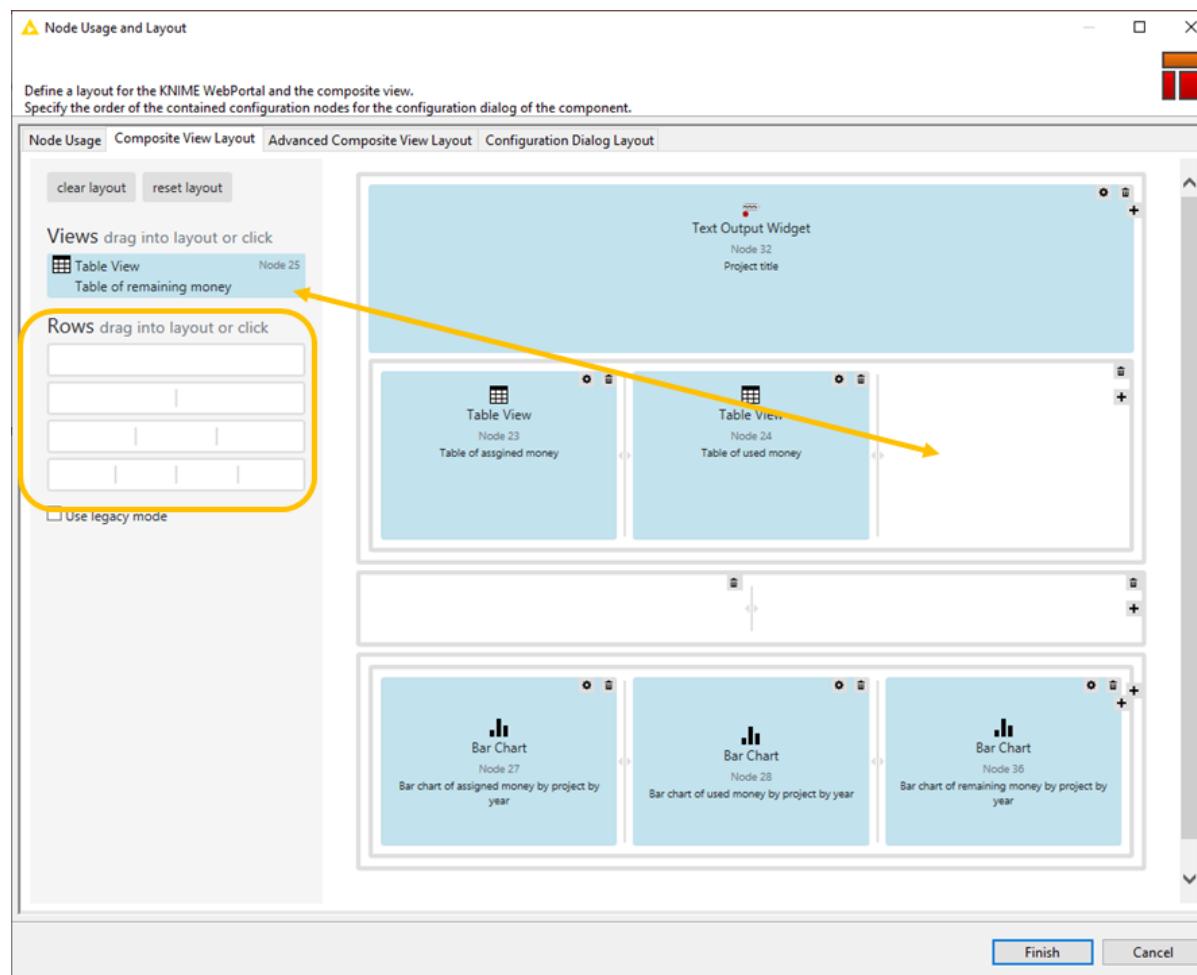
En el editor de diseño verá a la izquierda todos los elementos de la vista pertenecientes a los nodos de visualización dentro del componente y un conjunto de posibles cuadrículas de filas. Todos estos elementos pueden ser movidos a la vista final mediante drag&drop. Nos gustaría tener

- una fila completa dedicada al título, que es la vista producida por el nodo Text Output Widget
- una fila con tres celdas para las tres tablas
- una fila con tres celdas para los tres gráficos de barras

Arrastramos y soltamos la fila con una sola celda en la parte superior; luego la fila con tres celdas; luego de nuevo, la fila con tres celdas. A continuación, arrastramos y soltamos el elemento "Widget de salida de texto" en la primera fila, la "Table View" del nodo "Table of Assigned Money" en la primera celda de la segunda fila, y así sucesivamente hasta que todas las celdas estén pobladas como se desea.

Observe la papelera, el signo más y la rueda de ajuste en la esquina superior derecha de cada celda y fila, respectivamente, para añadir una celda/fila más, eliminar la celda/fila actual y personalizarlas.

#### 6.10. El editor de diseño



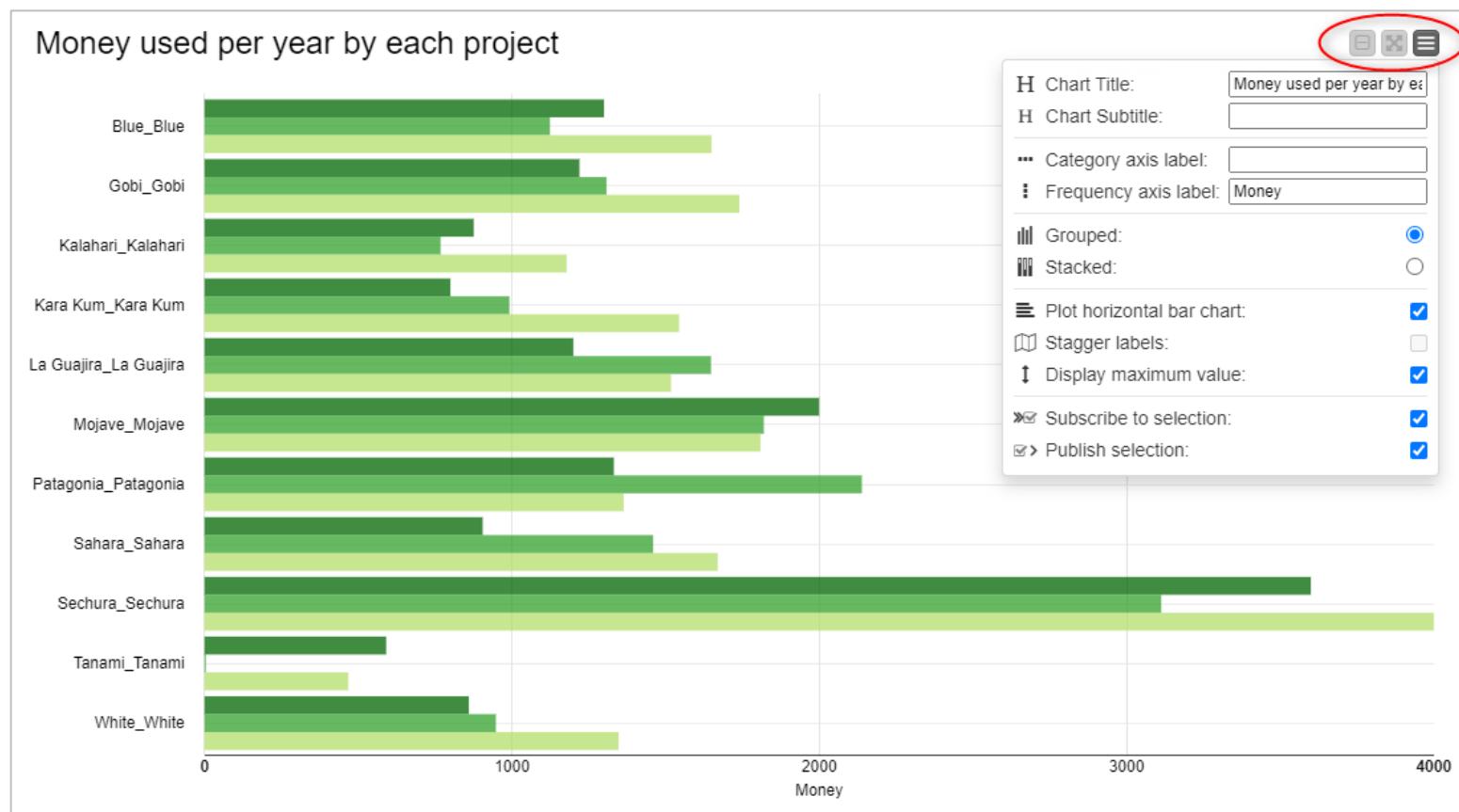
La vista ahora con el nuevo diseño debería parecerse al tablero mostrado en la Fig. 6.1, con un añadido importante: la interactividad. Veamos las opciones de interactividad derivadas de los ajustes de "Interactivity" en las ventanas de configuración.

Para abrir la vista compuesta de un componente, de nuevo, haga clic con el botón derecho del ratón en el componente y seleccione "Vista interactiva: <name of component>". La vista se abre y contiene las vistas individuales de cada uno de los nodos de visualización dentro del componente.

### Personalizar cada elemento de la vista compuesta

Si nos centramos en el elemento de la vista individual en la vista compuesta, observamos tres botones en la esquina superior derecha.

6.11. Uno de los gráficos de barras en la vista compuesta del componente "Dashboard"



Desde la derecha, el primer botón permite cambiar todos los ajustes del gráfico/tabla/gráfico, como los títulos y etiquetas, el modo de visualización, etc... En algunos gráficos incluso permite cambiar las columnas informadas en los ejes x e y.

El segundo botón permite ampliar el gráfico actual a pantalla completa.

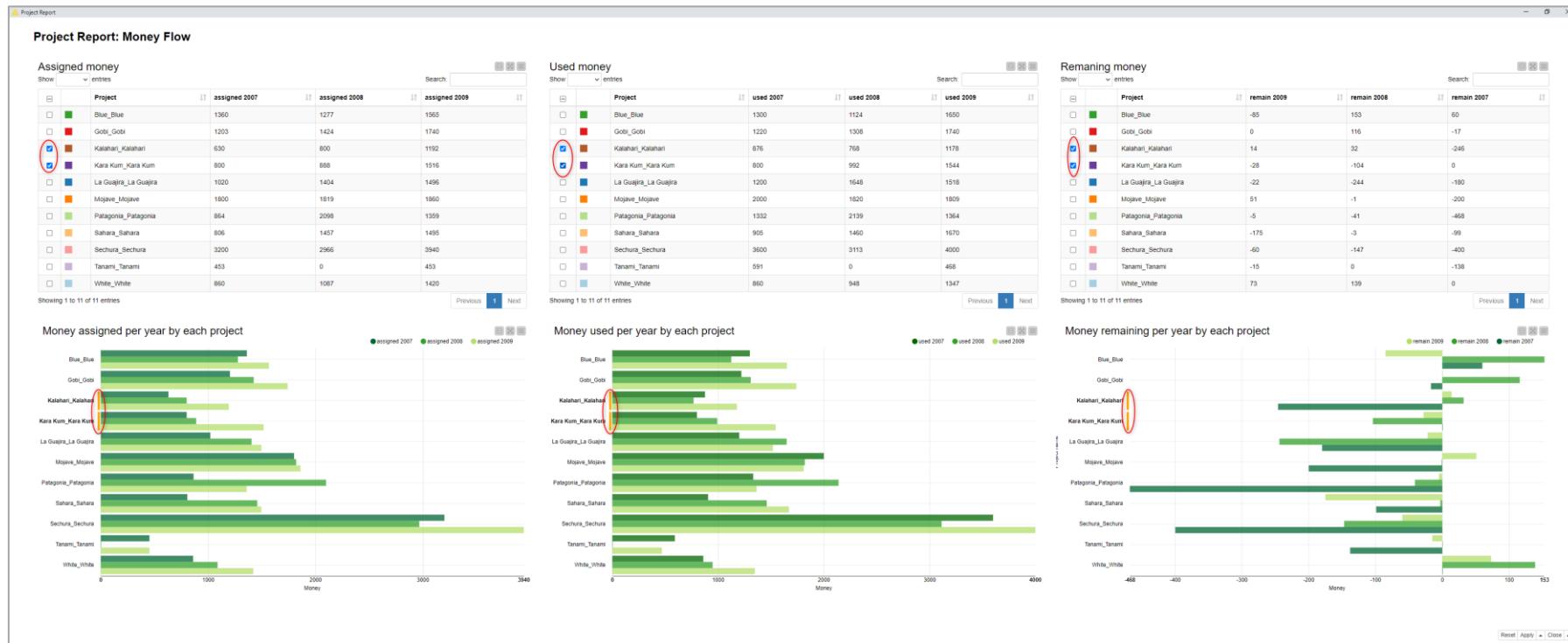
El tercer botón borra todos los ajustes anteriores.

### **Selecting and visualizing data rows**

Muchos gráficos y diagramas ofrecen información sobre herramientas al pasar el ratón por encima del área del gráfico o diagrama. En el caso del gráfico de barras, al pasar el ratón por encima de las barras aparece un tooltip con el número exacto que representa la barra.

Además, las barras/puntos/filas de un gráfico/parcela/tabla pueden seleccionarse y la misma selección aparece en todos los demás elementos de la vista si la suscripción y la publicación están activadas en sus ajustes de configuración. Por ejemplo, he seleccionado el proyecto Kalahari y Kara Kum en la tabla central. La misma selección aparece automáticamente en todas las demás tablas, así como en todos los gráficos de barras. Además, muchas vistas simples, como la vista de tabla de esta vista compuesta, ofrecen la opción de visualizar sólo las filas seleccionadas en el menú de configuración del botón superior derecho.

## 6.12. Selección de puntos en todos los elementos de la vista



## 6.6. En el WebPortal

Este flujo de trabajo, al igual que todos los demás flujos de trabajo, puede ser transferido al servidor KNIME para su producción. Allí, las vistas compuestas de los componentes se convierten en páginas web en el KNIME WebPortal.

Después de iniciar la sesión en un servidor KNIME desde un navegador web a través de su WebPortal y después de iniciar la ejecución del flujo de trabajo, la vista compuesta del primer componente del flujo de trabajo aparece en el navegador web en forma de página web; después de interactuar con la página si es necesario y pulsar "Siguiente", aparece la vista compuesta del siguiente componente del flujo de trabajo, y así sucesivamente.

En nuestro caso, nuestro flujo de trabajo (Fig. 6.3) sólo tiene un componente con una vista compuesta. Es decir, iniciar el flujo de trabajo desde el KNIME WebPortal nos llevará a la única página web generada por la vista compuesta del componente "Dashboard". Aquí podemos interactuar con todos los elementos de la vista exactamente como lo hicimos con la vista local de la vista compuesta. Observe que podrían introducirse nodos de visualización de datos y widgets más complejos en el componente para obtener una experiencia aún más interactiva. Dado que este libro sólo pretende introducir al lector en las características básicas de KNIME Analytics Platform, nos detendremos aquí. Sin embargo, tenga en cuenta que características más avanzadas - como barras de deslizamiento para filtrar, botones de actualización, y más - también podrían ser implementadas dentro del componente.

## 6.7. Ejercicios

Los ejercicios de este capítulo son la continuación de los ejercicios del capítulo 5. En concreto, requieren la elaboración de un informe para los conjuntos de datos creados en los ejercicios del capítulo 5.

### Ejercicio 1

Utilizando el flujo de trabajo construido en el Capítulo 5-Ejercicio 1, construya un informe con:

- Un título "ingresos por clase de trabajo"
- Una tabla en el lado izquierdo como:

Work Class	Income <= 50K	Income > 50K
[workclass]	[no <= 50K]	[no > 50K]

- Con colores asignados a cada clase de trabajo.
- Un gráfico de barras a la derecha con:
  - Clase de trabajo en el eje x
  - Income <= 50K" and "Income > 50K" on the y-axis
  - Legend available en el eje de ordenadas
  - Leyenda disponible
  - Sin título
  - Sin títulos en los ejes
  - Color verde para las barras <=50K y color rojo para las barras >50K

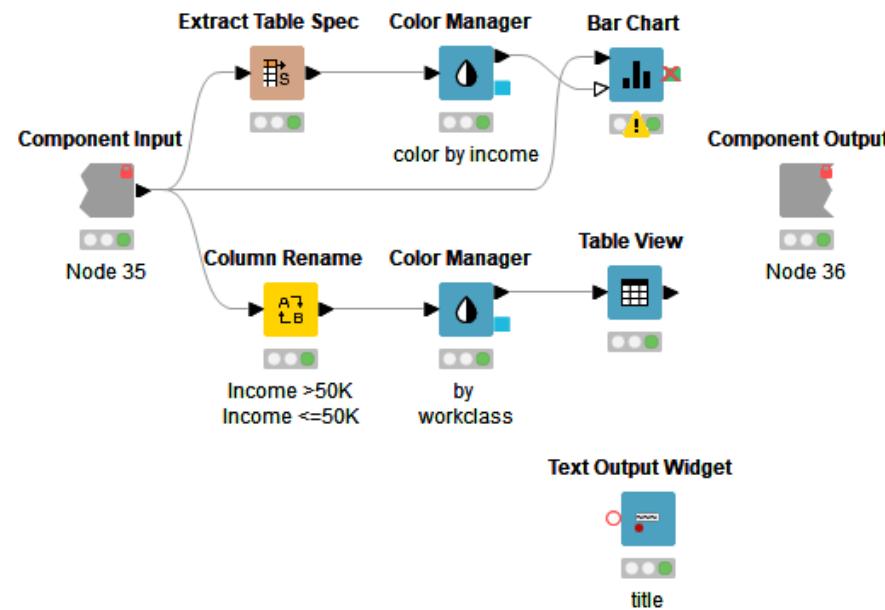
- Seleccione "nunca trabajó" "never worked" tanto en la tabla como en el gráfico de barras

## Solución al Ejercicio 1

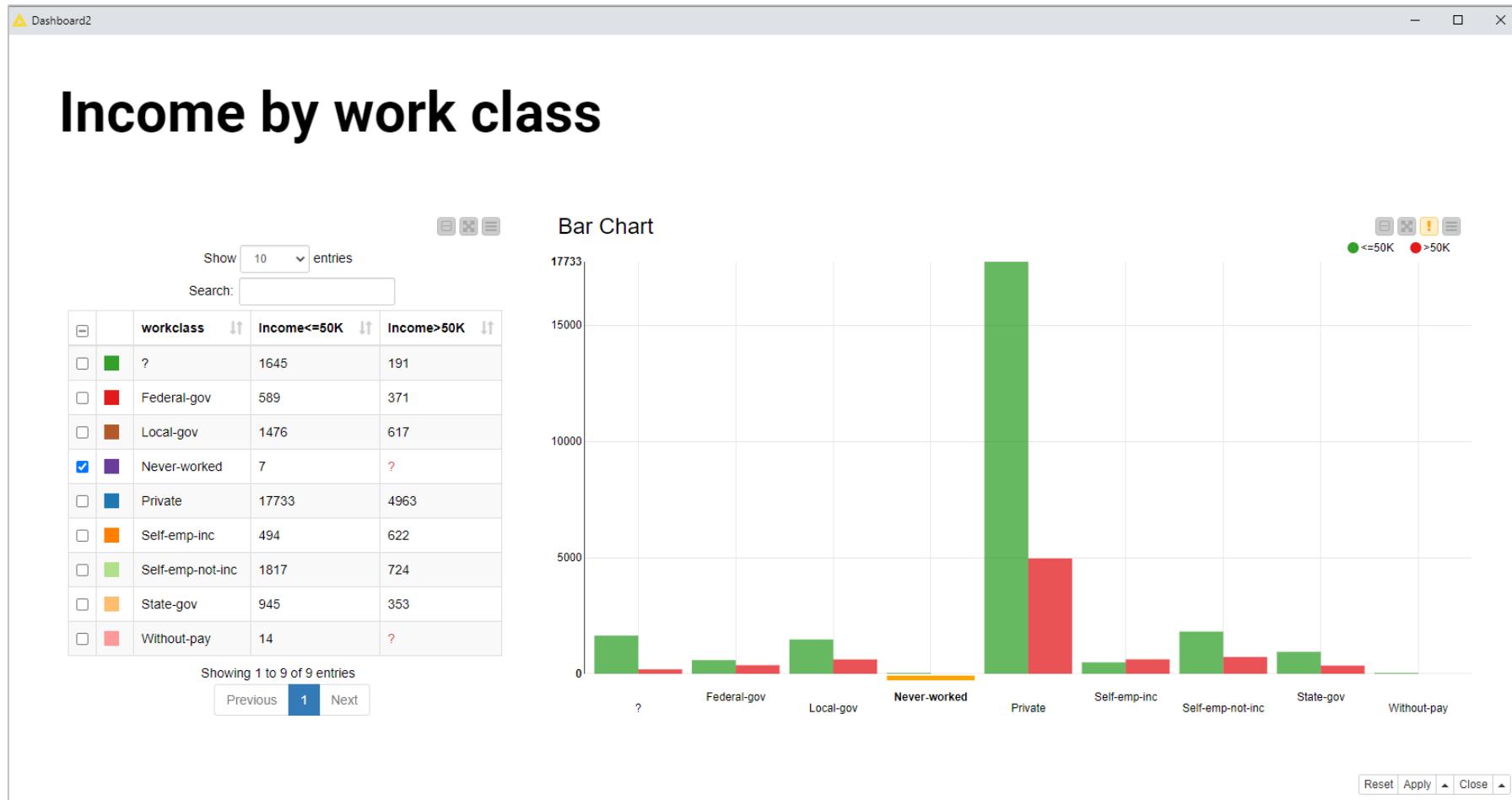
Se construye un Componente "Dashboard2" con:

