Tarea 21 de abril: Árboles de decisión.

Para realizar esta tarea necesitaremos el programa Knime.

Comenzamos abriéndolo, y en la página principal que nos aparece seleccionamos File→ New File.

Dejamos la opción por defecto y hacemos click en Next y luego Finish. De esta manera habremos creado nuestro espacio de trabajo.

Para esta tarea necesitaremos crear e interconectar 6 nodos. Cada nodo tiene una aplicación concreta:

Nodo 1- Lectura de datos.

Nodo 2- Asignación de colores.

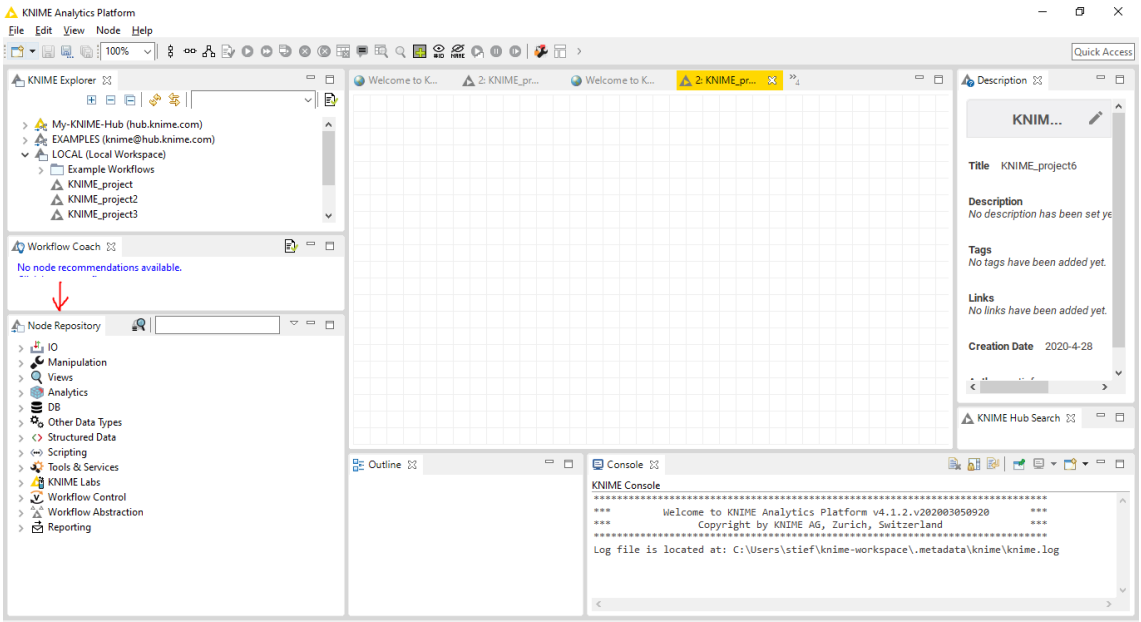
Nodo 3- Subdivisión de datos.

Nodo 4- Generación árbol de decisión.

