

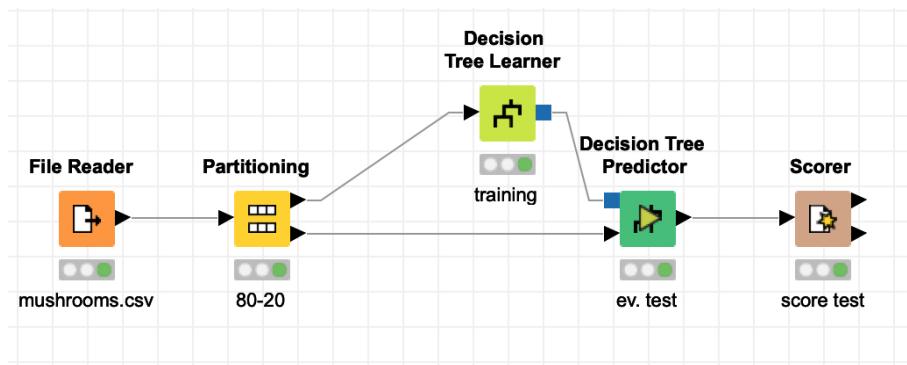


## 1. Objetivo

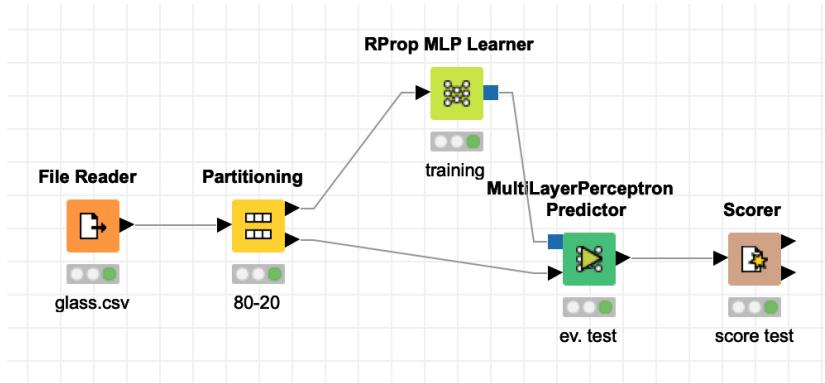
En esta práctica veremos el uso de algoritmos de aprendizaje supervisado de clasificación como herramienta para realizar análisis predictivo. Concretamente, se trabajará con un conjunto de datos reales sobre el que se emplearán diferentes algoritmos de clasificación (para su comparación) y a la luz del conocimiento descubierto se podrán concluir estrategias para resolver el problema.

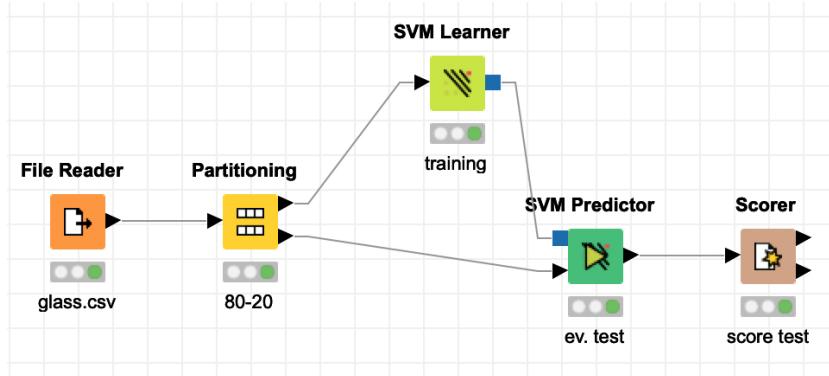
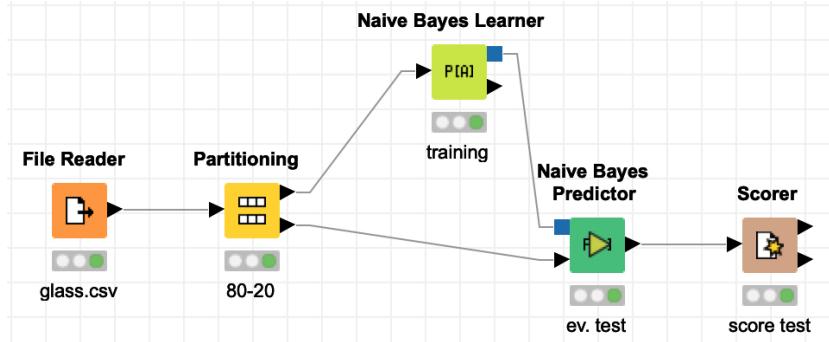
## 2. Validación cruzada