- Un nodo Table View
- Un nodo A Bar Chart
- Un nodo A Text Output Widget

6.13. Componente "Dashboard2" para aplicar la vista compuesta requerida



6.14. La vista compuesta final con la clase "Never-worked" seleccionada tanto en la tabla como en el gráfico de barras

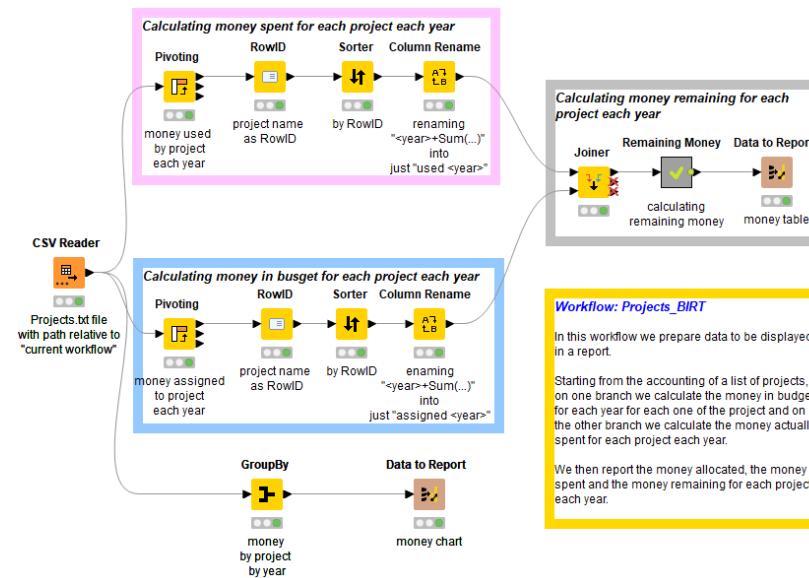


# Capítulo 7. Elaboración de informes con BIRT

## 7.1. Informes con BIRT

KNIME analytics Platform también ofrece una integración con la versión de código abierto de la herramienta de reporte [BIRT \(Business Intelligence and Reporting Tool\)](#). La integración de BIRT está contenida en una extensión de KNIME llamada "Report Designer". Esta extensión instala la interfaz entre KNIME Analytics Platform y BIRT, así como la versión de código abierto de BIRT. Por lo tanto, no es necesario tener una versión preinstalada de BIRT funcionando en su máquina para utilizar la extensión, como es el caso de las otras herramientas de BI. No es necesario comprar una licencia, lo que hace que la integración entre BIRT y KNIME Analytics Platform sea mucho más fácil de manejar. Por todo ello, en este capítulo nos centraremos en cómo construir un informe utilizando la Extensión del Diseñador de Informes (integración con BIRT). En este capítulo utilizaremos la integración BIRT para construir un informe similar al cuadro de mando construido en el capítulo anterior.

7.1. El flujo de trabajo "Reporting\_w\_BIRT"



En la carpeta del capítulo 7 del material descargado, puede encontrar una serie de flujos de trabajo -que se originan en el flujo de trabajo "Projects\_final" del capítulo anterior- para exportar los datos a diferentes herramientas de BI. En particular, los flujos de trabajo nº 1 y nº 2 también producen un cuadro de mando similar para BIRT y el KNIME Webportal, respectivamente.

Todos estos flujos de trabajo funcionan con el archivo "Projects.txt" disponible en la carpeta de datos del libro "KBLData". El archivo "Projects.txt" contiene una lista de nombres de proyectos con la correspondiente cantidad de dinero asignada y utilizada para cada trimestre de cada año entre 2007 y 2009. Todos los flujos de trabajo construyen una tabla dinámica con los nombres de los proyectos y la suma del dinero asignado, la suma del dinero utilizado y la suma del dinero restante (= asignado - utilizado) para cada proyecto y para cada año entre 2007 y 2009. Los datos resultantes producen algunas tablas y dos gráficos de barras en el informe asociado.

### 7.2. La tabla de datos, del nodo "money table", del flujo de trabajo "Reporting\_w\_BIRT"

Row ID	S name	I used 2007	I used 2008	I used 2009	I assigned 2007	I assigned 2008	I assigned 2009	D remain 2007	D remain 2008	D remain 2009
Row0_Row0	Blue	1300	1124	1650	1360	1277	1565	60	153	-85
Row1_Row1	Gobi	1220	1308	1740	1203	1424	1740	-17	116	0
Row2_Row2	Kalahari	876	768	1178	630	800	1192	-246	32	14
Row3_Row3	Kara Kum	800	992	1544	800	888	1516	0	-104	-28
Row4_Row4	La Guajira	1200	1648	1518	1020	1404	1496	-180	-244	-22
Row5_Row5	Mojave	2000	1820	1809	1800	1819	1860	-200	-1	51
Row6_Row6	Patagonia	1332	2139	1364	864	2098	1359	-468	-41	-5
Row7_Row7	Sahara	905	1460	1670	806	1457	1495	-99	-3	-175
Row8_Row8	Sechura	3600	3113	4000	3200	2966	3940	-400	-147	-60
Row9_Row9	Tanami	591	0	468	453	0	453	-138	0	-15
Row10_Row10	White	860	948	1347	860	1087	1420	0	139	73

## 7.2. Instalación de la extensión del diseñador de informes (BIRT)

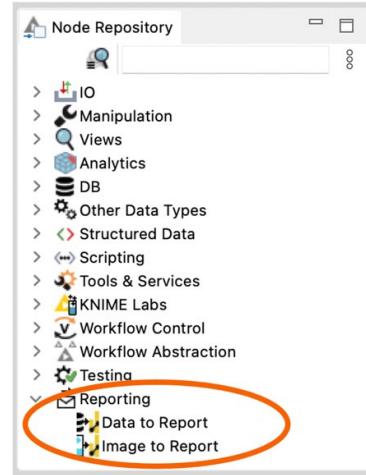
El paquete "KNIME Report Designer" no está incluido en la versión básica independiente de KNIME Analytics Platform. Puede descargarse como un paquete de extensión independiente desde el enlace "KNIME & Extensions" en "File" → "Install KNIME Extensions".

Para instalar la extensión "KNIME Report Designer":

- Inicie la plataforma de análisis KNIME
- En el menú superior haga clic en "Archivo" "Instalar extensiones KNIME..."
- En la ventana "Software disponible", expanda "KNIME & Extensiones" y baje hasta "KNIME Report Designer". Alternativamente, busque "KNIME Report Designer" en el cuadro de búsqueda de la parte superior.
- Seleccione la extensión "KNIME Report Designer".
- Haga clic en el botón "Siguiente" de la parte inferior y siga las instrucciones de instalación

Si la instalación se ejecuta correctamente, después de reiniciar KNIME Analytics Platform, debería tener una nueva categoría "Reporting" en el panel "Node Repository" con dos nodos: "Data To Report"" e "Image To Report"".

### 7.3. La categoría de informes ("Reporting") en el repositorio de nodos



## 7.3. Marcado de datos en el flujo de trabajo

La herramienta de informes KNIME es una aplicación diferente (BIRT) de KNIME Analytics Platform. La idea es que el flujo de trabajo KNIME prepara los datos para el diseñador de informes KNIME, mientras que el diseñador de informes KNIME muestra estos datos en un diseño gráfico.

Las dos aplicaciones, el editor de flujo de trabajo y la herramienta de informes, necesitan comunicarse entre sí; en particular, el flujo de trabajo necesita pasar los datos a la herramienta de informes. Esta comunicación de datos entre el flujo de trabajo y la herramienta de informes se produce a través del nodo "Data to Report".

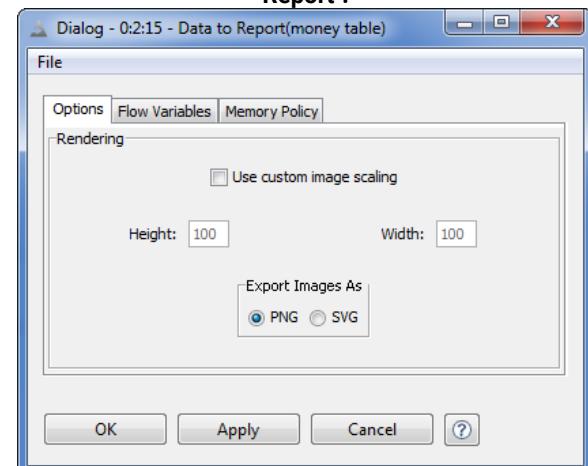
## Nodo "Data to Report"

El nodo "Data to Report" se encuentra en el panel "Node Repository" en la categoría "Reporting". El nodo "Data to Report" marca la tabla de datos KNIME en el puerto de entrada como un conjunto de datos para el KNIME Report Designer.

Al pasar del editor de flujos de trabajo a la herramienta de informes, todas las tablas de datos marcadas por un nodo "Data to Report" se importan automáticamente como conjuntos de datos en la herramienta de informes. Cada conjunto de datos lleva el nombre como el texto debajo del nodo "Data to Report" de origen. Por lo tanto, el texto debajo del nodo "Data to Report" es importante. Tiene que ser un texto significativo para facilitar la identificación del conjunto de datos en el entorno del informe.

Dado que el nodo "Data to Report" es sólo un marcador para una tabla de datos, no necesita mucha configuración. La ventana de configuración sólo contiene una bandera "use custom image scaling" para escalar las imágenes de los datos a un tamaño personalizado. El tamaño de la imagen por defecto es el tamaño del renderizador.

7.4. Ventana de configuración del nodo "Data To Report".



En nuestro flujo de trabajo utilizamos dos nodos "Data to Report". Uno está conectado a la secuencia de nodos "Math Formula" - en el metanodo llamado "Remaining Money" - y exporta los datos para las tablas del informe. El segundo está conectado al nodo "GroupBy" con el texto "Remaining Money" y exporta los datos para las tablas del informe. Añadimos un texto "money table" bajo el primer nodo de Data to Report y un texto "money chart" bajo el segundo nodo de Data to Report. Así, al pasar a la herramienta de informes, encontraremos allí dos conjuntos de datos llamados "money table" y "money chart" respectivamente. Sabremos inmediatamente qué datos utilizar para las tablas y qué datos para los gráficos.

En la categoría "money table" and "money chart" también se encuentra el nodo "Image to Report". El nodo "Image to Report" funciona de forma similar al nodo "Data to Report", sólo se aplica específicamente a las imágenes.

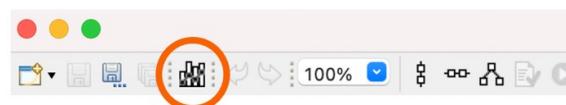
## 7.4. De KNIME a BIRT y viceversa

En KNIME Analytics Platform desarrollamos flujos de trabajo para la manipulación y el modelado de datos. En BIRT creamos y damos forma al informe para representar los datos de los flujos de trabajo. Sólo se asocia un informe a un flujo de trabajo y viceversa. No es posible asociar más de un informe a un flujo de trabajo. Cuando nos movemos en el entorno BIRT, abrimos el informe asociado al flujo de trabajo activo. Desde un flujo de trabajo KNIME abierto en el banco de trabajo KNIME, se puede pasar al entorno BIRT y abrir el informe asociado mediante:

- Abrir el flujo de trabajo desde el panel "KNIME Explorer" al editor de flujos de trabajo
- Haciendo clic en el ícono "Reporting" de la barra de herramientas.

El editor de reports de BIRT abre entonces el informe asociado al flujo de trabajo seleccionado. El editor de reportes crea una nueva pestaña en la ventana del editor de flujos de trabajo KNIME.

**7.5. El ícono del Reporte en la Herramienta**



**7.6. La nueva pestaña del editor de flujo de trabajo KNIME para el reporte seleccionado**



Para volver del reporte al editor de flujo de trabajo, puede seleccionar la pestaña de flujo de trabajo o hacer clic en el ícono de KNIME en la barra de herramientas. Esto le llevará de vuelta al entorno más familiar de KNIME. Si es la primera vez que abre el reporte asociado al flujo de trabajo, éste estará vacío

**7.7. El ícono de KNIME en la barra de herramientas cuando se abre un reporte**

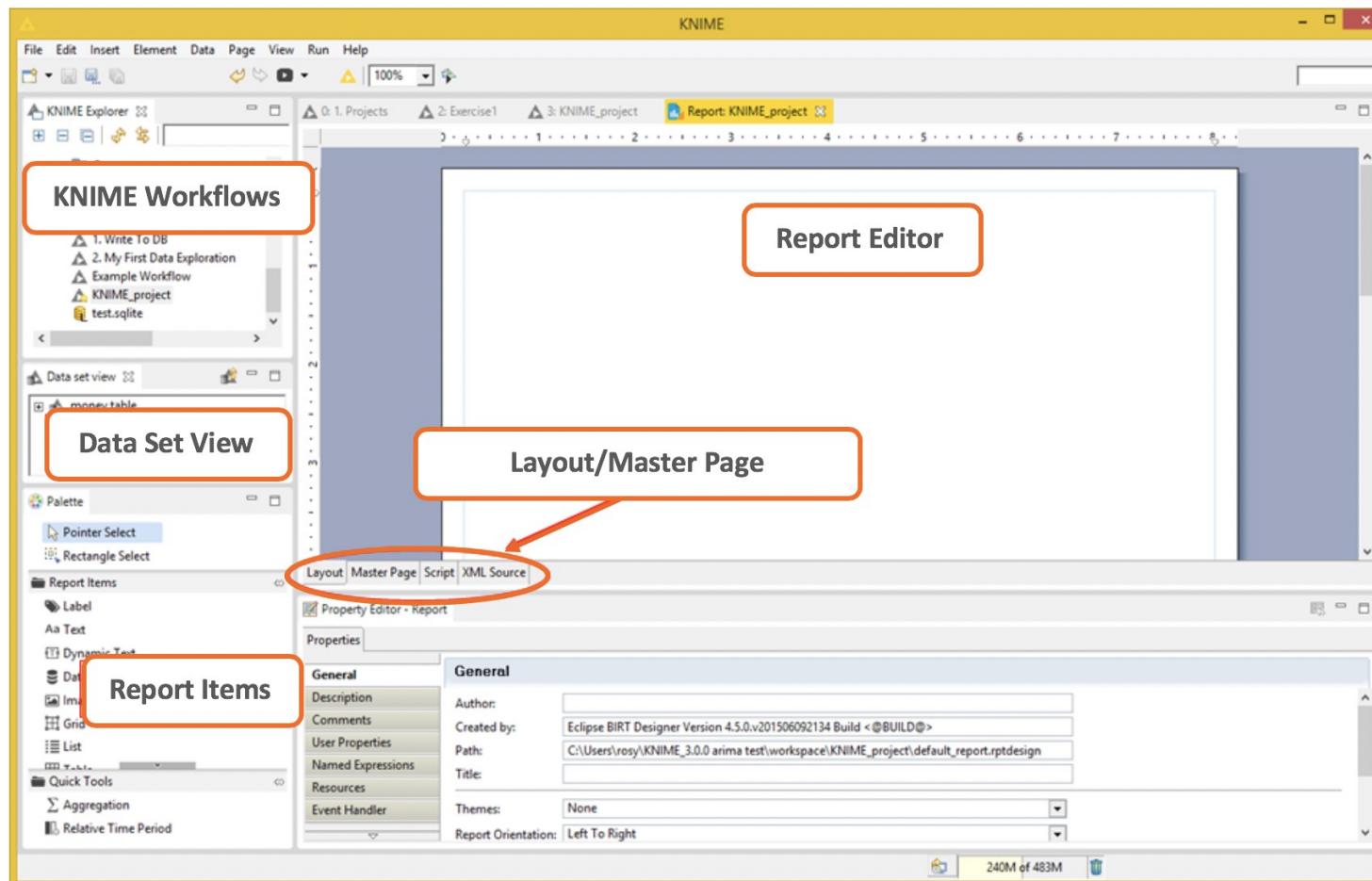
**Nota.** Si el flujo de trabajo no tiene un nodo de datos para reportar, el ícono de "Report" no está presente en la barra de herramientas.

Haga doble clic en el flujo de trabajo "Reporting\_w\_BIRT" en el panel "KNIME Explorer" para abrirlo; a continuación, seleccione el ícono del informe en la barra de herramientas. Esto le lleva al entorno BIRT y al reporte asociado.

## 7.5. El entorno BIRT

BIRT está desarrollado como un plug-in de Eclipse, al igual que KNIME Analytics Platform. Esto significa que ambos heredan algunas propiedades y herramientas de la plataforma Eclipse. Como consecuencia, el editor de reportes de BIRT y el editor de flujos de trabajo de KNIME son muy similares, lo que nos facilita el proceso de aprendizaje de la herramienta de reportes. En esta sección ofrecemos una rápida visión general del editor de informes BIRT. Para más información sobre el software BIRT, el libro que aparece en [2] ofrece una visión detallada de las potencialidades de BIRT. Veamos las diferentes ventanas del entorno BIRT con un informe vacío.

7.8. El editor Report Editor en el entorno BIRT



El panel "**KNIME Explorer**" sigue estando en la esquina superior izquierda y sigue conteniendo la lista de flujos de trabajo KNIME disponibles.

Bajo el panel "**KNIME Explorer**", encontramos el panel "**Data Set View**". Este panel contiene todos los conjuntos de datos que están disponibles para el reporte.

Bajo el panel "**Data Set View**", encontramos la lista de todos los "**Report Items**" disponibles para crear nuestro reporte, como Tabla, Etiqueta, Gráfico, etc.