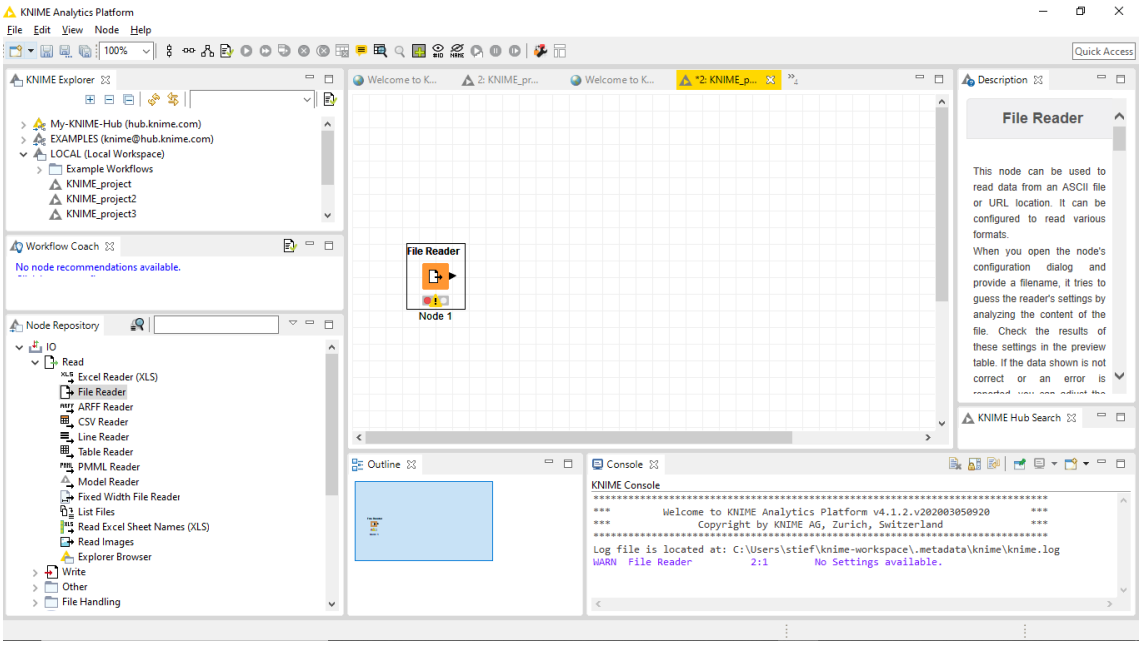
Nodo 5- Testing del modelo.

Nodo 6- Matriz de confusión.

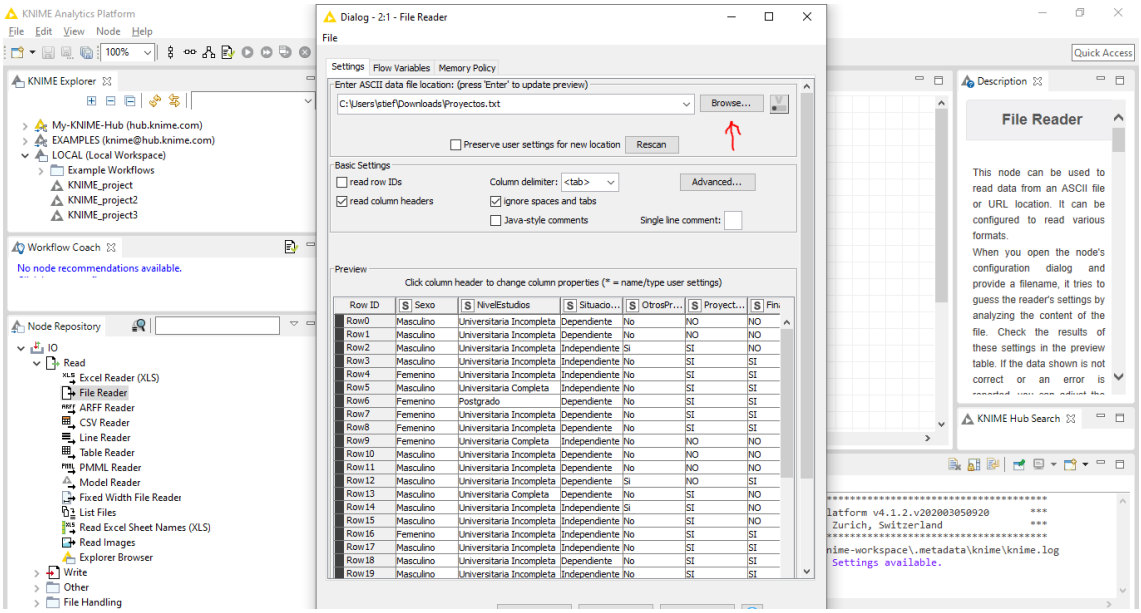
Para ubicar los nodos necesarios en el espacio de trabajo, iremos a Node Repository.