En prácticas anteriores describíamos cómo realizar una partición del conjunto en dos partes para realizar el proceso de aprendizaje del modelo con uno de ellos y la evaluación del modelo con el resto (hold-out). Así, nuestro flujo de datos podría tener un aspecto parecido a siguiente, donde se ha realizado una validación del modelo entrenado (árbol de decisión).

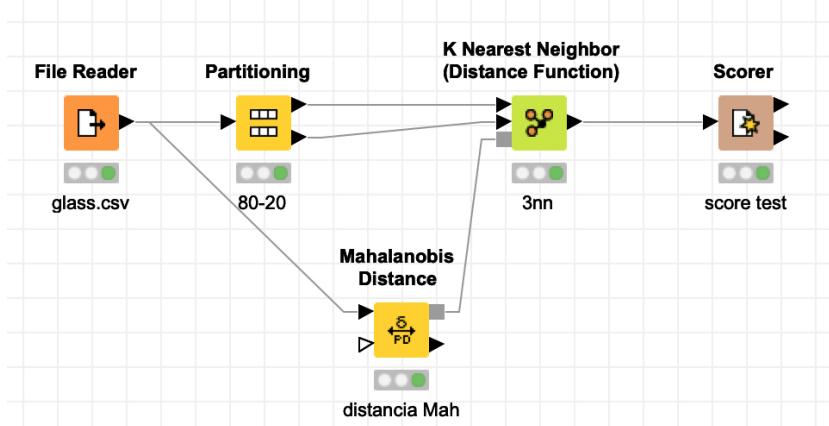
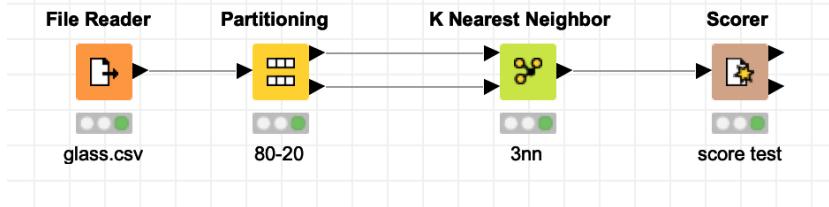


En general, los modelos de clasificación siguen este esquema. Necesitamos un nodo con el que entrenar el modelo (learner) que devuelve como salida el modelo entrenado. Este, junto con los datos de validación, es la entrada del nodo encargado de clasificar dichos datos (predictor). Así, para otros clasificadores el flujo quedaría:



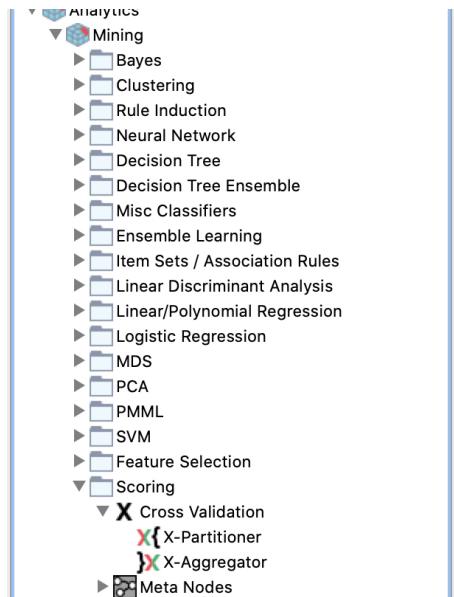


Simplemente hay que tener cuidado con los tipos de datos de los atributos, puesto que algunos nodos de entrenamiento no aceptan cualquier tipo de dato (por ejemplo, el entrenamiento con máquinas de vector soporte no acepta datos nominales). La única excepción a este esquema son los clasificadores perezosos (lazy), como kNN. En este caso, el flujo quedaría como

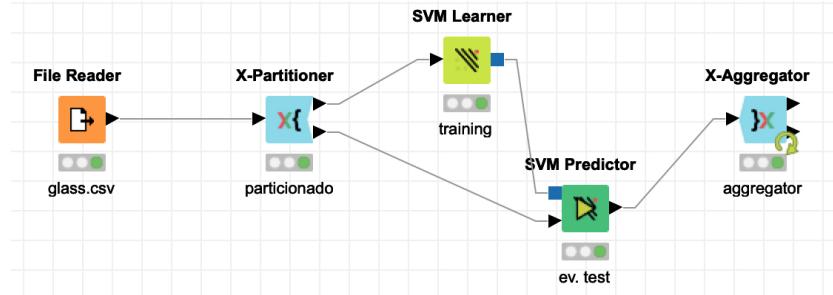


Veamos ahora cómo realizar una validación cruzada. Para ello, utilizaremos los nodos en

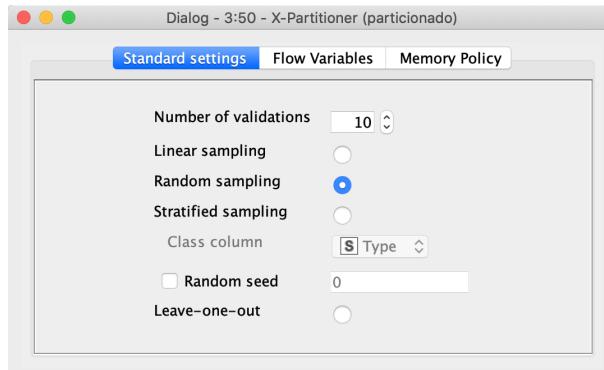
## Analytics/Mining/Scoring/X Cross Validation sobre validación cruzada



La disposición básica de los nodos es la siguiente:



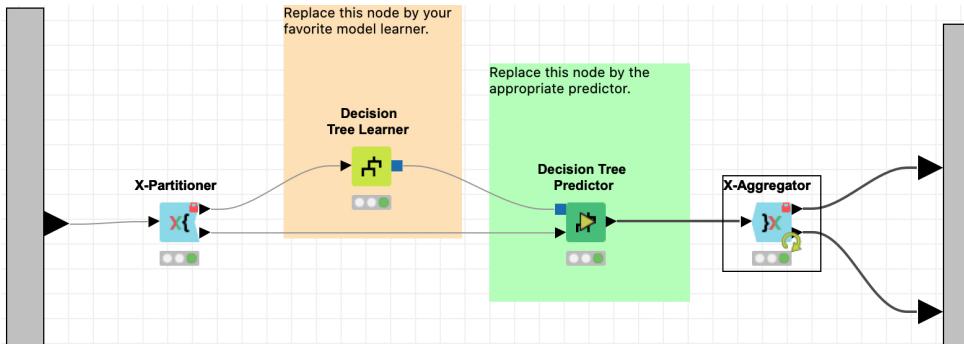
donde, en el nodo **X-partitioner** podemos configurar el número de particiones y la forma de seleccionar las particiones (incluso realizar una validación leave-one-out).