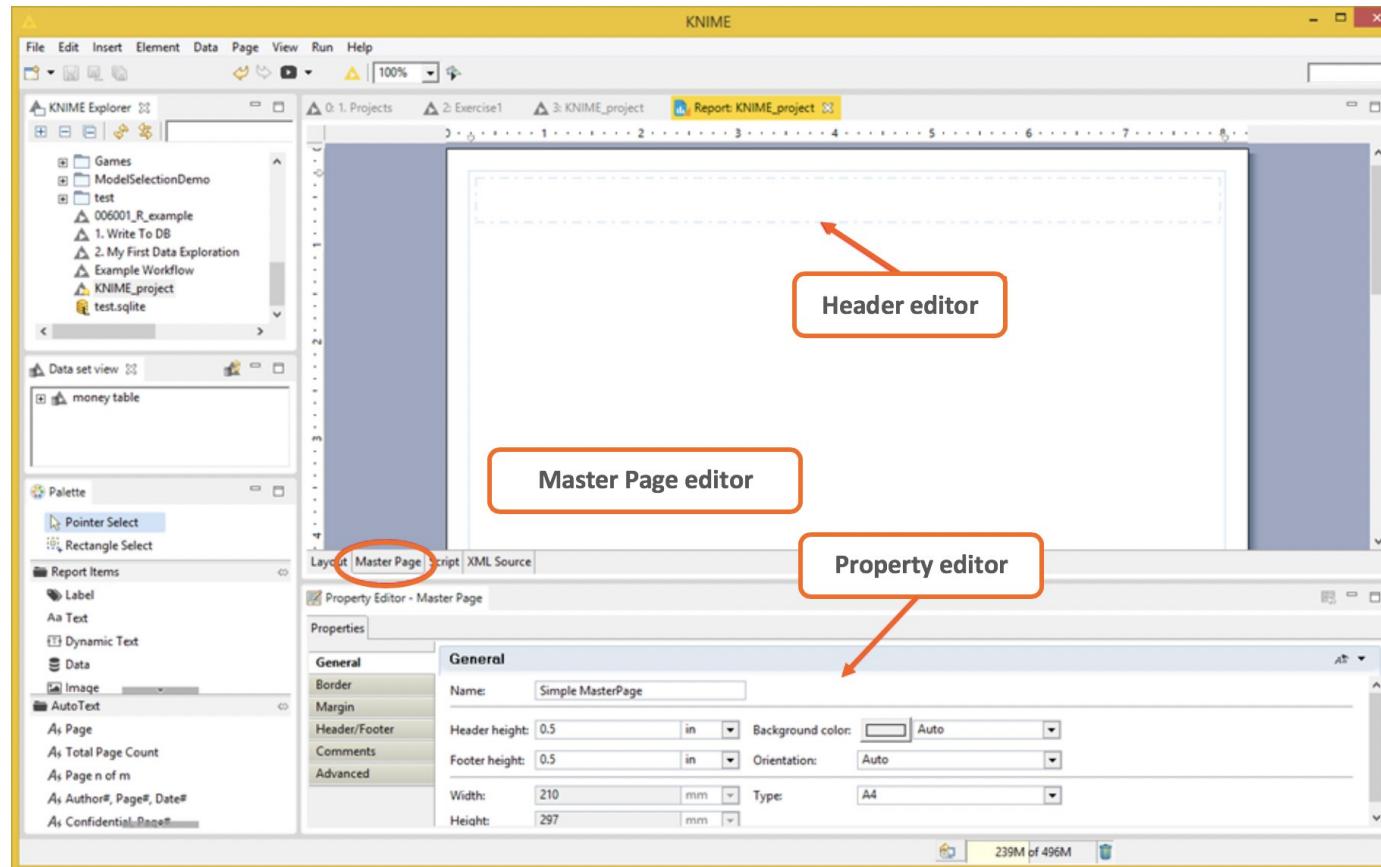
En el centro, al igual que en el editor de flujo de trabajo de KNIME, encontramos el editor de reportes. Al igual que en KNIME, donde construimos los flujos de trabajo "arrastrando y soltando" los nodos en el editor de flujos de trabajo, aquí podemos componer el informe "arrastrando y soltando" los elementos del informe en el editor de informes.

Finalmente, en la parte inferior central de la ventana hay unas cuantas pestañas, de las cuales sólo dos son interesantes para nuestro trabajo: Layout and Master Page

**Layout** es el editor de páginas, donde se procesa la página de reporte individual.

**Master Page**, como en PowerPoint Master Page, define una plantilla para cada página del reporte. Aquí se diseñan el encabezado y el pie de página.

## 7.9. El Header Editor dentro del editor de la Master Page Editor



### Master Page (Pagina Maestra)

Justo debajo del editor de reportes, hay algunas pestañas: "Layout", "Master Page", y otras. Seleccionemos la pestaña "Master Page".

Ahora el editor de reportes en el centro se ha convertido en el editor de la Página Maestra y, debajo de las pestañas, puede ver el Editor de Propiedades de la Página Maestra. Hay 6 grupos de propiedades: "General", "Border", "Margin", "Header/Footer", "Comments", y "Advanced".

Queremos preparar un reporte para exportarlo en diapositivas en formato PowerPoint. También queremos tener un título corrido con un logo en todas las diapositivas.

Generalmente las diapositivas de PowerPoint tienen una orientación horizontal. Para cambiar la orientación del papel, vamos al campo "Orientation" en la propiedad "General". Lo cambiamos a "Landscape".

Para crear un título, debemos cambiar la cabecera en la Página Maestra. La propiedad "Encabezado/Pie de página" sólo ofrece casillas de verificación para mostrar o no el encabezado y el pie de página. Para poder cambiar realmente la cabecera y el pie de página, debemos trabajar en el propio editor de la Página Maestra. En la parte superior del editor de la página maestra hay un rectángulo punteado. Este es el editor de la cabecera.

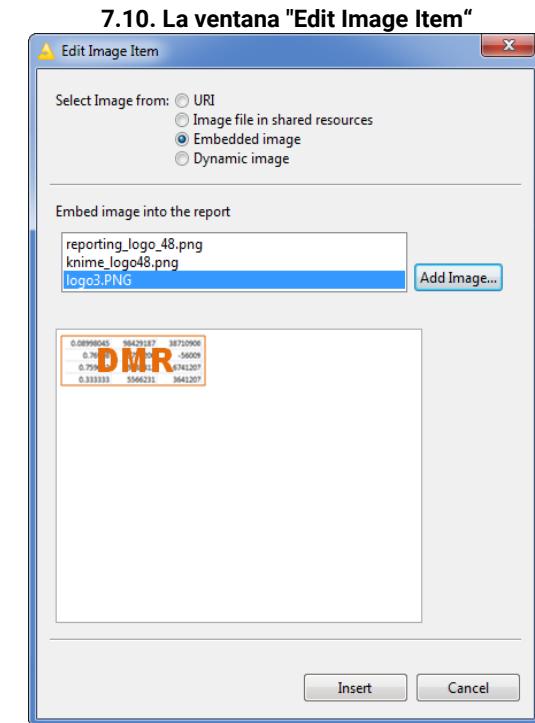
Para insertar un logo en el encabezado de cada diapositiva se vá al editor de la Master Page:

- Haga clic con el botón derecho del ratón en el editor de cabeceras
- Seleccione "Insertar".
- Seleccione "Imagen".
- En la ventana "Editar elemento de imagen" cargue su imagen, por ejemplo, como un archivo incrustado

La imagen del logotipo aparecerá en la esquina superior izquierda del editor de cabecera.

En lugar de una imagen, puede insertar una "Label" en el editor de cabecera para tener un título corrido en sus diapositivas. También puede combinar ambos, un título y un logotipo, en el editor de cabecera. Sin embargo, sólo puede combinar más elementos de reporte uno al lado del otro utilizando el elemento de reporte "Grid".

Para ver cómo será el reporte, debe seleccionar "Run" → "View Report" en el menú superior y luego su formato de salida. Esto genera el reporte real. Para una vista previa rápida puede elegir "En el visor web" para una creación rápida de la página de reporte HTML. Por el momento es sólo el logotipo que hemos introducido en la esquina superior izquierda y el pie de página con el logotipo de KNIME.



## The Data Sets

El panel denominado "Data set view" contiene los datos disponibles para el reporte. Cada reporte está vinculado a un solo flujo de trabajo. En la integración de BIRT dentro de KNIME, los conjuntos de datos se importan automáticamente desde las tablas de datos marcadas por un nodo "Data to Report" en el flujo de trabajo subyacente. En la versión integrada, no hay otra forma de generar conjuntos de datos en el entorno de los informes.

Veamos los conjuntos de datos disponibles para el reporte del flujo de trabajo "Reporting\_w\_BIRT".

En el panel "Data Set View" debería ver dos conjuntos de datos, llamados "money chart" y "money table". Estos eran los nombres de los dos nodos "Data Set View" en el flujo de trabajo "Reporting\_w\_BIRT". En efecto, al pasar del editor de flujos de trabajo KNIME al editor de informes BIRT, los datos de los nodos "Data to Report" se exportan automáticamente como conjuntos de datos al entorno del informe.

Si ha utilizado nombres oscuros de conjuntos de datos y no puede recordar de qué nodo "Data to Report" se ha generado el conjunto de datos o para comprobar que el conjunto de datos se ha exportado correctamente, puede que tenga que previsualizar los datos del conjunto de datos. Para ello:

- Haga doble clic en el conjunto de datos O haga clic con el botón derecho del ratón en el conjunto de datos y seleccione "Edit"
- En "Edit Data Set" seleccione "Preview Results"

7.11. "Preview Results" muestra el contenido del conjunto de datos

Row ID	name	reference year	Sum(money assign...)	Sum(money u...)
Row0	Blue	2007	1360.0	1300.0
Row1	Blue	2008	1277.0	1124.0
Row2	Blue	2009	1565.0	1650.0
Row3	Gobi	2007	1203.0	1220.0
Row4	Gobi	2008	1424.0	1308.0
Row5	Gobi	2009	1740.0	1740.0
Row6	Kalahari	2007	630.0	876.0
Row7	Kalahari	2008	800.0	768.0
Row8	Kalahari	2009	1192.0	1178.0
Row9	Kara Kum	2007	800.0	800.0
Row10	Kara Kum	2008	888.0	992.0
Row11	Kara Kum	2009	1516.0	1544.0
Row12	La Guajira	2007	1020.0	1200.0
Row13	La Guajira	2008	1404.0	1648.0
Row14	La Guajira	2009	1496.0	1518.0
Row15	Mojave	2007	1800.0	2000.0
Row16	Mojave	2008	1819.0	1820.0
Row17	Mojave	2009	1860.0	1809.0
Row18	Patagonia	2007	864.0	1332.0

Total 33 record(s) shown.

OK Cancel

## 7.6. La Plantilla (The Layout)

Empecemos a crear el reporte. Haga clic en la pestaña "Layout" para salir del editor de la Master Page y volver al editor del reporte. Lo que vemos ahora es una página vacía. En primer lugar, nos gustaría tener un título para nuestro reporte, algo así como "Project Report: Money Flow", por ejemplo. Flujo de dinero", por ejemplo. Vamos a colocar tablas, gráficos y etiquetas más explicativas bajo el título principal.

### El Título

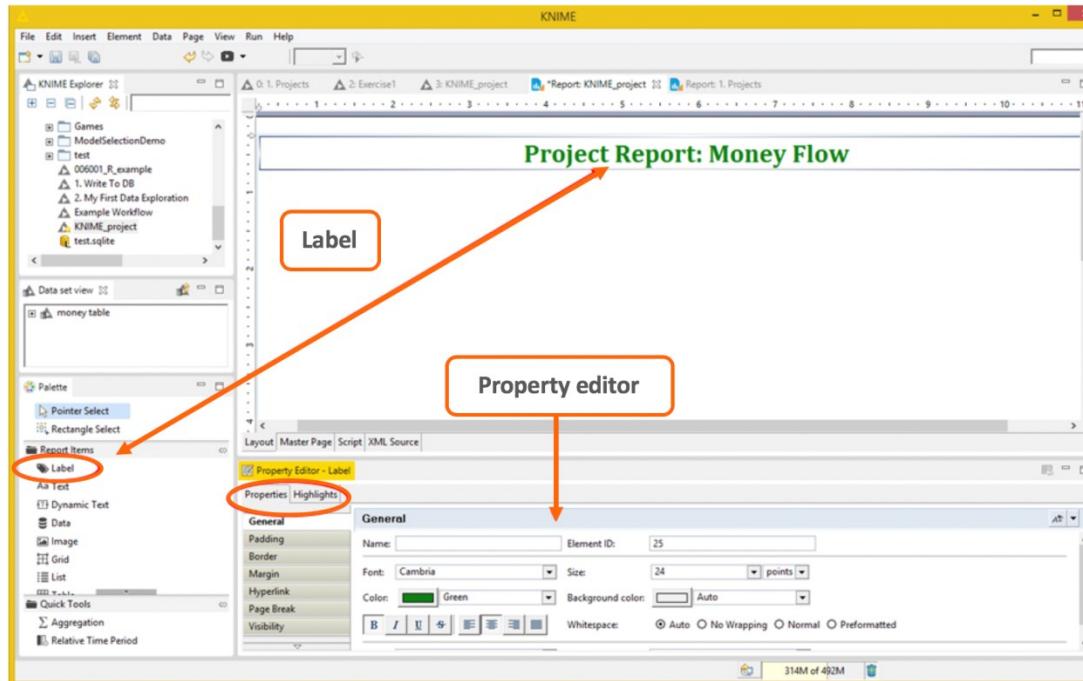
Para crear un título:

- Arrastre y suelte el elemento de informe "Label" desde el panel "Report Items" en la esquina inferior izquierda al editor de informes
- Haga doble clic en la etiqueta e introduzca el título: "Project Report: Money Flow"
- Seleccione toda la etiqueta haciendo clic en su contorno externo
- En el editor de "Property" bajo el editor de Informes, vaya a la pestaña llamada "General" y seleccione las propiedades para su título: fuente, tamaño de fuente, estilo de fuente, color de fuente, color de fondo, etc.

Hemos seleccionado "Cambria", color "green", tamaño "24 points", estilo "bold", y ajustes "centered".

**Nota.** La configuración del tamaño de la fuente consta de 2 parámetros: el número y la unidad de medida (% , cm, in, puntos, etc...). Asegúrese de ajustar ambos parámetros de forma coherente. Si establece el número en 24 y la unidad en "%", ya no verá la etiqueta del título y se preguntará qué ha pasado con ella.

#### 7.12. Arrastre y suelte un elemento "Label" en el Editor de Informes para crear el título del reporte



Seguro que se ha dado cuenta de que la etiqueta del título se ha colocado automáticamente en la parte superior de la página y que ocupa todo el ancho de la misma. No puede moverla para colocarla en otro lugar ni encogerla para que ocupe sólo una parte del ancho de la página. Este ajuste automático (todo el ancho de la página y el primer lugar disponible en la página desde la parte superior) afectará a todos los elementos del reporte que se arrastren desde el panel "Report Items" y se suelten directamente en el editor de reportes. Para el elemento del título esto no es tan malo, ya que el título suele abarcar todo el ancho de la página y se coloca en la parte superior de la misma. Sin embargo, no es deseable para la mayoría de los demás elementos del reporte.

## **La Matrix (The Grid)**

En nuestro reporte nos gustaría tener tres tablas: dos tablas en la parte superior que describan la cantidad de dinero asignada y utilizada para cada proyecto cada año, y una tabla en el medio de la página debajo de las dos tablas anteriores para mostrar el dinero restante. También estaría bien que todas las tablas tuvieran el mismo tamaño, es decir, algo menos de la mitad del ancho de la página. Debajo de las tablas nos gustaría colocar dos gráficos de barras uno al lado del otro para mostrar respectivamente cómo se ha asignado y utilizado el dinero. Para tener la libertad de colocar los elementos del reporte en cualquier parte de la página del reporte y darles un tamaño arbitrario, necesitamos colocarlos dentro de una matriz "Grid".

La "Grid" es un elemento del reporte, algo así como una tabla que crea celdas en la página del reporte con ubicación y tamaño personalizables para contener otros elementos del reporte.

Para nuestro reporte, necesitamos:

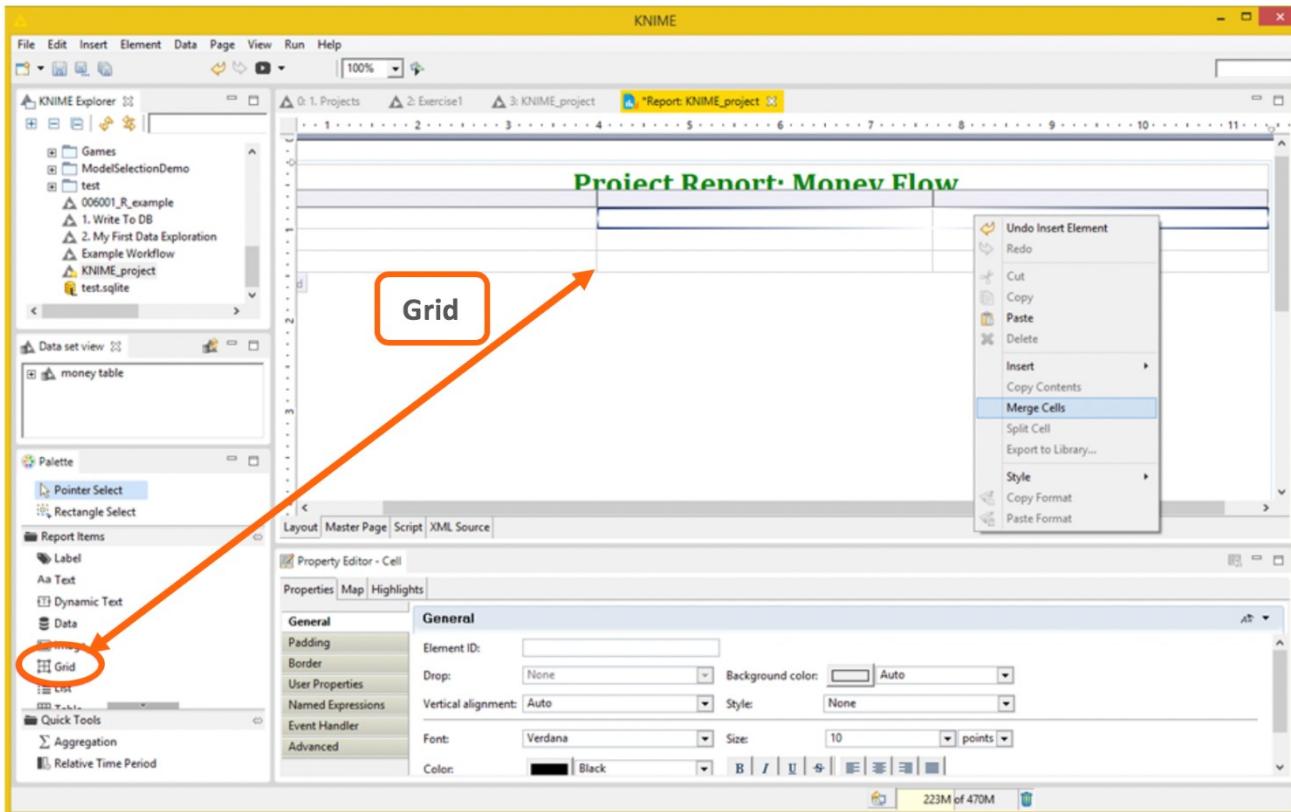
- una fila con dos celdas: una para la tabla de dinero asignada y otra para la tabla de dinero utilizada
- una fila con una sola celda para la tabla de dinero restante
- una fila con dos celdas de nuevo para los 2 gráficos de barras

Por lo tanto, queremos crear una "Grid" con 3 filas y 2 columnas y fusionar las dos celdas de la segunda fila en una sola celda.

Para crear la matriz ("Grid"):

- Arrastre y suelte el elemento de informe "Grid" del panel "Report Items List" al editor de informes bajo la etiqueta de título
- Introduzca 2 para el número de columnas y acepte 3 para el número de filas
- Seleccione las dos celdas de la segunda fila haciendo clic en el borde externo izquierdo de la fila
- Haga clic con el botón derecho en la selección de dos celdas
- Seleccione la opción "Merge cells".

7.13. Arrastre y suelte el elemento de informe "Grid" en el editor de informes, seleccione 3 filas y 2 columnas, y fusione las dos celdas de la fila central



**Nota.** A veces utilizamos cuadrículas demasiado detalladas. Esto significa que definimos cuadrículas con más columnas y filas de las necesarias. Esto nos da más libertad para ajustar las distancias entre los elementos del reporte y otros márgenes.