Una vez aquí podemos buscar en el buscador el nodo que queremos, o podemos buscar manualmente. El primer nodo que usaremos para leer el archivo Proyectos.txt se llama File Reader y lo podemos encontrar en la carpeta I/0.



Como ya vimos en la tarea anterior, ahora es el momento de configurar el nodo. Esto consiste, en este caso, en hacer click con el botón derecho en el nodo y elegir configurar. En la nueva pantalla que aparecerá navegamos por las carpetas hasta encontrar el archivo que queremos leer.



Una vez seleccionado el archivo a leer, pulsamos ok. De esta manera volvemos a nuestro espacio de trabajo, observando que el semáforo del archivo está en naranja ahora.

Para ponerlo en verde simplemente deberemos hacer click derecho en el nodo y pulsar ejecutar.

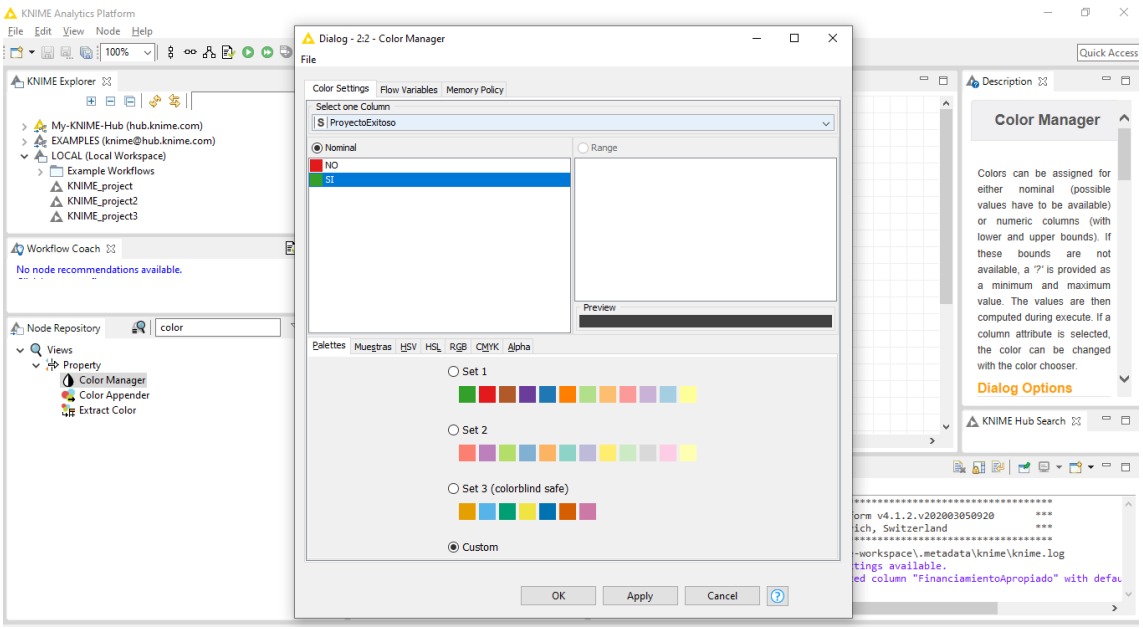
Ya tenemos listo en nodo de lectura del archivo.

El próximo nodo que usaremos es el nodo Color Manager, que nos permitirá seleccionar los colores que queremos que aparezcan. Por ejemplo, si queremos saber si el proyecto fue exitoso o no, podemos hacer que SI salga de color azul, y NO salga de color rojo.

Buscamos en Node Repository Color Manager y lo metemos en nuestro espacio de trabajo.

Conectamos ambos nodos.

Para configurarlo hacemos click con el botón derecho en el nodo y seleccionamos Configuration. En esta ventana, podemos elegir el atributo que queremos tener como explicado, en nuestro caso ProyectoExitoso, y también el color que queramos que tenga cada una de las posibilidades del atributo ProyectoExitoso. En nuestro caso hemos decidido que los colores sean rojo para el NO y verde para el SI.