El funcionamiento es sencillo, en el proceso de aprendizaje se realiza el número de particiones que hayamos seleccionado en el nodo **X-partitioner** y los resultados de los tests se acumulan en el nodo **X-Aggregator**. En el nodo **learner** seleccionado para la validación cruzada se van sobreescribiendo los distintos modelos de las particiones y al final, lo que queda almacenado es sólo el modelo de la última iteración.

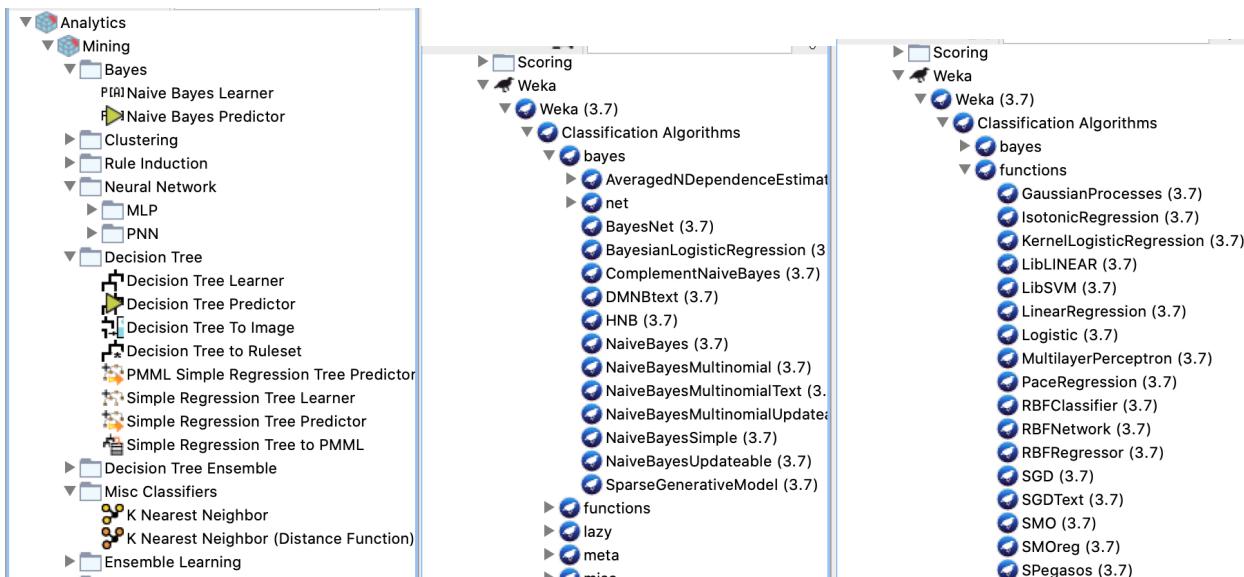
Knime también contiene por defecto un metanodo para añadir la validación cruzada. Para ello, se debe añadir un nodo de tipo **Analytics/Mining/Scoring/X Cross Validation/Meta Nodes -> Cross Validation** al proyecto. Tras añadirlo, hacer doble click sobre él y se abrirá un cuadro en el que se podrá definir el flujo interno