## 7.7. Las Tablas (The Tables)

Para crear una tabla podemos seguir el procedimiento estándar:

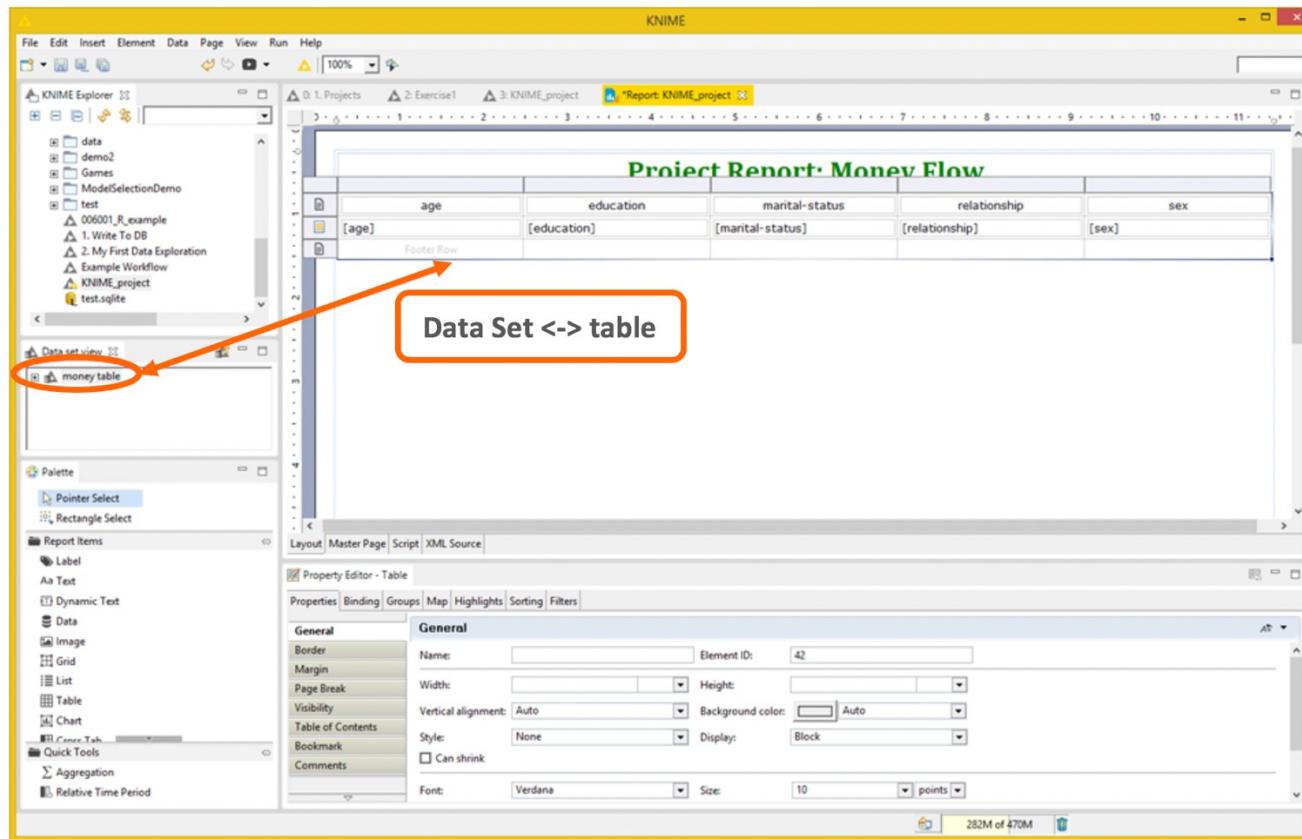
- Arrastrar y soltar el elemento de informe "Table" en el editor de informes
- Vincular la "Table" a un conjunto de datos
- Vincular cada celda de datos a un campo del conjunto de datos

O podemos:

- Arrastre y suelte el conjunto de datos en el editor de reportes
- En la siguiente ventana, seleccione las columnas de datos que desea que aparezcan en el reporte final

El segundo método es más fácil, especialmente para las tablas grandes.

#### 7.14. Arrastre y suelte un conjunto de datos desde el panel "Vista del conjunto de datos" para producir una tabla con tantas columnas como campos tenga el conjunto de datos



En el diseño del reporte una tabla se compone de tres filas

- una fila de cabecera
- una fila de celdas de datos

- una fila de pie de página

La fila de cabecera y la fila de pie de página contienen sólo etiquetas u otros elementos estáticos del reporte y aparecen en el reporte final sólo una vez al principio y al final de la tabla respectivamente. La fila de celdas de datos contiene los campos del conjunto de datos. En el informe real, la fila de celdas de datos se multiplica en tantas filas como haya en el conjunto de datos.

Después de arrastrar y soltar el conjunto de datos en el editor del reporte, vemos una tabla con tantas columnas como campos hay en el conjunto de datos. Las cabeceras de las columnas se configuran automáticamente como etiquetas con el nombre del campo del conjunto de datos. La fila del pie de página está vacía. La fila de la celda de datos contiene los campos del conjunto de datos. Ahora vamos a ajustar el aspecto de la tabla.

#### *Eliminar las columnas no deseadas*

- Seleccione toda la tabla. Si pasa el ratón por encima de la esquina inferior izquierda de la tabla, aparece un pequeño rectángulo gris con la palabra "Table". Para seleccionar toda la tabla, haz clic en ese rectángulo.
- Seleccione la columna no deseada. Para seleccionar toda una columna, haga clic en el rectángulo gris situado encima de la cabecera de la columna.
- Haga clic con el botón derecho del ratón en la parte superior de la columna no deseada
- Seleccione "Delete".

#### *Cambiar la cabecera de la columna*

- La cabecera de cada columna es una etiqueta editable
- Haga doble clic en la etiqueta de la cabecera
- Cambie el texto

#### *Cambiar la posición de las columnas*

- Seleccione toda la tabla
- Haga clic con el botón derecho del ratón en la parte superior de la columna (el rectángulo gris) que desea mover
- Seleccione "Cut".
- Seleccione la columna que desea posicionar a la izquierda; para ello, haga clic con el botón derecho del ratón en el rectángulo gris de la parte superior de la misma
- Seleccione "Insert Copied Column"

#### *Cambiar las propiedades de la fuente*

- En cuanto a las "Labels", en la ventana de "Properties" (pestaña "General") puede cambiar la fuente, el tamaño de la fuente, la alineación, el estilo, etc.

#### *Formatear el número*

- Seleccione una celda que contenga un número
- En el editor de "Properties", seleccione la pestaña "Format Number".
- Elija el formato del número en su celda

#### *Defina el alto y el ancho*

- Seleccione una fila o una columna
- En la ventana "Properties", vaya a la pestaña "General" y cambie el alto y el ancho

#### *Establezca los bordes*

- Seleccione el elemento que necesita bordes (tabla completa, fila o celda única).
- En el editor de "Properties", seleccione la pestaña "Border".
- Elija el borde deseado

**Note.** La propiedad "Border" no está disponible para las columnas.

#### *Establecer el tamaño de la mesa*

- Seleccione toda la tabla
- En la ventana "Properties", seleccione la pestaña "General".
- Elija el ancho y el alto deseados

**Note.** Para la fuente, la celda y el tamaño de la tabla, el alto y el ancho pueden expresarse en diferentes unidades de medida. Compruebe que la unidad que está utilizando es una unidad significativa. BIRT realiza algún tipo de ajuste automático en el ancho y alto de las celdas. Para que el alto y ancho de las celdas individuales sea efectivo, debe definir primero una altura y ancho adecuados para la tabla completa.

Arrastramos y soltamos el conjunto de datos de la "money table" en cada una de las dos celdas de la primera fila y en la única celda de la segunda fila de la "Grid". La tabla de la izquierda de la primera fila mostrará el dinero asignado. A continuación, eliminamos todas las columnas "\*used\*" y "\*remain\*". La tabla de la derecha de la primera fila mostrará el dinero utilizado. A continuación, borramos todas las columnas "\*assigned\*" y "\*remain\*\*". La tabla de la segunda fila mostrará los valores restantes. Aquí borramos todas las columnas "\*used\*" y "\*assigned\*".

En cada tabla, la columna "RowID" contiene el nombre del proyecto. Por lo tanto, cambiamos la etiqueta de la cabecera por "Nombre". Los datos y la celda de cabecera de la columna "Name" estaban alineados a la izquierda, mientras que las 3 últimas celdas estaban alineadas a la derecha. Las tablas tenían un borde verde a su alrededor y también un borde verde entre la fila del encabezado y la de los datos.

El tamaño de las dos primeras tablas se fijó en el 80% (= 80% de la celda de la cuadrícula) y el tamaño de la tercera tabla, que en una celda de la cuadrícula es el doble del tamaño de las dos anteriores, se fijó en el 40% (= 40% de la celda de la cuadrícula). La propiedad de alineación de las tres celdas de la cuadrícula se estableció en "Center".

En la primera tabla, se fijó la fuente en "Cambria" y el tamaño de la fuente en "10 puntos" tanto en las celdas de encabezado como en las de datos. El estilo de la fuente del encabezado también se fijó en "bold" y el color en "green". Por último, las celdas de datos que contenían números se formatearon con "Format Number" establecido en "Fixed" con 2 decimales y separador de 1000. Todas estas operaciones deben repetirse también para la segunda y la tercera tabla

## Toggle Breadcrumb

En la barra superior puede encontrar el botón "Toggle Breadcrumb".

This button shows the hierarchy of a report element on the layout, for example the hierarchy of the "assigned 2008" data cell as:

Grid → Row → Cell → Table → Row → Cell → <data set field name>

**7.15. Toggle Breadcrumb**

The screenshot illustrates the 'Toggle Breadcrumb' feature in the Crystal Report Designer. The toolbar at the top includes a magnifying glass icon, which is circled in red. The breadcrumb navigation path shows the report structure: 0: 1. Projects > 2: Exercise1 > Report: 1. Projects. The 'Report' item in the path is also circled in red. The main workspace displays a report titled 'Project Repo' with a table titled 'Assigned money'. The table has columns for 'Name', '2009', '2008', and '2007'. The '2008' column contains a data cell with the value '[assigned 2008]', which is also circled in red. The left side of the interface shows the report structure tree, indicating the hierarchy from Grid to Cell.

### Previsualización del Reporte

Ahora vamos a crear el reporte (desde el menú superior "Run" → "View Report" → "In Web Viewer") para tener una idea aproximada del aspecto que tendrá el reporte. Probablemente el tamaño de fuente "large" que hemos elegido para las celdas de datos y las celdas de cabecera será demasiado grande para que las tablas quepan bien en una página. Podemos reducir fácilmente el tamaño de la fuente ajustándolo a "small" en una o ambas hojas de estilo (Style Sheets). Esto se aplicará automáticamente a todas las celdas de la tabla que hayan sido formateadas por estas Hojas de Estilo (Style Sheets). Esta es una de las grandes ventajas de usar Hojas de Estilo (Style Sheets).

Coloquemos una etiqueta en la parte superior de cada tabla para decir qué representa la tabla: "assigned money", "used money", y "remaining money". Luego podemos cambiar los encabezados de las columnas de "<assigned/used/remain> <year>" a "<year>", por ejemplo "assigned 2009" at "2009" y así sucesivamente. También vamos a añadir unas cuantas etiquetas vacías después de cada tabla para que el diseño del reporte sea más espacioso. Si ahora ejecutamos una vista previa, el reporte se verá similar al que se muestra a continuación.

#### **7.16. Vista del reporte en un navegador web después de crear y formatear las tres tablas**

0.08998045	58429187	387120908	
0.79	100	56009	
0.799	111	6743207	
0.333333	5566231	3641207	
<b>DMR</b>			
<b>Project Report: Money Flow</b>			
<b>Assigned money</b>			
Name	2009	2008	2007
Blue	1,565	1,277	1,360
Gobi	1,740	1,424	1,203
Kalahari	1,192	800	630
Kara Kum	1,516	888	800
La Guajira	1,496	1,404	1,020
Mojave	1,860	1,819	1,800
Patagonia	1,359	2,098	864
Sahara	1,495	1,457	806
Sechura	3,940	2,966	3,200
Tanami	453	0	453
White	1,420	1,087	860
<b>Used money</b>			
Name	2009	2008	2007
Blue	1,650	1,124	1,300
Gobi	1,740	1,308	1,220
Kalahari	1,178	768	876
Kara Kum	1,544	992	800
La Guajira	1,518	1,648	1,200
Mojave	1,809	1,820	2,000
Patagonia	1,364	2,139	1,332
Sahara	1,670	1,460	905
Sechura	4,000	3,113	3,600
Tanami	468	0	591
White	1,347	948	860
<b>Remaining money</b>			
Name	2009	2008	2007
Blue	-85	153	60
Gobi	0	116	-17
Kalahari	positive	32	-246
Kara Kum	-28	-104	0
La Guajira	-22	-244	-180
Mojave	positive	-1	-200
Patagonia	-5	-41	-468
Sahara	-175	-3	-99
Sechura	-60	-147	-400
Tanami	-15	0	-138
White	positive	139	0

## Mapeo

A veces, podemos querer asignar valores numéricos a valores descriptivos. Por ejemplo, en un reporte financiero, podemos asignar una columna con valores numéricos como:

Valores < 0 a "negative"

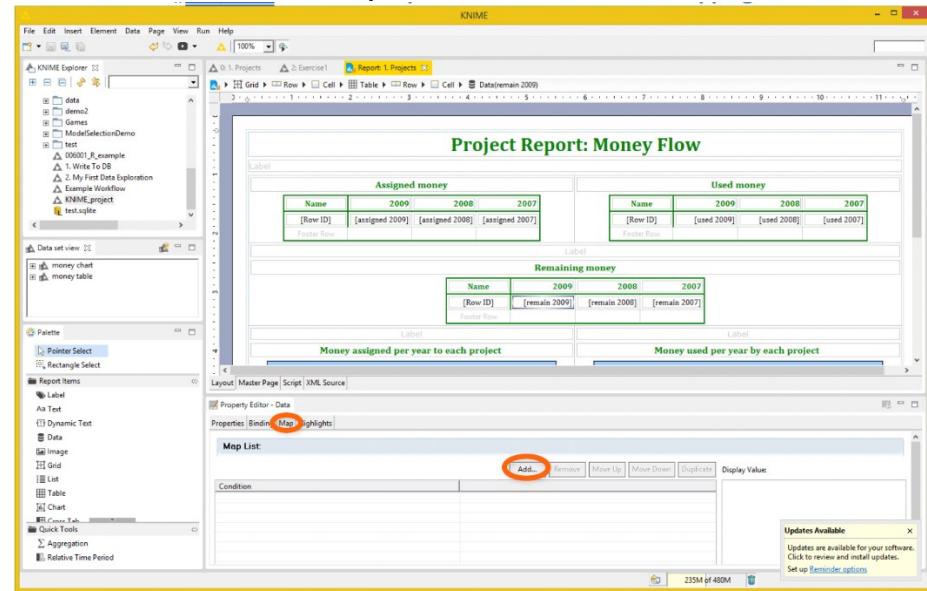
Valores = 0 a "zero"

Valores > 0 a "positive"

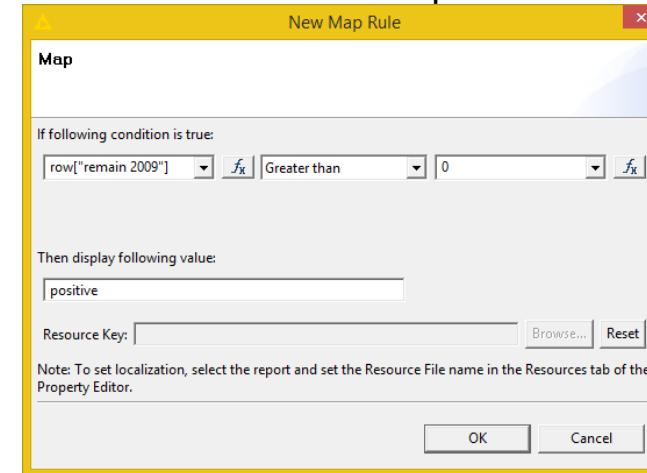
La funcionalidad de mapeo se encuentra en la pestaña "Maps" del editor de "Properties" de los elementos del reporte de la tabla; es decir, celdas, filas, columnas e incluso toda la tabla.

- Seleccione la celda de datos, la fila, la columna o la tabla a la que desea aplicar su mapeo
- Seleccione la pestaña "Maps" en el editor de "Properties".
- Haga clic en el botón "Añadir" para añadir una nueva regla de asignación
- Se abre el editor "New Map Rule".
- Construya su condición en el "Map Rule Editor", por ejemplo:
  - $\text{row}["remain 2009"] \text{ Greater than } 0 \rightarrow \text{"positive"}$
- Clickea "OK"

7.17. La pestaña "Mapas" del editor de propiedades de la tabla define la asignación de texto para un elemento de la tabla



7.18. El editor "New Map Rule"



## Highlights (Destacados-Resaltados)

La propiedad "Highlights" funciona de forma similar a la propiedad "Maps", sólo que afecta al diseño de las celdas y filas en lugar del contenido del texto de las celdas.

La propiedad "Highlights" se encuentra en la pestaña "Highlights" del editor de propiedades de los elementos del reporte "Table": celdas, filas, columnas y toda la tabla.

Por ejemplo, queremos marcar en rojo todas las celdas con un valor de "remain 2009" inferior a 0.

- Seleccione la celda de datos [remain 2009] (u otra celda, una fila o una columna en la que deba producirse el resaltado)
- Haga clic en la pestaña "Highlights" del editor de propiedades
- Haga clic en el botón "add".

Se abre el editor "New Highlight".

En la sección "Condition" :

- Introduzca la regla para el resaltado, por ejemplo:

*Row[remain 2009] smaller than 0*

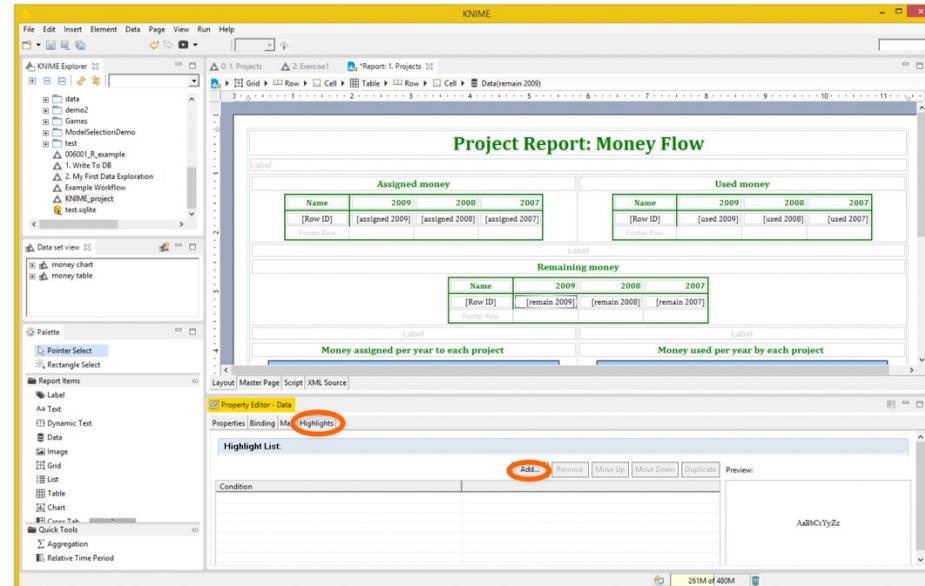
Para construir la regla también puede utilizar el "Expression Editor" que se explica más adelante en este capítulo.

- En la sección "Format", introduzca el formato que desea que se aplique, cuando la condición sea verdadera.

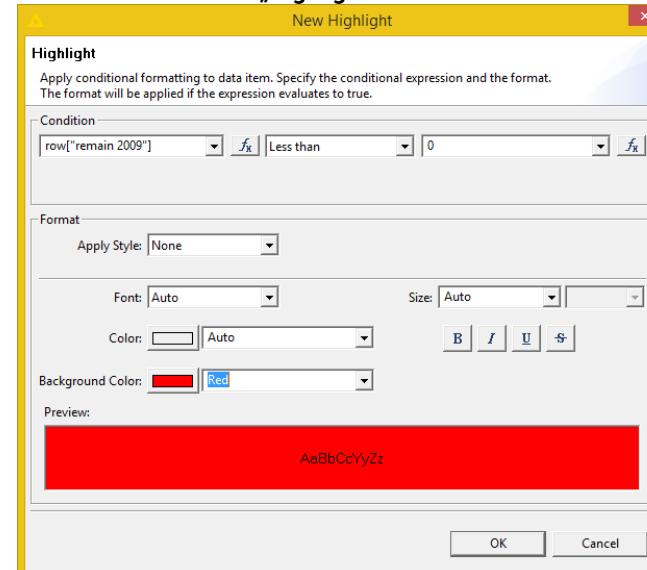
Para ello, haga clic en el botón situado junto a "Background Color" y seleccione el color rojo en el diálogo de colores.

- Haga clic en "OK".

7.19. La pestaña "Highlights" del Editor de Propiedades define las propiedades condicionales de un elemento de la tabla



7.20. El „Highlights Rule Editor“



Después de cerrar el cuadro de diálogo Resaltar, ejecute una vista del documento para ver las nuevas celdas resaltadas en rojo.

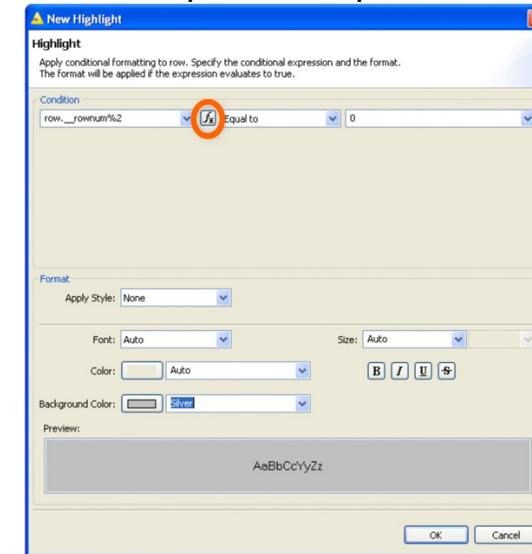
### El estilo zebra (Zebra style)

El estilo zebra es muy popular para las tablas de los reportes. En él, las filas de la tabla tienen colores alternos. Para producir una tabla de estilo zebra, necesita añadir la siguiente condición en el editor "New Highlight":

- Seleccione toda la fila de datos de la tabla, seleccionando el rectángulo gris a la izquierda de la fila de la tabla
- Seleccione la pestaña "Highlights" en el editor de propiedades
- Seleccione el icono "Expression Builder". Este es el icono con "fx" cerca de la caja de entrada "Condition"
- En el cuadro de diálogo "Expression Builder", seleccione "Available Column Bindings" y luego "Table"
- Haga doble clic en "RowNum" en la columna derecha de la tabla "Expression Builder"
- row.\_rownum aparece en el "Editor del Constructor de Expresiones"
- Escriba "row.\_rownum % 2" en el cuadro de diálogo "Expression Builder" y haga clic en "OK"
- Seleccione "Equal to" e introduzca "0" en el editor de "New Highlights"
- En la sección "Format", seleccione el color de fondo "gray" o "silver" en el campo consecuente de la regla
- Haga clic en "OK".

Ejecute una vista previa del documento para ver la tabla de estilo zebra.

7.21. El ícono para abrir el "Expression Builder"



7.22. La tabla "zebra style"

Net money			
Name	2009	2008	2007
Blue	-85.00	153.00	60.00
Gobi	0.00	116.00	-17.00
Kalahari	14.00	32.00	-246.00
Kara Kum	-28.00	-104.00	0.00
La Guajira	-22.00	-244.00	-180.00
Mojave	51.00	-1.00	-200.00
Patagonia	-5.00	-41.00	-468.00
Sahara	-175.00	-3.00	-99.00
Sechura	-60.00	-147.00	-400.00
Tanami	-15.00	0.00	-138.00
White	73.00	139.00	0.00

## 7.8. Salto de página

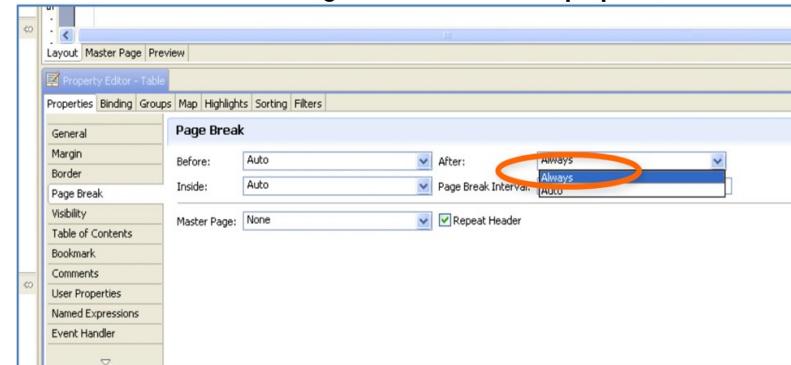
Queremos exportar el reporte final a PowerPoint. Esta primera parte de nuestro reporte encaja muy bien en una diapositiva de PowerPoint. Un salto de página en este punto sería muy útil para evitar efectos de formato de página no deseados en el documento final.

Para insertar un salto de página después de un elemento del reporte:

- Seleccione el elemento del reporte
- En el editor de propiedades, seleccione la pestaña "Page Break"
- Establezca el salto de página cambiando la opción de salto de página de "Auto" a "Always"

En el flujo de trabajo del ejemplo, el salto de página se ha establecido después de la tabla de "remaining money" ..

7.23. Pestaña "Page Break" del editor de propiedades



## 7.9. Los Gráficos (Charts)

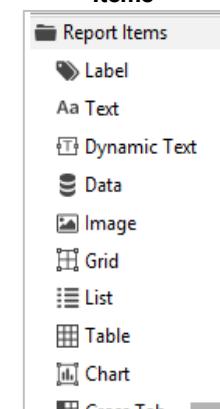
La parte final del informe consiste en dos gráficos que se colocan uno al lado del otro en la última fila de la cuadrícula. Un gráfico muestra el dinero asignado a lo largo de los años y el otro gráfico muestra el dinero utilizado a lo largo de los años. Los dos gráficos deben tener un aspecto idéntico. Para crear un gráfico, arrastre y suelte el elemento de informe "Chart" de la "Report Item List" en el editor de informes. Despues de soltar el gráfico, se abre el "Chart Wizard" para guiarle en la configuración de las propiedades adecuadas para el gráfico.

El "Chart Wizard" abarca tres pasos principales para todos los tipos de gráficos:

- Seleccionar el tipo de gráfico
- Seleccionar los datos
- Formatear el gráfico

El "Chart Wizard" puede reabrirse en cualquier momento haciendo doble clic en el gráfico

7.24. Elemento de informe "Chart" en el panel "Report Items"



# Seleccionar tipo de gráfico

El primer paso del "Chart Wizard" consiste en seleccionar el tipo de gráfico.

Hay muchos tipos de gráficos disponibles y cada tipo de gráfico tiene un número de subtipos de gráficos.

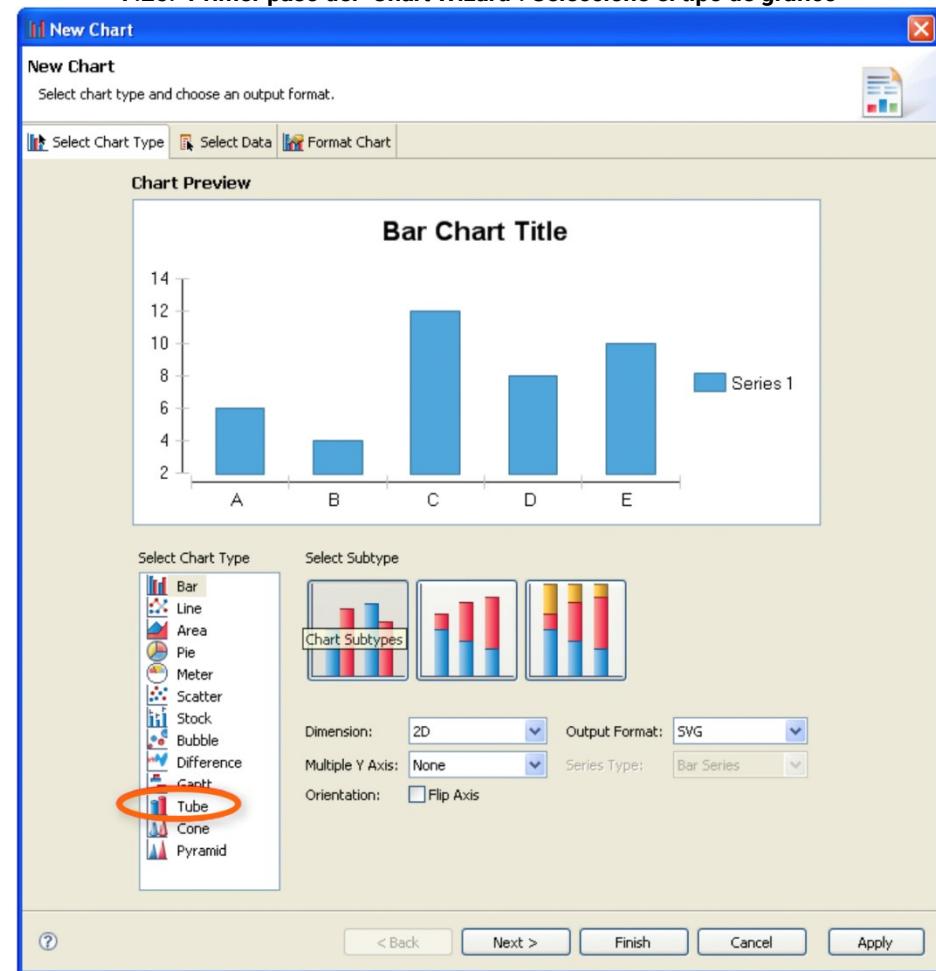
Además, cada gráfico puede representarse en 2D, en 2D con profundidad o en 3D.

Voltear el eje cambiará la orientación del gráfico. El eje X será entonces vertical y el eje Y horizontal.

- Seleccione el tipo de gráfico
- Haga clic en "Next" para pasar al siguiente paso del asistente de gráficos.

Elegimos un tipo de gráfico "Tube" en una dimensión 2D simple.

7.25. Primer paso del "Chart Wizard": Seleccione el tipo de gráfico

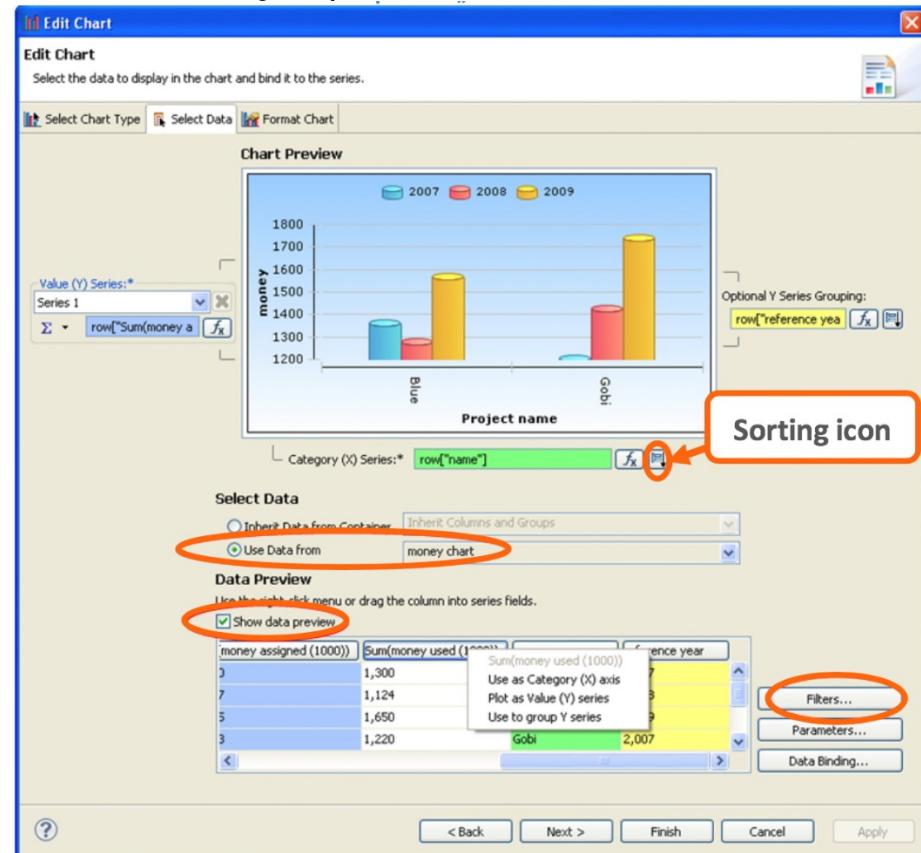


## Selecciona “Data”

El segundo paso es conectar el gráfico a un conjunto de datos.

- Vincule el gráfico con un Conjunto de Datos con la opción "Use Data from".
- En la tabla de vista previa de datos seleccione los datos de la columna para que estén en el eje X o en el eje Y o para que funcionen como datos de grupo. Haga clic con el botón derecho del ratón en la cabecera de la columna y seleccione una de esas opciones:
  - Usar como eje de categoría (X)
  - Graficar como serie de valores (Y)
  - Usar para agrupar series Y
- Si necesita series Y adicionales, seleccione "<New Series ...>" en el menú llamado "Value (Y) Series".
- Los datos de la categoría se ordenan en el eje X en orden descendente por defecto. Si no desea ninguna ordenación, haga clic en el icono de ordenación (el que tiene la flecha hacia abajo en el lado del cuadro de texto "Category (X) Series:" text box)) y desactive la "Grouping".
- A veces no es necesario mostrar todas las filas de datos del conjunto de datos en un gráfico. Para filtrar filas del conjunto de datos, haga clic en el botón "Filters" de la parte inferior derecha y añada reglas para incluir o excluir filas del conjunto de datos (véase más abajo).
- Haga clic en "Next" para pasar al siguiente paso del asistente.

7.26. Segundo paso del "Chart Wizard": Seleccionar Data



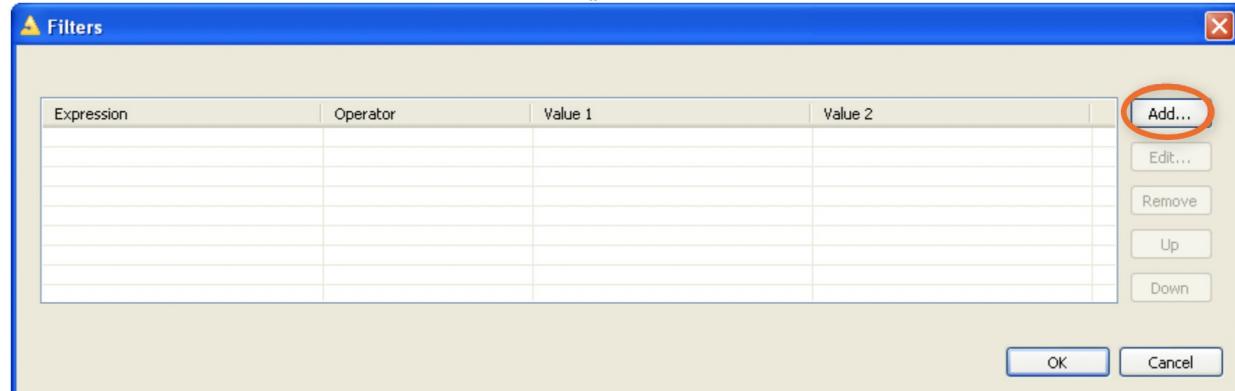
Para filtrar filas en el data set:

- Clickee el boton "Filters"
- En "Filters", clickee el botón "Add"

La ventana "New Filter Condition" aparecerá

Inserte su regla de filtrado en la ventana "New Filter Condition"

7.27. La ventana „Filters“



Aquí a la derecha hay un ejemplo de una regla de filtrado que excluye todas las filas en las que la columna "name" = "total". Observe que "total" está entre comillas. No olvide las comillas en una comparación de cadenas, ya que BIRT necesita las comillas para reconocer las cadenas.

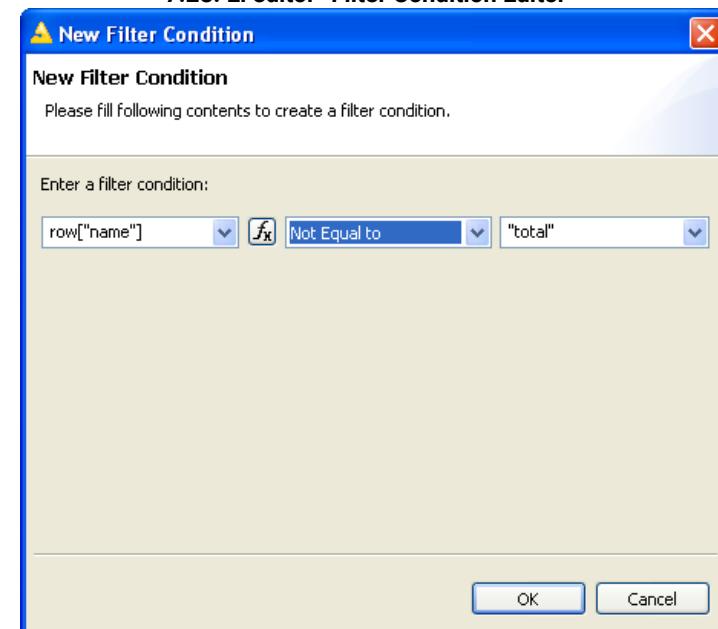
El primer gráfico debe mostrar el dinero asignado a lo largo de los años.

Seleccionamos:

- Conjunto de datos "money chart"
- Columna "name" como Serie de categoría (eje X) sin clasificar
- Columna "Sum(assigned money(1000))" como Serie Y
- Columna "reference year" para agrupar las series Y

Sólo hemos representado una serie Y en este gráfico y no se ha aplicado ningún filtro a las filas del conjunto de datos.

7.28. El editor "Filter Condition Editor"



# Formatear el Gráfico

El último paso del Asistente le guía en la configuración del diseño del gráfico.

A la izquierda, un árbol muestra las opciones de formato del gráfico.

En **"Series"** puede cambiar el nombre de la serie Y. Los nombres por defecto son "Series 1", "Series 2", etc....

En **"Value (Y) Series"** puede añadir y formatear etiquetas encima de cada punto del gráfico.

Under **"Chart Area"** puede definir el color y el estilo del fondo

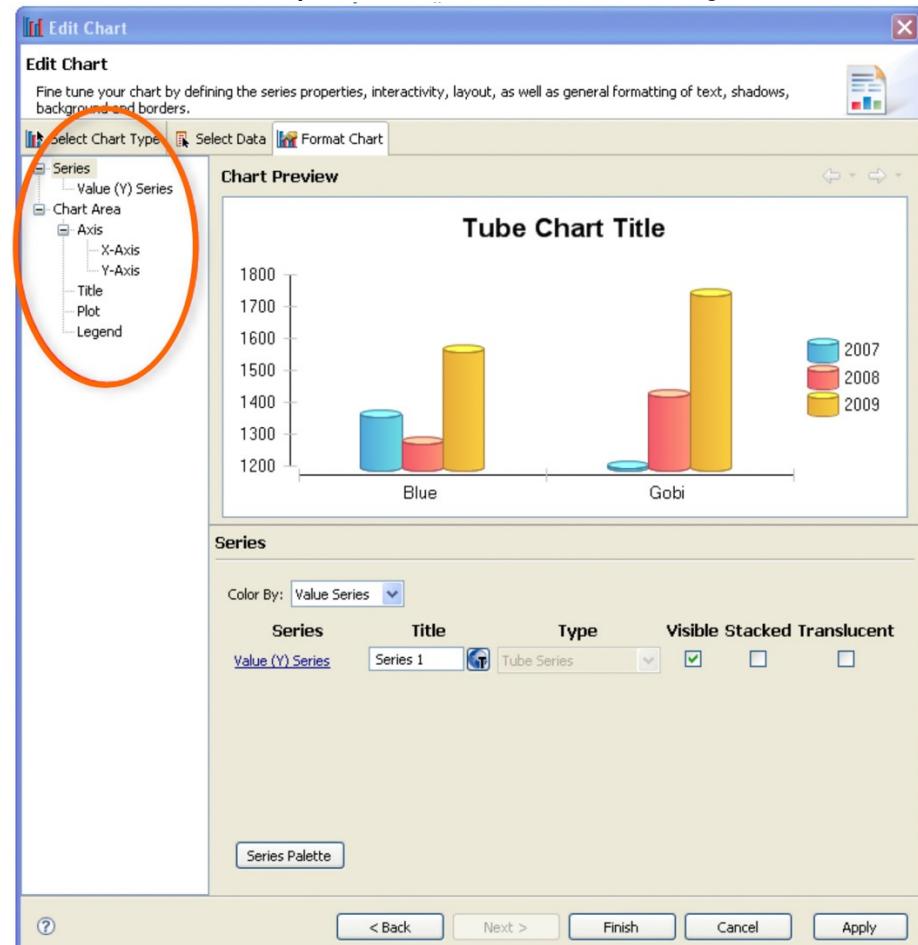
Under **"Axis"**, puede definir las etiquetas, la escala, las líneas de la cuadrícula y todo lo relacionado con el eje del gráfico (X-axis or Y-axis).