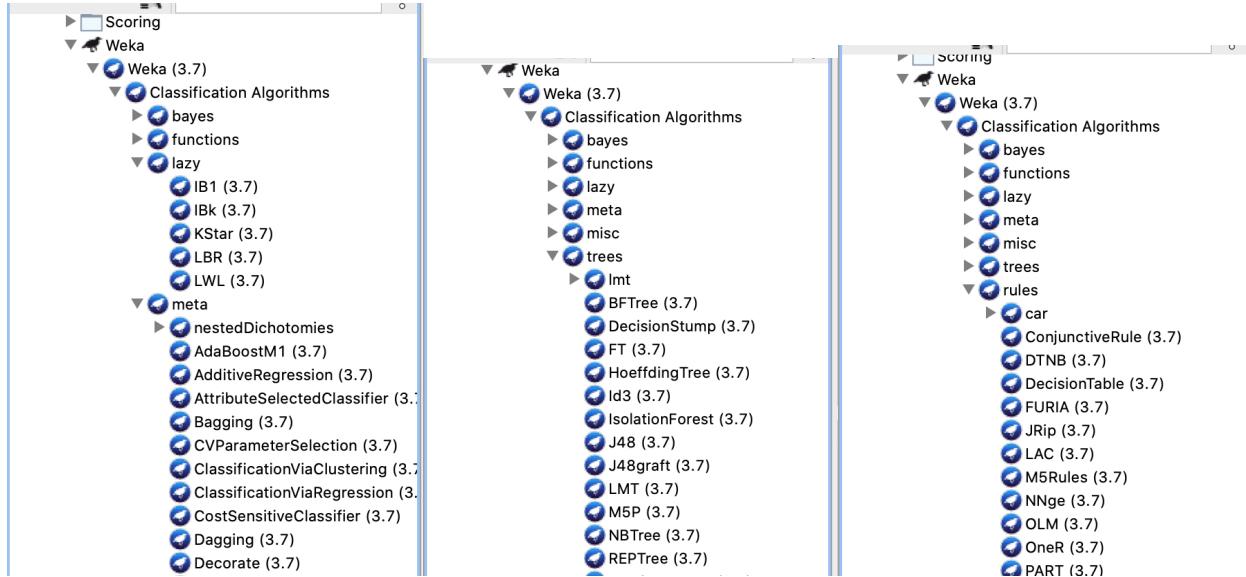
de este metanodo. Por ejemplo, tras cambiar los nodos internos, podría quedar



### 3. Sobre los modelos de clasificación

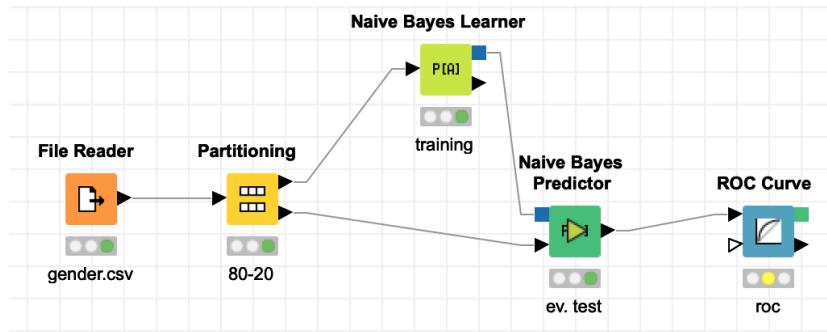
Knime y sus extensiones contienen nodos para aprender los modelos más usuales de clasificadores. Excepto para el sistema de reglas, en Analytics/Mining se encuentra, al menos, un nodo para cada uno de estos modelos. En los nodos de Weka se tiene una amplia colección de algoritmos para aprender todos los modelos.