**"Title"** tiene opciones para el texto del título, el diseño y la fuente.

**"Plot"** es similar a "Chart Area", pero se refiere sólo al espacio de trazado.

**"Legend"** le ayuda con la posición, el diseño, las propiedades de la fuente y todo lo relacionado con la leyenda del gráfico.

7.29. Tercer paso del "Chart Wizard": Formatear el gráfico



## Series

En "Series" puede cambiar el nombre (etiquetado como "Title") de cada serie Y. Los nombres por defecto son simplemente "Serie 1", "Serie 2", etc.... que no son muy significativos. Las series Y pueden ocultarse desactivando la casilla "Visible" situada a la derecha del cuadro de texto "Title".

El botón "Series Palette" permite elegir los colores de la serie Y. Puede seleccionar un color diferente para cada uno de los valores de la serie Y.

Cambiamos el nombre de la serie Y de "Serie 1" a "money assigned". Este nombre aparecerá en la leyenda. Mantenemos la paleta de series por defecto.

### Value (Y) Series

En la "Value (Y) Series" puede añadir etiquetas encima de cada punto del gráfico, activando la opción "Show Series Labels".

El botón "Labels" abre la ventana "Series Labels" para formatear las etiquetas de la serie.

### Ventana "Series Label"

La ventana "Series Labels" nos ayuda a formatear las etiquetas encima de cada punto del gráfico, siempre que decidamos hacerlas visibles.

Aquí puede definir la posición de la etiqueta, la fuente, el fondo, la sombra, el contorno e incluso los puntos de inserción.

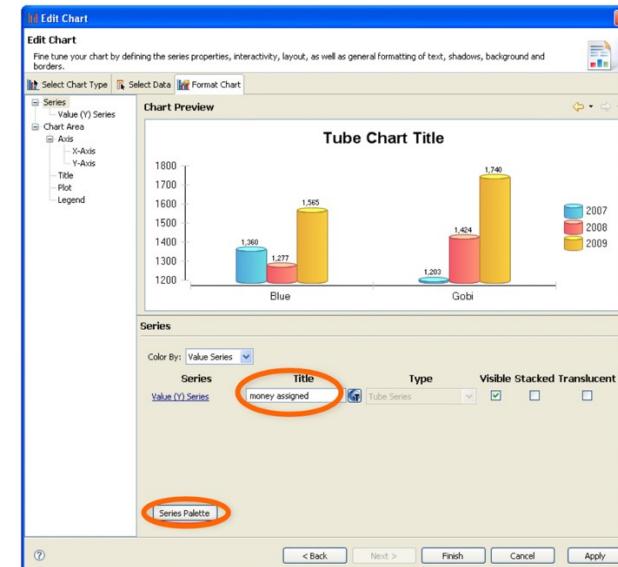
También puede definir qué valores quiere que se muestren encima de cada punto: valor Y actual, valor Y porcentual, valor X o nombre de la serie. La etiqueta también puede construirse alrededor del valor mostrado con un prefijo, un sufijo y un separador.

El pequeño botón con una "A" dentro lleva al "Editor de fuentes".

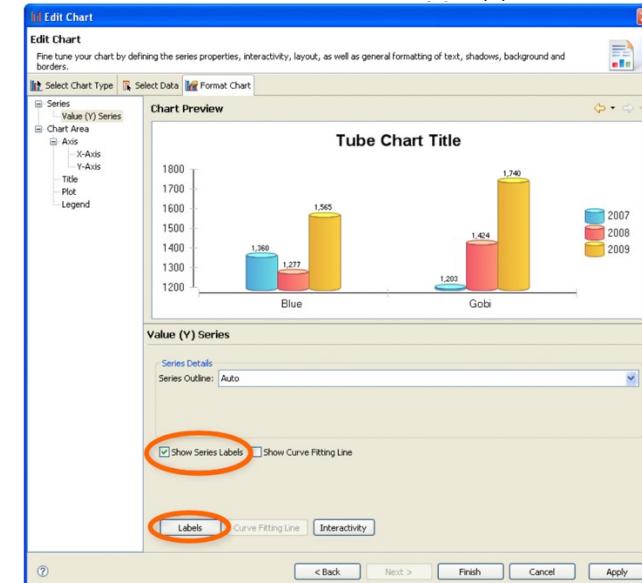
El botón "Format" lleva al "Format Editor" (debe seleccionar un elemento en la lista de "Values" para activar este botón).

No hay ningún botón de "OK" o "Cancel" en este diálogo de "Series Labels". Los nuevos ajustes se aplican inmediatamente. Para el informe "Projects" hemos decidido hacer visibles las etiquetas de las series.

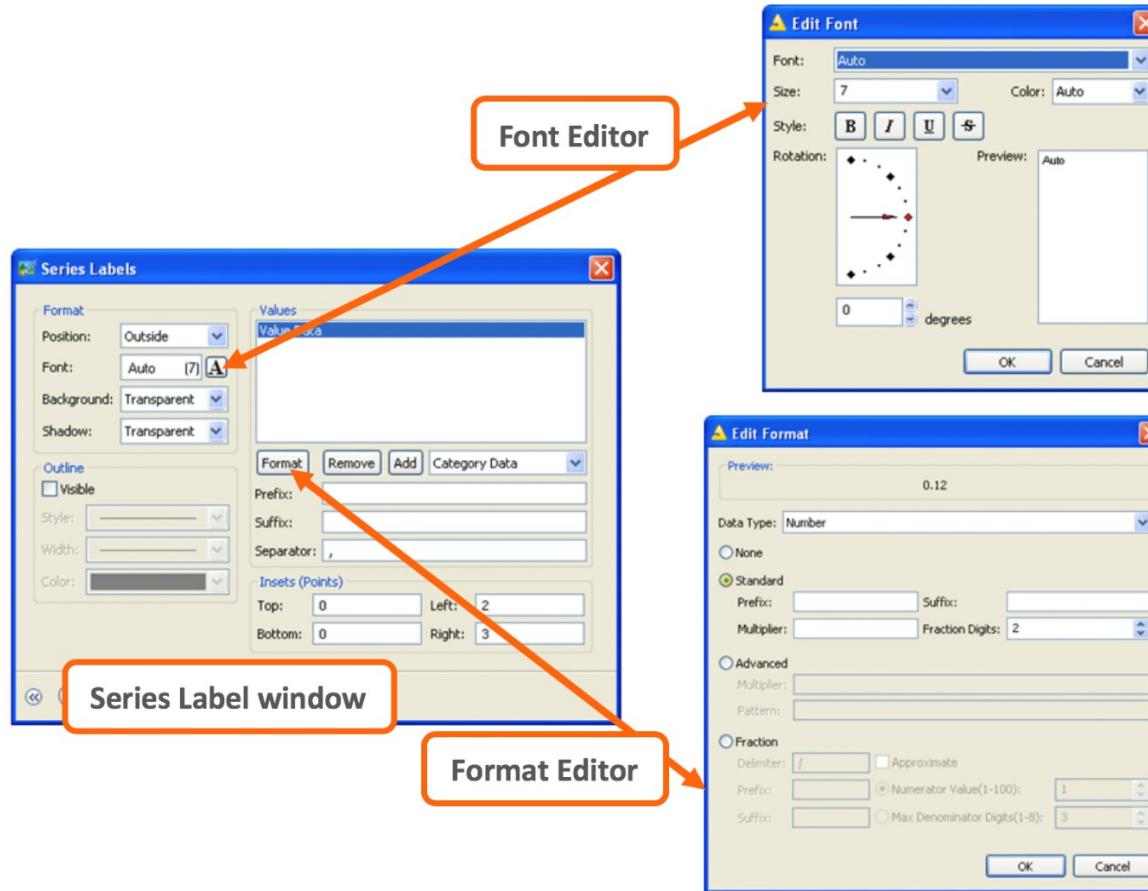
### 7.30. Formato : Series



### 7.31. Formato: Series "Value (Y)"



7.32. La ventana "Series Labels". El botón "A" abre el "Font Editor". El botón "Format" abre el "Format Editor".



### Editor de Fuentes (Font editor)

El "Font Editor" es una ventana estándar que encontrará en el paso "Format Chart" en cualquier lugar, donde es posible cambiar un formato de fuente. Contiene las opciones habituales de formato de fuente: nombre de la fuente, tamaño, estilo, color. La nueva opción es "Rotation".

"Rotation" hace girar la etiqueta el número de grados requerido. "0 grados" (= el punto rojo en el taquímetro) corresponde a etiquetas escritas horizontalmente. "-90 degrees" escribe las etiquetas verticalmente de arriba a abajo. "+90 degrees" escribe las etiquetas todavía verticalmente pero de abajo hacia arriba. "-45 degrees" escribe las etiquetas en una línea pendiente de 45 grados desde arriba hacia abajo.

Y así sucesivamente... La opción "Rotation" es muy útil para gráficos abarrotados o para etiquetas muy largas. For the charts in the report "Projects" the only setting we made was to specify the series labels font size as 7.

## Editor de formatos

El "Format Editor" se utiliza para formatear valores numéricos, fechas e incluso cadenas. Sin embargo, el uso más común es para formatear números.

Hay 4 formatos numéricos posibles: ninguno, estándar, avanzado, fracción. El multiplicador se utiliza para representar números con cadenas más pequeñas, por ejemplo el dinero en millones de unidades en lugar de en moneda real. Los dígitos de la fracción son los que van después de la coma. El prefijo y el sufijo también están disponibles para formatear cadenas y se utilizan para construir una etiqueta alrededor del valor básico.

En nuestro gráfico hemos formateado las etiquetas de las series en "Value data" (es decir, los datos de la serie) utilizando 2 dígitos decimales.

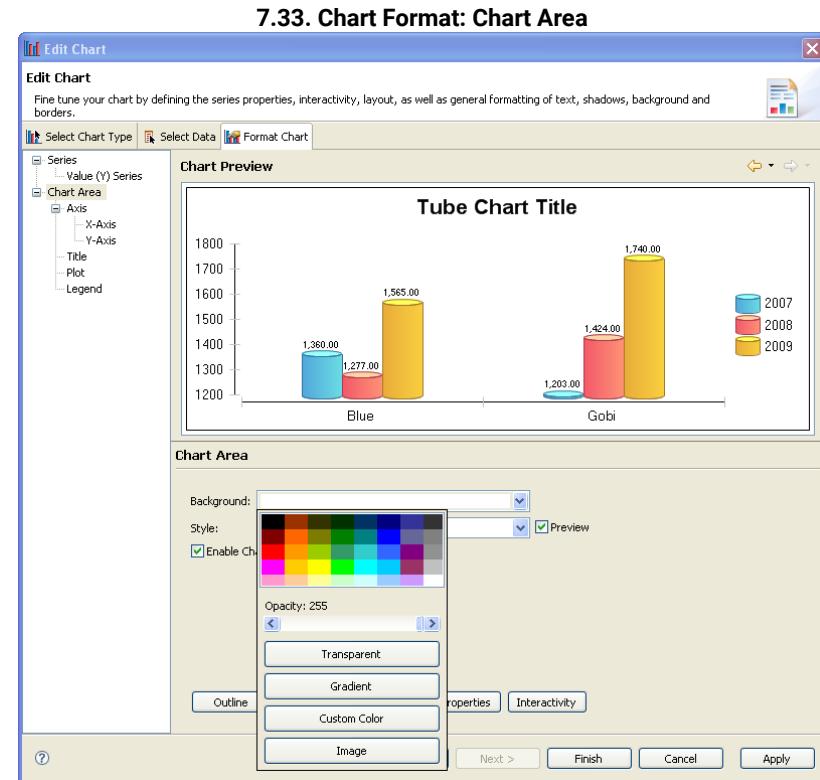
## Área del Gráfico

En el "Chart Area" puede definir el color y el estilo del fondo del gráfico.

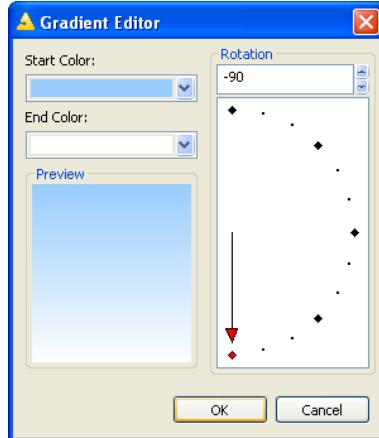
Si hace clic en el menú "Background", se le mostrarán una serie de opciones que puede utilizar para definir el fondo:

- Un color simple
- "Transparent", es decir, sin color de fondo
- Un degradado entre dos colores
- Un color personalizado
- Una imagen

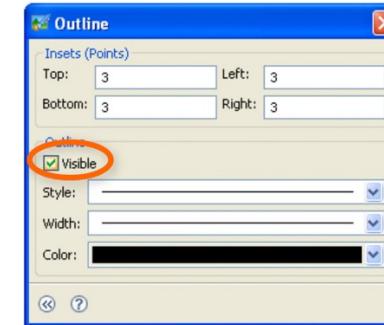
Seleccionamos la opción "Gradient". El "Gradient" necesita el color inicial y final y la dirección del degradado expresada en grados. Por último, hacemos visible el contorno del gráfico haciendo clic en el botón "Outline" y activando la opción "Visible" en el "Outline Editor".



**7.34. El Editor "Gradient"**



**7.35. El "Outline Editor"**



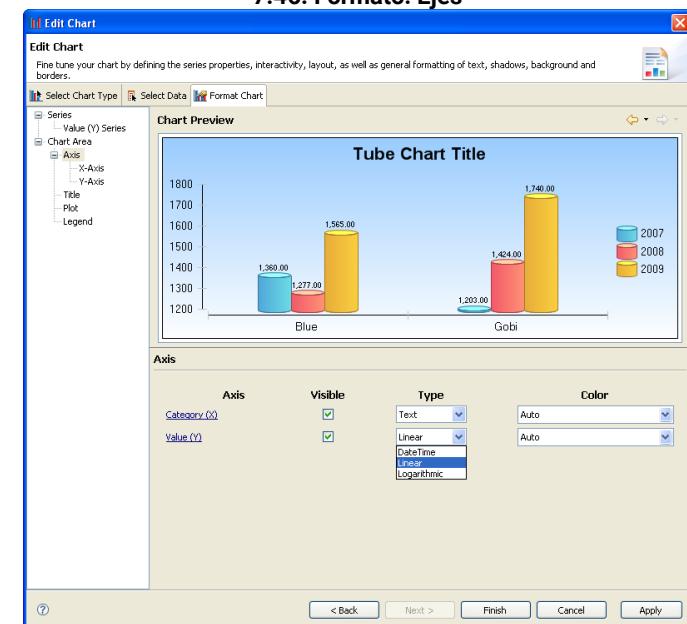
## Axis

En "Axis", puede definir el tipo y el color de los ejes X e Y. Hay varios tipos de ejes disponibles en función de los tipos de valores mostrados en el eje (Texto, Número o Fecha). Los ejes lineales y logarítmicos sólo se aplican a los valores numéricos.

Dejemos la escala lineal por defecto para el eje de valores (Y).

Todos los demás ajustes de los ejes, como los tipos de letra, las líneas de cuadrícula y la escala, pueden definirse para cada eje por separado. Las dos ventanas para los ajustes del eje X y del eje Y son casi idénticas, además de dos opciones de categoría en el marco del eje X.

**7.46. Formato: Ejes**



## Ejes X/Y (X-Axis / Y-Axis)

Aquí el usuario puede establecer un título apropiado y hacerlo visible.

La parte más importante es definir las etiquetas de los ejes: formato, fuente y diseño. El botón habitual "A" lleva al usuario al "Font Editor".

El botón con el icono del formato lleva al "Font Editor".

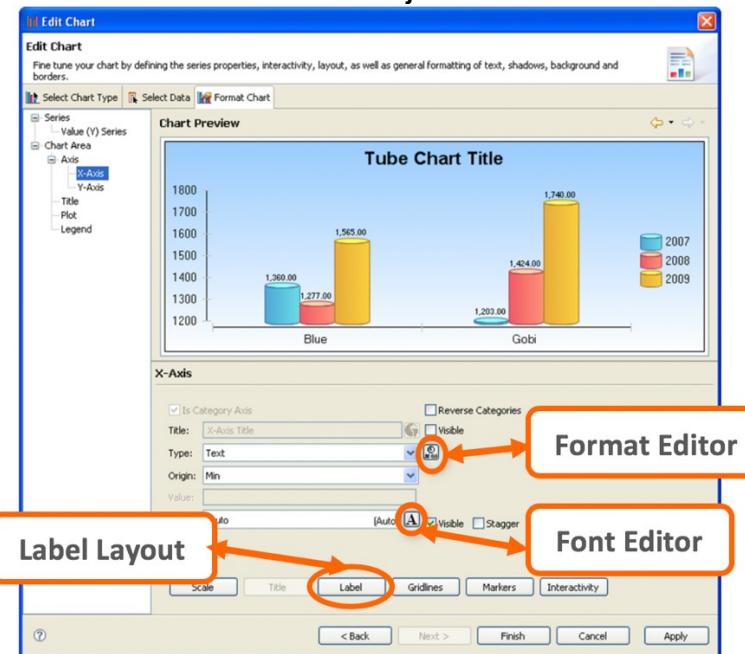
El botón "Label" lleva al "Label Layout Editor", donde podemos definir la posición de la etiqueta, el fondo, el contorno, etc....

El botón "Scale" define el tamaño del paso para los valores numéricos en el eje. Está desactivado para los valores de texto.

El botón "Title" define el tipo de letra y el diseño del título del eje si se ha activado la casilla para que el título sea visible.

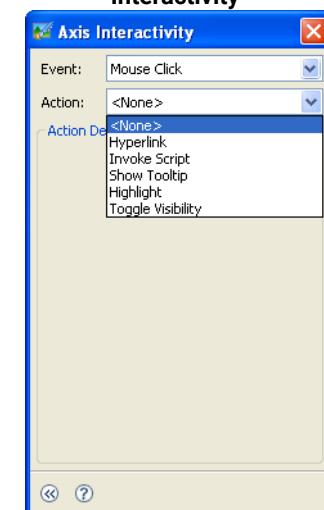
El botón "Marker" introduce líneas para marcar áreas del gráfico.

## 7.37. Formato Ejes: X-axis



El botón "Interactivity" abre la ventana "Axis Interactivity", donde se puede establecer una acción para seguir un evento. Esto se utiliza para los cuadros de mando o los informes html. Por ejemplo, un clic del ratón puede iniciar un script Java. Están disponibles muchos eventos, como el clic del ratón, y muchas acciones, como un hipervínculo o un script.

## 7.38. La ventana "Axis Interactivity"

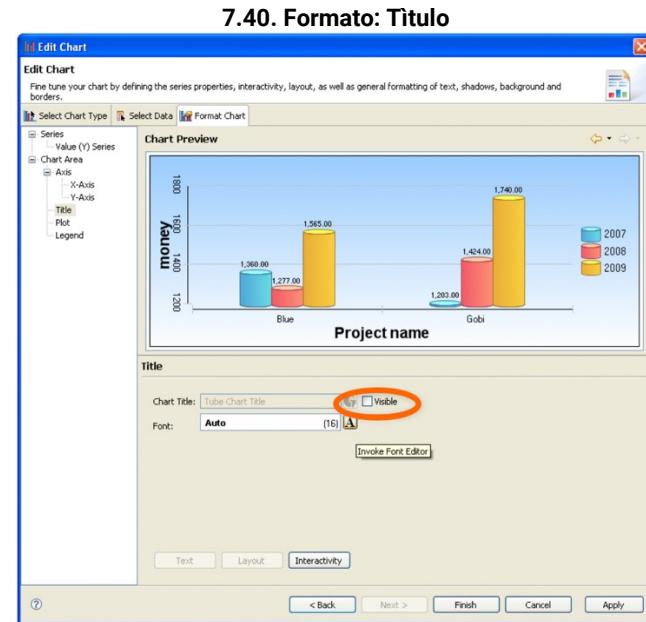


El botón "Gridlines" abre la ventana "Axis Gridlines" para activar las líneas de cuadrícula para este eje; es decir, líneas de cuadrícula horizontales para el eje Y y líneas de cuadrícula verticales para el eje X.

Hay cuadrículas mayores y menores en el gráfico, así como marcas en el eje.

Para el informe "Projects" hemos habilitado lo siguiente:

- Líneas de cuadrícula en el eje Y, cuadrícula mayor y ticks en la cuadrícula mayor solamente. Pasamos por alto la rejilla menor para no hacer el gráfico demasiado lleno.
- Etiquetas con tamaño de fuente 7 y giradas a -90 grados en el eje X
- Título visible en ambos ejes con "Project name" como texto para el eje X y "money" para el eje Y, tamaño de fuente 8 y girado a -90 grados en el eje Y



#### 7.40. Formato: Título

"Title" establece un título en el gráfico. Si habilita que el título sea visible, el marco "Title" tiene opciones para el diseño del título, la fuente y la interactividad. Por lo general, no configuro el título para que sea visible, porque le quita espacio al gráfico. Utilizo una etiqueta en la parte superior del gráfico en el diseño del informe para que actúe como título del gráfico.



## Plot

"Plot" es similar a "Chart Area", pero se refiere sólo al espacio de trazado.

## Legend

"Legend" establece la posición, el diseño, las propiedades de la fuente y todo lo relacionado con la leyenda del gráfico.

Si decide incluir una leyenda en el gráfico, en primer lugar debe hacerla visible en el marco de la leyenda (casilla "Visible" al principio del marco "Legend").

A continuación, debe definir el diseño de la leyenda (botón "Diseño") y las propiedades de la fuente (botón "Entradas").

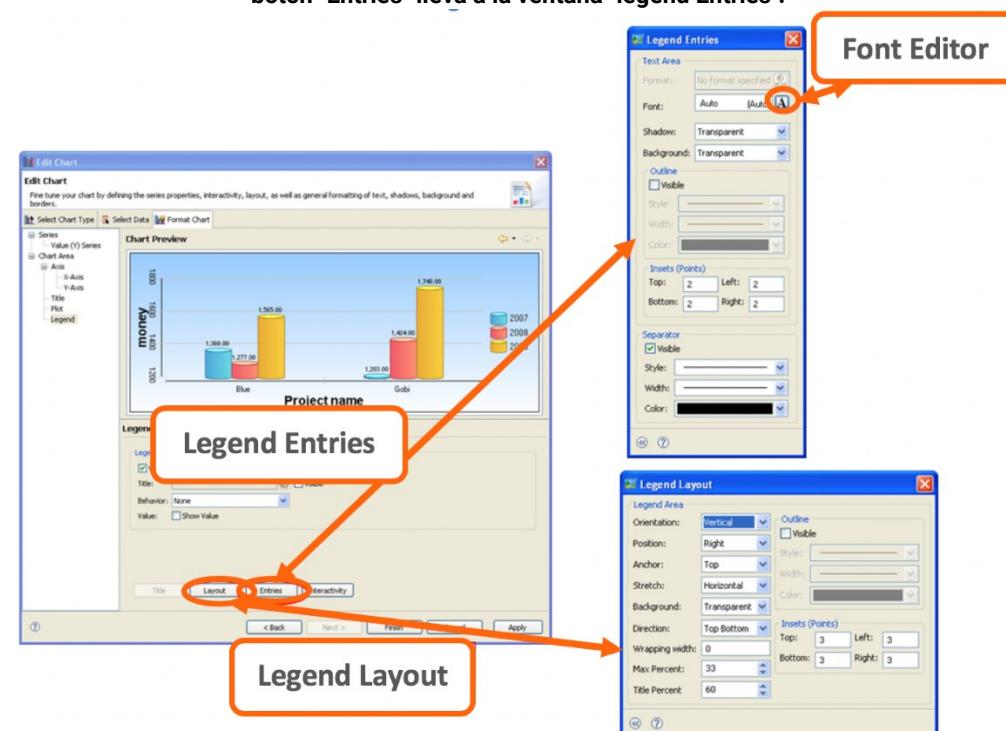
En "Projects" establecemos las siguientes propiedades para la leyenda:

- Font size: 7
- Orientation: horizontal
- Direction: left to right
- Position: above

Cuando haya terminado de formatear el gráfico, haga clic en "Finish". El asistente de gráficos le llevará de nuevo al informe.

Cambie el tamaño del gráfico para que se ajuste a la celda de la cuadrícula. Inserte una etiqueta encima del gráfico para hacer el título del mismo, por ejemplo donde el texto es "money assigned per year to each project".

**7.41. Formato del gráfico: Legend.** El botón "Layout" lleva a la ventana "Legend Layout". El botón "Entries" lleva a la ventana "legend Entries".



## Cambiar una propiedad de formato

Ejecute una vista previa del documento. Si no le gusta el aspecto del gráfico, vuelva a la pestaña "Layout", haga doble clic en el gráfico y cambie la configuración que no le haya gustado. En el informe "Projects", por ejemplo, las "Series Labels" parecen demasiado recargadas.

Para desactivar las "Series Labels":

- Haga doble clic en el gráfico
- En la parte superior, seleccione la pestaña "Format Chart".
- Seleccione "Value (Y) Series".
- Desactive la casilla "Show Series Labels".
- Haga clic en el botón "Finish"

## Change data assignment

We need to create an identical chart on the right cell of the grid, but with reference to the money used instead of the money assigned.

- Copiar y pegar el gráfico y su etiqueta de título de la celda de la izquierda a la celda de la derecha
- Haga doble clic en el gráfico de la derechaSelect the "Select Data" tab
- En "Chart Preview", haga clic con el botón derecho del ratón en la cabecera de la columna "Sum(money used (1000))"
- Seleccione "Plot as Value Y Series"
- Clickee en el botón "Finish"

## 7.10. Hojas de estilo (Style Sheets)

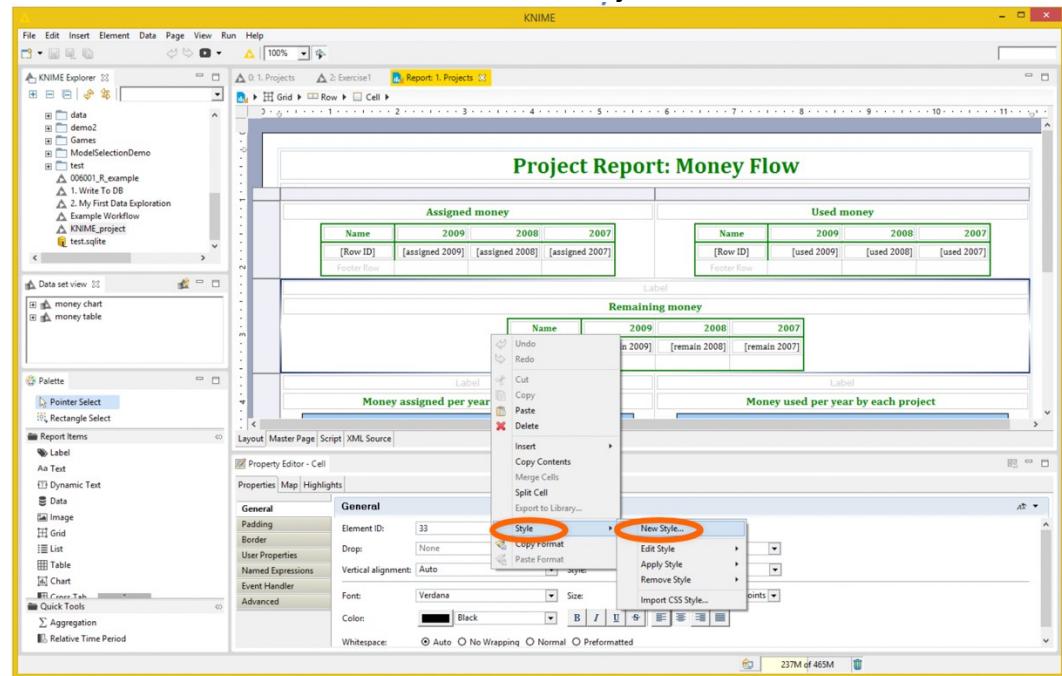
A veces puede resultar tedioso dar formato a todos los elementos individuales de un elemento del informe, especialmente si muchos de estos elementos del informe tienen que ser formateados con el mismo estilo. Por ejemplo, en la sección anterior debíamos dar formato a todas las celdas de datos y de cabecera de tres tablas de la misma manera. Para evitar tener que repetir esas tediosas operaciones, podemos utilizar las hojas de estilo.

Las hojas de estilo se utilizan ampliamente en la programación web para compartir las especificaciones de estilo entre los numerosos elementos de las páginas web. Del mismo modo, la herramienta de informes KNIME admite hojas de estilo que pueden utilizarse para aplicar atributos de estilo a múltiples elementos del informe.

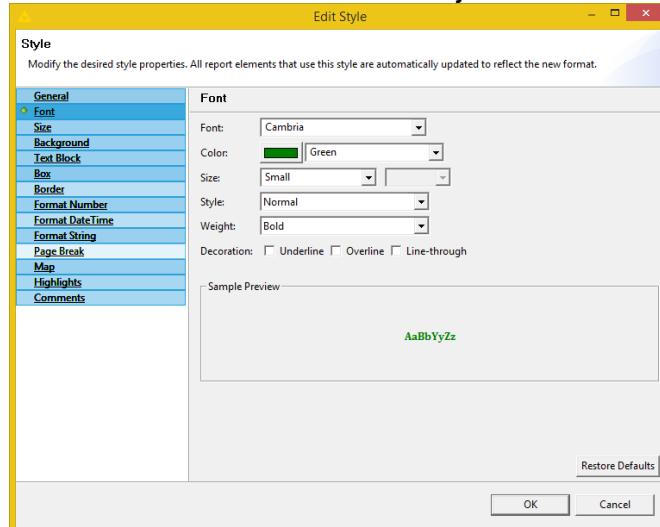
# Crear un Nuevo estilo

- Haga clic con el botón derecho del ratón en cualquier lugar del editor de informes
- Seleccione "Style".
- Seleccione "New Style".
- Se abre la ventana "New Style" ..

7.42. Crear una nueva Hoja de Estilo



7.43. La ventana "Edit Style"



En la ventana "New Style", hay que definir:

- El nombre de la hoja de estilo en la pestaña "General".
- Las propiedades de la fuente en la pestaña "Font".
- Las propiedades de los números en la pestaña "Format Number"

Y así con más propiedades en otras pestañas

Tomando como ejemplo las tablas de la sección anterior, es fácil ver que hay dos grupos de celdas para cada tabla:

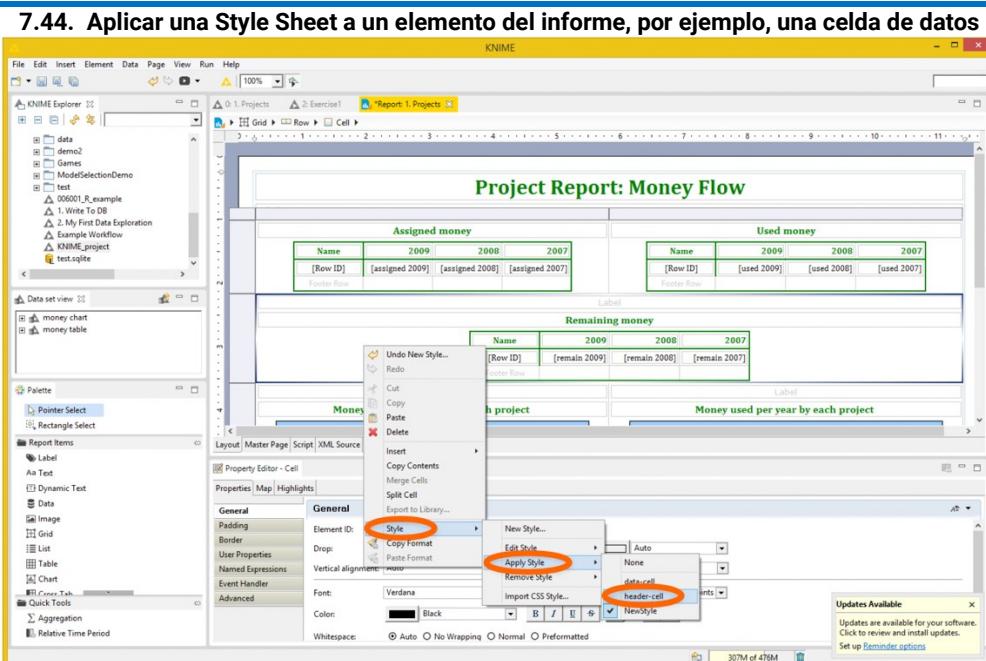
- Celdas de encabezado con fuente "Cambria", tamaño de fuente "10 puntos", estilo de fuente "negrita" y color de fuente "verde"
- Celdas de datos con fuente "Cambria", tamaño de fuente "10 puntos", y formato de número con 2 decimales y separador de 1000s

A continuación, construimos dos hojas de estilo, una para las celdas de datos y otra para las celdas de cabecera con las propiedades indicadas anteriormente. Elegimos un tamaño de fuente "large" para ambas hojas de estilo, las llamamos "data cell" y "header cell" y las aplicamos a cada celda de cabecera y a cada celda de datos de las tres tablas.

**Nota.** No todos los tamaños de fuente están disponibles en el editor de la Hoja de Estilo como en el Editor de Propiedades. Sólo se pueden utilizar algunos tamaños de fuente predefinidos en una Hoja de Estilo.

## Aplicando un Style Sheet

- Haga clic con el botón derecho del ratón en el elemento del informe (celda de la tabla, etiqueta, etc.)
- Seleccione "Style".
- Seleccione "Apply Style"
- Seleccione el nombre de Style Sheet que desea aplicar



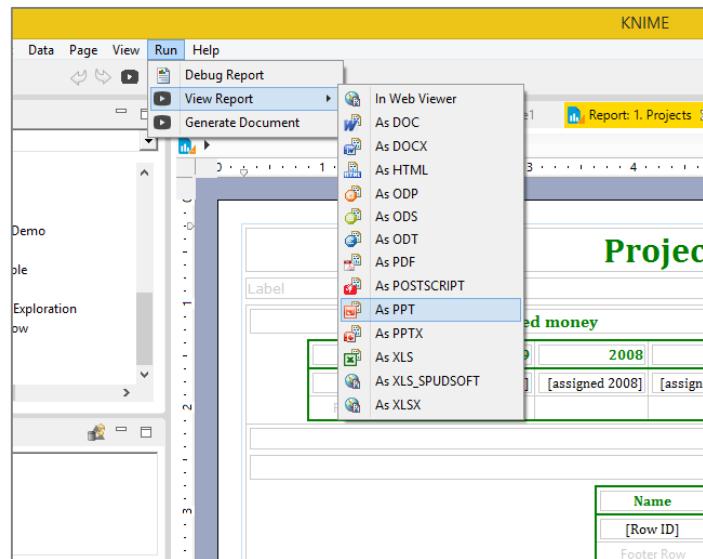
## 7.11. Generar el documento final

In order to generate the final document, go to the Top Menu:

- Seleccione "Run"
- Seleccione "View Report"
- Seleccione el formato de su informe, por ejemplo "PPT" for Powerpoint
- BIRT genera el documento en el formato deseado.

Alternativamente, "Run" → "Generate Document" genera directamente el archivo en el formato preferido

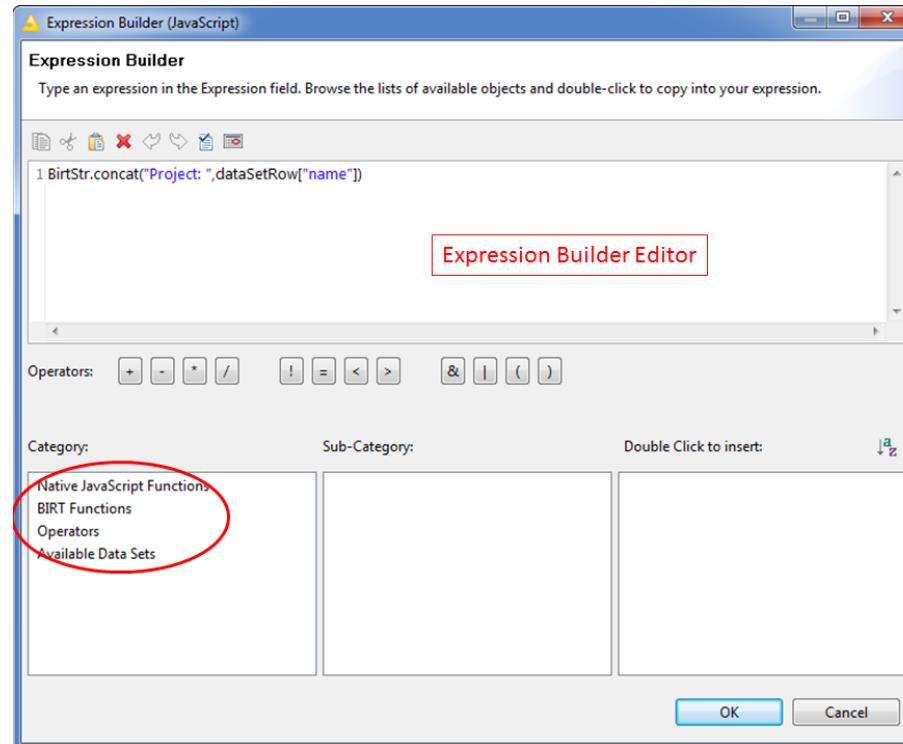
7.45. Generar el documento final



## 7.12. Texto Dinámico

Un elemento de informe dinámico es el "Dynamic Text", que se encuentra en la lista de "Report Items", en el panel inferior izquierdo. El elemento "Report Items" muestra un pequeño texto, construido con el "Expression Builder". La ventana del "Expression Builder" ofrece la lista de operadores y funciones de BIRT y Java Script. Si el texto dinámico está incluido en una tabla, aparecen algunas opciones adicionales en el "Expression Builder", como por ejemplo los "Available data Sets".

7.46. Ventana "Expression Builder"



**"Operators"** ofrece una serie de operadores matemáticos/lógicos. En el panel central se muestra un extracto con los operadores más utilizados.

**"BIRT Functions"** La categoría incluye una serie de funciones de BIRT específicamente diseñadas en los campos de las finanzas, la manipulación de la fecha/hora, la duración, las matemáticas, la manipulación de cadenas y la comparación de cadenas o números..

**"Native JavaScript Functions"** incluye una serie de funciones de JavaScript. Éstas resultan especialmente útiles cuando el informe se crea utilizando el formato HTML.

Al hacer doble clic en un elemento de un subpanel, como el nombre de una columna, una función BIRT, un operador o una función Java Script, se inserta automáticamente el elemento en el editor de Expression Builder de arriba.