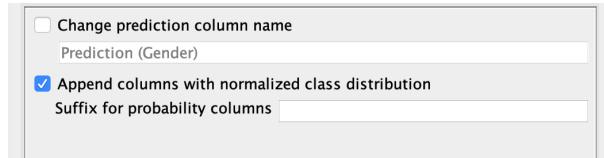


## 4. Evaluación

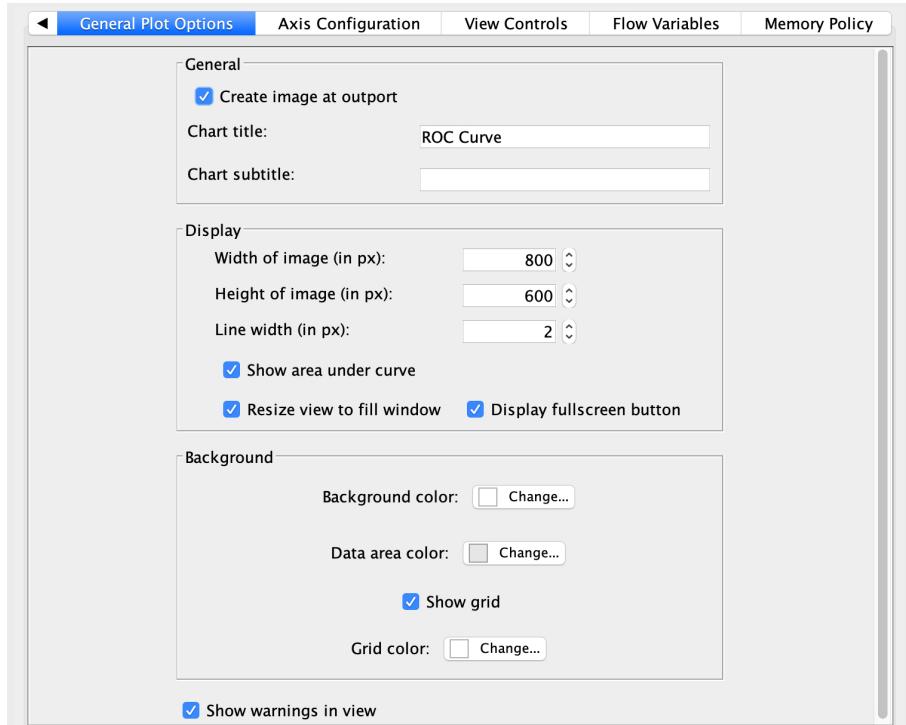
A la hora de evaluar los modelos podemos utilizar, como se muestra en los gráficos anteriores, el nodo **scorer** que calcula la matriz de confusión y las medidas de evaluación usuales (recall, precision, medida F, etc). También podemos calcular la curva ROC utilizando el nodo **ROC Curve** en **Views/JavaScript**.



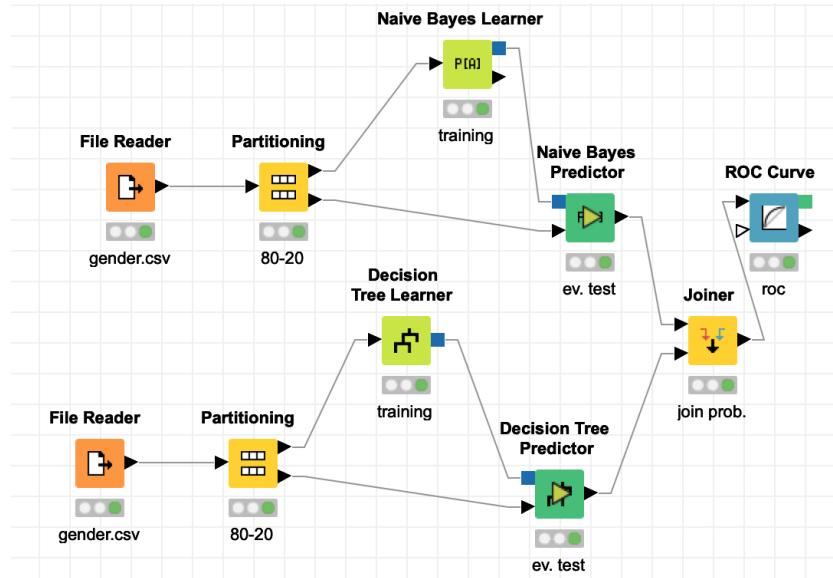
Se debe utilizar para clasificación binaria, y necesitamos calcular la probabilidad de ser un caso positivo para cada instancia del conjunto de test. Esto último lo puede calcular el propio nodo **predictor** marcando la opción en su configuración