**"BIRT Functions"** y **"JavaScript Functions"** Las categorías albergan algunas funciones útiles, por ejemplo las funciones de fecha y hora y las funciones matemáticas. Podríamos insertar la fecha actual como un título en curso en el informe, a la derecha del logotipo.

En el entorno de los informes, en la pestaña "Master Page", en el marco superior de la cabecera, insertamos una cuadrícula con dos columnas y una fila. La celda de la izquierda ya incluía el logotipo. En la celda de la derecha queríamos insertar un elemento de "Dynamic Text" que mostrara la fecha actual en la creación del informe. Arrastramos y soltamos un elemento de "Dynamic Text" desde el panel de lista "Report Items" de la parte inferior izquierda a la celda derecha de la cuadrícula en el marco superior de la cabecera de la pestaña "Master Page". Se abre la ventana del "Expression Builder". Para mostrar la fecha actual, tenemos muchas opciones.

Podemos elegir, por ejemplo, una sencilla function "BIRT Functions" → "BirtDateTime" → "Today()" . "Today()" devuelve una fecha (que es la medianoche de la fecha actual).

#### 7.47. Fecha actual de la función BIRT "Today()



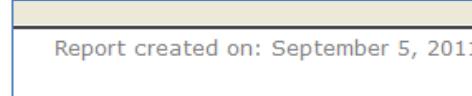
Sep 5, 2011 12:00 AM

La función "Today()" no ofrece ninguna opción de formato. Si queremos tener la fecha actual en formato personalizado debemos construirlo nosotros mismos. "BIRT Functions" → "BirtDateTime" ofrece una serie de funciones para extraer componentes de un DateTime object, tal como day(DateTime), month(DateTime), year(DateTime), etc.. Podríamos extraer los componentes de la fecha y combinarlos con una función BirtStr.concat() para obtener el formato de fecha deseado. Tras extraer las partes de la fecha del resultado de la función Today() y combinarlas con una función concat(), obtenemos, por ejemplo, la siguiente fórmula en la ventana "Expression Builder":

```
BirtStr.concat( "Report created on: ",  
                BirtDateTime.month(BirtDateTime.today(), 2), " ",  
                BirtDateTime.day(BirtDateTime.today()), " ",  
                BirtDateTime.year(BirtDateTime.today()))
```

y el siguiente formato de fecha actual en el título del informe:

#### 7.48. Personalización de la fecha actual a partir de la función BIRT "Today()"



Report created on: September 5, 2011

"BIRT Functions" → "BirtDateTime" ofrece una serie de funciones para sumar y restar tiempo a un objeto DateTime. Por ejemplo, el título de la carrera podría utilizar la siguiente fórmula con la función addQuarter():

```
BirtStr.concat("Report valid from: ",  
                BirtDateTime.today(),  
                " to: ",  
                BirtDateTime.addQuarter(BirtDateTime.today(),1)  
            )
```

Que produce una fecha en el título en curso como la siguiente.

#### 7.49. Utilización de la función "addQuarter()" en el título de ejecución

Report valid from: Mon Sep 05 00:00:00 CEST 2011 to: Mon Dec 05 00:00:00 CET 2011

**Nota.** El cambio en la configuración regional de la fecha se debe a la introducción de la función concat(), que también establece automáticamente la configuración regional de los valores DateTime.

**Nota.** En el editor de Expression Builder las cadenas de texto deben escribirse entre comillas (estilo Java). Los elementos de texto en una función concat() deben estar separados por una coma.

En esta sección sólo hemos mostrado las funciones BIRT relacionadas con el objeto DateTime, porque son las más utilizadas. Sin embargo, la categoría "Funciones BIRT" ofrece muchas funciones incorporadas para expresiones matemáticas, cantidades financieras, manipulación de cadenas, etc...

Si el documento de salida del informe es HTML, también podríamos aprovechar las funciones incorporadas de JavaScript, que son más articuladas y variadas que las funciones incorporadas de BIRT.

## 7.13. Informes con otras herramientas

De forma similar a la utilización de BIRT para construir su informe, también puede utilizar cualquier otra herramienta de informes disponible en el mercado. La mayoría de estas herramientas requieren una licencia de pago, por lo que no se describirán aquí en detalle. En el material que ha descargado, encontrará en la carpeta Capítulo7 unos cuantos flujos de trabajo que muestran los nodos KNIME dedicados a exportar los datos a la herramienta de elaboración de informes elegida.

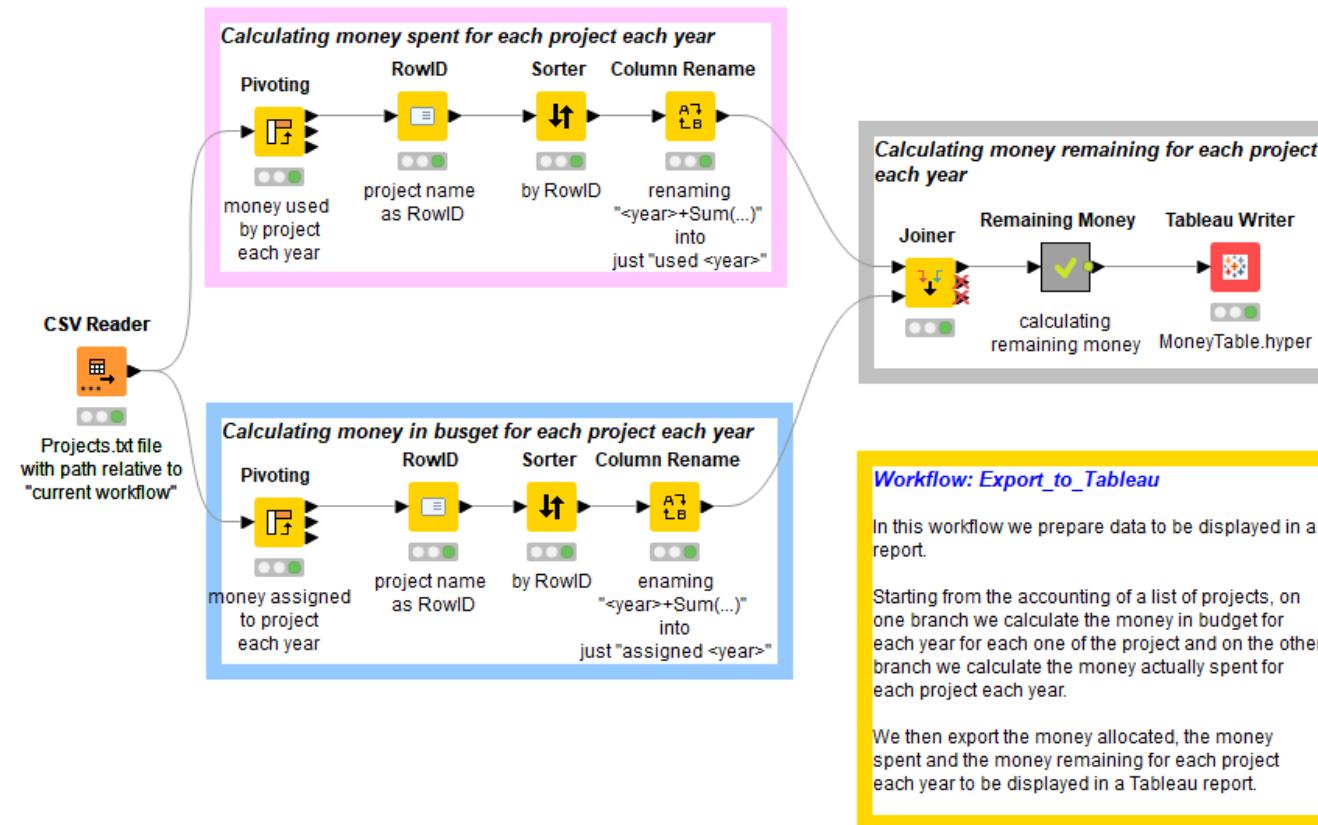
El flujo de trabajo "Export\_to\_Tableau" incluye el nodo "Tableau Writer" para escribir los datos en un archivo con formato Tableau que posteriormente se importará en la plataforma Tableau para construir el informe. Otro nodo llamado "Send to Tableau Server" permite la transferencia directa de datos desde KNIME Analytics Platform a Tableau.

El flujo de trabajo "Export\_to\_Spotfire", de forma similar, utiliza el nodo "TIBCO Spotfire File Writer" para exportar los datos a un archivo con formato TIBCO Spotfire que posteriormente se importará en Spotfire para construir el informe. Además del nodo "TIBCO Spotfire File Writer",

un nodo "TIBCO Spotfire File Reader" permite leer archivos con formato TIBCO Spotfire y un nodo "TIBCO Spotfire Information Link Reader" conecta directamente con el servidor Spotfire.

El flujo de trabajo "Export\_to\_PowerBI" utiliza el nodo "Send to Power BI", previa autentificación de Microsoft, para transferir los datos directamente a un servidor PowerBI.

**7.50. El flujo de trabajo "Export\_to\_Tableau" exporta los datos a un archivo con formato de Tableau para ser importado posteriormente en la plataforma de informes de Tableau**



## 7.14. Ejercicios

Los ejercicios de éste capítulo son la continuación de los ejercicios del capítulo 5. En particular, requieren la elaboración de un diseño de informe para los conjuntos de datos construidos en los ejercicios del Capítulo 5.

### Ejercicio 1

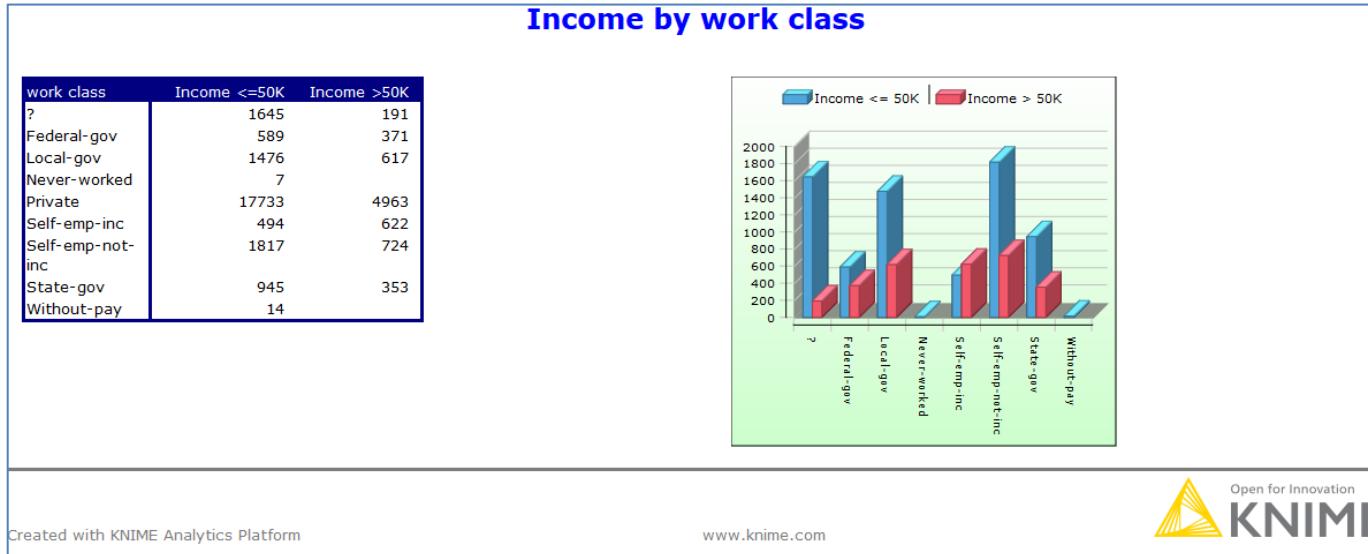
- Utilizando el flujo de trabajo construido en el Capítulo 5/Ejercicio 1, construya un informe BIRT con:
  - Un título "income by work class"
  - Una tabla en el lado izquierdo como

Work class	Income <= 50K	Income > 50K
[work class]	[nr <= 50K]	[nr > 50K]

- Un gráfico de barras con:
  - Work class en el eje X ( X-axis)
  - "Income <= 50K" y "Income > 50K" en el Y-axis
  - Estilo gradiente
  - Font size 7 on the axis
  - Font size 8 in the legend
  - Leyenda situada encima de la parcela y en horizontal
  - Sin título
  - Sin títulos en los ejes
- Exportar como documento Word

## Solución

### 7.51. Ejercicio 1: Reporte



# Referencias

- [1] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Koetter, T. Meinl, P. Ohl, C. Sieb, and B. Wiswedel, "KNIME: The Konstanz Information Miner". KDD 2006 ([http://www.kdd2006.com/docs/KDD06\\_Demo\\_13\\_Knime.pdf](http://www.kdd2006.com/docs/KDD06_Demo_13_Knime.pdf))
- [2] D. Peh, N. Hague, J. Tatchell, "BIRT. A field Guide to Reporting", Addison-Wesley, 2008
- [3] C.M. Bishop, "Pattern Recognition and Machine Learning", Springer (2007)
- [4] M.R. Berthold, D.J. Hand, "Intelligent Data Analysis: An Introduction", Springer Verlag, 1999
- [5] M.R. Berthold, C. Borgelt , F. Höppner, F. Klawonn, "Guide to intelligent data analysis", Springer 2010
- [6] D. L. Olson, D. Delen, "Advanced Data Mining Techniques" Springer; 2008
- [7] D.G. Altman, J.M. Bland, "Diagnostic tests. 1: Sensitivity and specificity" *BMJ* 308 (6943): 1552; 1994
- [8] J.R. Quinlan, "C4.5 Programs for machine learning", Morgan Kaufmann Publishers Inc. , 1993
- [9] J. Shafer, R. Agrawal, M. Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining", Proceedings of the 26th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc. ,1996 (<http://citeseer.ist.psu.edu/shafer96sprint.html>)
- [10] B. Wiswedel, M.R. Berthold, "Fuzzy Clustering in Parallel Universes" , International Journal of Approximate Reasoning, Elsevier Inc., 2007
- [11] R. Silipo, J. Prinz, "KNIME Advanced Luck", KNIME Press (2019) (<https://www.knime.com/knimepress/knime-advanced-luck>)

# Índice de temas

## A

Accuracy.....	176
<b>Accuracy Measures</b> .....	175
Aggregations.....	125
Annotations.....	44
Artificial Neural Network .....	189

## B

Bar Chart .....	144
Bar Charts.....	143
Binning.....	125
BIRT Functions.....	291

## C

Case Converter.....	105
Cell Splitter.....	100
Cell Splitter by Position.....	99
Chart.....	275
Chart Format Axis.....	283
Chart Format Chart .....	279
Chart Format Chart Area .....	282
Chart Format Font Editor.....	281
Chart Format Format Editor .....	282
Chart Format Legend.....	286
Chart Format Plot.....	286
Chart Format Series.....	280
Chart Format Title.....	285
Chart Select Data .....	277
Cluster Assigner.....	201
Clustering .....	199
Cohen's kappa.....	176
Color Manager .....	137

column .....	71, 92, 214
Column.....	162
Column Combiner.....	107
Column Filter .....	71, 74
Column Resorter.....	109
Combine.....	156
comments.....	58
Community .....	17
Concatenate .....	160
configure.....	58
Confusion Matrix.....	174
Connector .....	116, 117, 120
courses .....	18
CSV Reader .....	62, 63
CSV Writer.....	81

## D

Data.....	47, 68
Data Models.....	169
<b>Data Sets</b> .....	262
Data To Report.....	257
data visualization .....	132
Database Connector.....	116, 117, 120
Database Driver .....	121
Database Reader .....	124, 125
Decision Tree .....	178
<b>Decision Tree Learner</b> .....	180
<b>Decision Tree Predictor</b> .....	181
<b>Decision Tree View</b> .....	186
delete workflow .....	56
Double To Int .....	113
Dynamic Text.....	291

## E

events .....	18
EXAMPLES server .....	40
execute .....	59
extensions .....	45

## F

file .....	61, 80, 192
final document .....	290
F-measure .....	176

## G

graphical properties .....	136
GroupBy .....	128, 129

## H

Histogram .....	144
Histograms .....	143
hotkeys .....	37
Hypothesis Testing .....	202

## I

install .....	23
---------------	----

## J

Java Snippet .....	222, 239
Java Snippet (simple) .....	221
Javascript .....	132
JavaScript Functions .....	291
Joiner .....	216
Joiner Settings .....	217

## K

k-Means .....	200
---------------	-----

knar file type .....

knime

    protocol .....

KNIME Community .....

KNIME Explorer .....

KNIME Extensions .....

KNIME Hub .....

KNIME Public Server .....

knime: .....

knwf file type .....

## L

launcher .....

Line Plot .....

**Linear Regression (Learner)** .....

## M

Math Formula .....

Math Formula (Multi Column) .....

Meta-node .....

Meta-node context menu .....

Misc .....

Missing Value .....

**Model Reader** .....

**Model Writer** .....

**Multilayer Perceptron Predictor** .....

## N

Naïve Bayes .....

**Naïve Bayes Predictor** .....

Neural Network .....

new node .....

new workflow .....

new workflow group .....

node .....

Node Monitor .....

Node Repository .....

Normalization Methods .....

Normalizer.....	166
Normalizer (Apply).....	167
Number To String.....	111
Numeric Binner .....	127

## P

Page Break.....	275
Parallel Coordinates.....	139, 142
Partitioning.....	158
Pivoting.....	130
PMML .....	163
<b>PMML Reader</b> .....	194
<b>PMML Writer</b> .....	192
Precision.....	176

## R

Recall.....	175
RegEx Split .....	101
Regression .....	197
<b>Regression (Predictor)</b> .....	199
Rename .....	93
reporting.....	237
resources.....	17
ROC Curve .....	187
row.....	75
Row .....	208
Row Filter.....	76
Row Filter criteria.....	78
Row Sampling .....	157
RowID .....	211
<b>RProp MLP Learner</b> .....	189
Rule Engine.....	96

## S

save workflow .....	56
scatter plot .....	133

Scorer.....	173
Sensitivity.....	175
Shuffle.....	159
Sorter.....	214
Specificity .....	175
split.....	98
Split .....	156
Statistics .....	195
string manipulation .....	102, 103
String Replacer .....	106
String To Number .....	112
Style Sheets .....	270, 287, 288, 289

## T

Table View .....	147
Tables .....	266
Title .....	263, 290
Top Menu.....	32
type conversion .....	110

## U

Unpivoting .....	212
------------------	-----

## V

view .....	132
views properties .....	136
visualization.....	132

## W

workbench .....	28, 30
workflow .....	26, 53
Workflow Annotations.....	44
Workflow Credentials .....	118
Workflow Editor .....	42, 44
workspace .....	24



## ***KNIME Beginner's Luck: A Guide to KNIME Analytics Platform for Beginners***

*This book is born from a series of lectures. It gives a quite detailed overview of the main tools and philosophy of KNIME Analytics Platform. The goal is to empower new KNIME users with the necessary knowledge to start analyzing, manipulating, and reporting even complex data. No previous knowledge is required. The book has been updated for KNIME 4.1. The book shows:*

- *how to move inside (and install) KNIME Analytics Platform (Chapter 1);*
- *how to transform data and build a workflow (Chapter 2);*
- *how to perform a visual data exploration (Chapter 3);*
- *how to build models from data (Chapter 4);*
- *how to prepare data for a report (Chapter 5);*
- *how to build a report (Chapter 6 ).*

### ***About the Authors***

*Dr Satoru Hayasaka was trained in statistical analysis of various types of biomedical data. Since his doctoral training, he has taught several courses on data analysis geared toward non-experts and beginners. In recent years, he taught introductory machine learning courses to graduate students from different disciplines. Recently he joined KNIME as part of the evangelism team, and he continues teaching machine learning and data mining using KNIME Analytics Platform.*

*Dr Rosaria Silipo has been mining data since her master's degree in 1992. She kept mining data throughout all her doctoral program, her postdoctoral program, and most of her following job positions. She has many years of experience in data analysis, reporting, business intelligence, training, and writing. In the last few years she has been using KNIME for all her data science work, becoming a KNIME trainer and evangelist.*