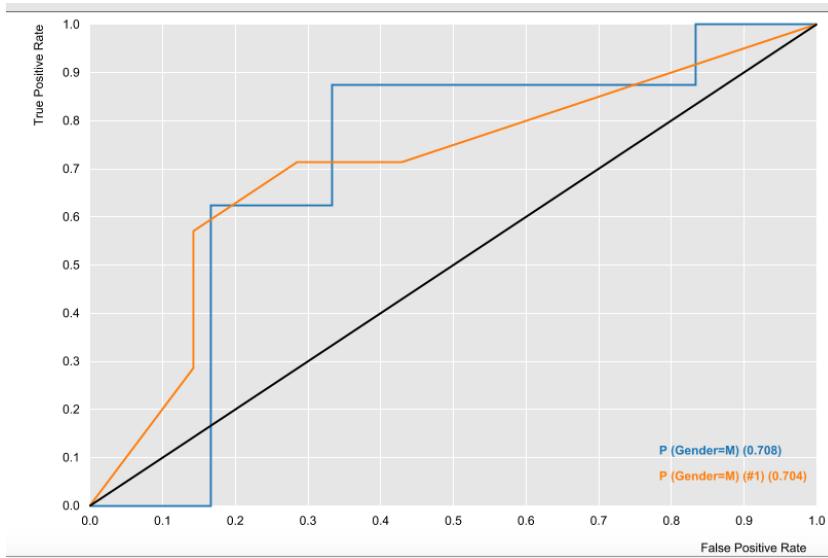


Si queremos obtener la imagen de la curva, debemos especificarlo en la configuración del nodo ROC Curve

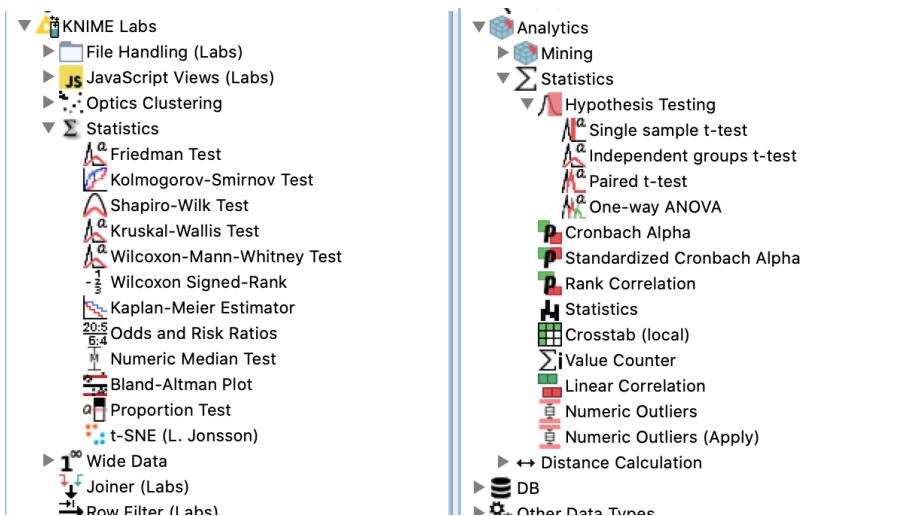


Para dibujar varias curvas ROC, debemos unir los datos de salida de las evaluaciones.

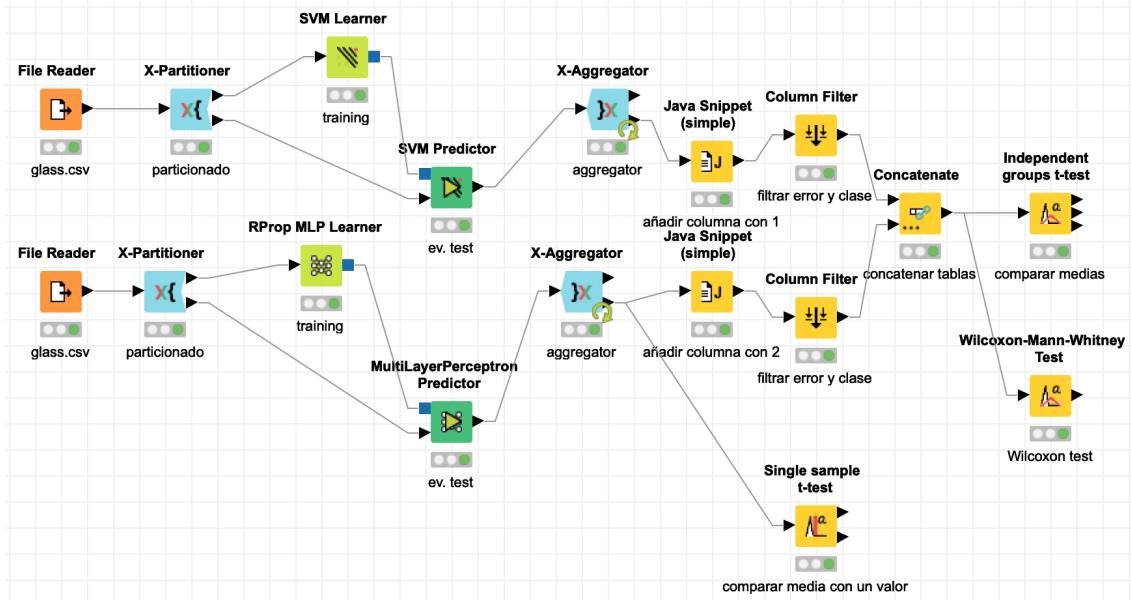




Por último, Knime también incorpora nodos para la comparación estadística (paramétrica y no paramétrica) en las carpetas **Analytics/Statistics/Hypothesis Testing** y **KNIME Labs/Statistics**.



Por ejemplo, podemos comparar los resultados de dos validaciones cruzadas utilizando el nodo **independent groups t-test** y el nodo **Wilcoxon-Mann-Whitney Test**.



## 5. Descripción del problema: subastas competitivas en eBay

La multinacional eBay Inc. (propietaria de la conocida web eBay.com para subasta de productos a través de internet) desea conocer mejor el comportamiento de las transacciones producidas en su web para, a la vista de los resultados, diseñar nuevos servicios que mejoren la experiencia de los usuarios vendedores para así incrementar las ventas y, por tanto, mejorar los ingresos de la compañía.

Así, se propone aplicar analítica empresarial de cara a extraer conocimiento útil para la toma de decisiones a partir de algunos datos disponibles. Concretamente, el objetivo es construir un modelo que clasifique las subastas entre competitivas y no competitivas. Una subasta se considera competitiva si recibe al menos dos ofertas sobre el objeto subastado. Los datos disponibles corresponden a 1972 subastas a través de eBay.com realizadas en los últimos dos meses. Se incluyen variables que describen el objeto (categoría), el vendedor (mediante su valoración o rating) y los términos de la subasta fijados por el vendedor (duración de la subasta, precio de inicio, moneda y día de la semana en el que finaliza la subasta). Además, disponemos del precio al cual se cerró la subasta. Los datos están contenidos en el fichero Excel **eBayAuctions**. Se pretende predecir si la subasta será o no competitiva así como comprender qué relaciones provocan dicho factor de cara a diseñar estrategias de negocio para que los clientes vendedores aumenten la tasa de subastas competitivas en sus productos.

Considerar varios modelos de clasificación distintos (con distintos parámetros cada uno) y entrenar los modelos. Para estos modelos realizar una evaluación mediante validación cruzada (si son mucho modelos los que se entrena, utilizar 5 particiones, por ejemplo). Para sustentar el análisis comparativo emplear tablas de errores, matrices de confusión y curvas ROC. Previamente, realizar un preprocesamiento adecuado de los datos. Basado en los resultados obtenidos, ¿qué se recomendaría a un vendedor para hacer que sus subastas tengan más probabilidad de ser competitivas? En función del conocimiento obtenido, ¿qué estrategias de negocio podría adoptar la empresa eBay para mejorar el resultado de las subastas?