Tras realizar estos cambios, pulsamos ok y ejecutamos(F6) el nodo.

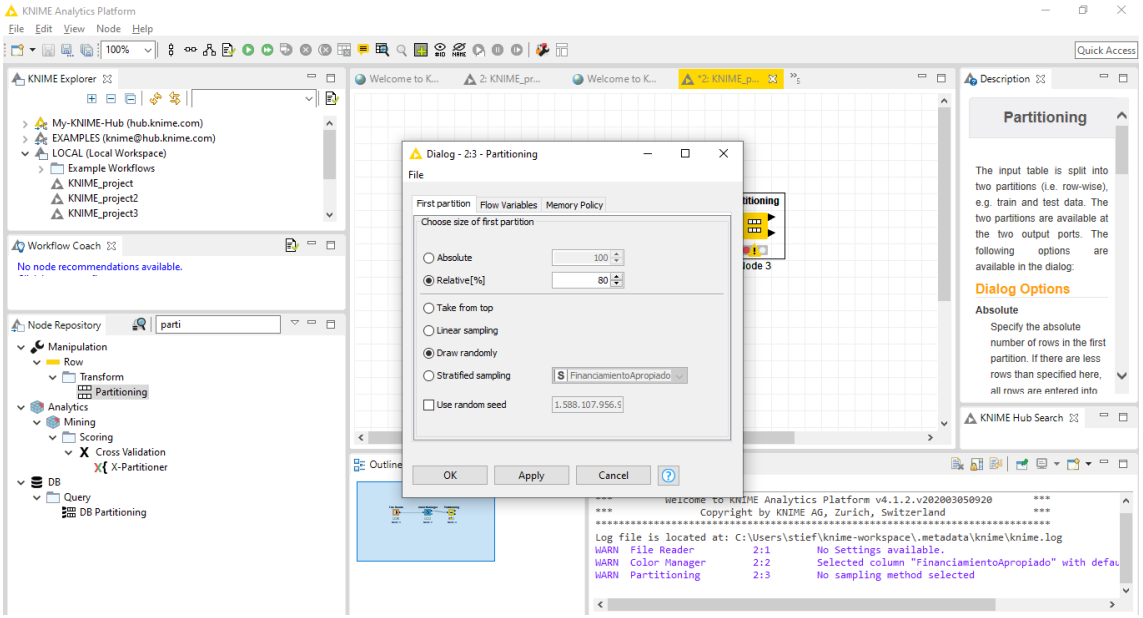
El siguiente paso a dar es dividir los datos, ya que al usar clasificación, usaremos partes de estos datos para realizar un modelo, y el resto lo usaremos para testear el modelo y comprobar que funcione correctamente. Esto lo hacemos porque si no podríamos caer en el sobreajuste. El sobreajuste se da cuando el modelo se adapta perfectamente a los datos sobre los que ha sido creado, pero si aparecen nuevos datos, el modelo no se ajusta.

Nuestro modelo debe ajustarse a la posibilidad de que haya nuevos datos, es por eso que dividimos los datos para crear un modelo y luego testearlo.

El nodo que usaremos para la partición de los datos se llama Partitioning y lo buscamos tal y como hicimos con el nodo Color Manager.

Conectamos el nodo Color Manager con el nodo Partitioning y procedemos a la configuración de este ultimo.

Para configurarlo, con el nodo seleccionado, pulsamos F6 para entrar en el panel de configuración. En esta pantalla elegiremos la opción relativo e indicaremos que queremos dividir los datos en el 80% para creación del modelo, y el 20% restante lo usaremos para testear el modelo y comprobar que funciona correctamente.



El resto de opciones las dejamos por defecto y pulsamos ok.

A la salida del nodo obtendremos las dos partes de datos. La salida de arriba del nodo comprende el 80% de los datos, que serán los que usaremos para crear el modelo, y la salida de abajo del nodo comprende el 20% restante de los datos.

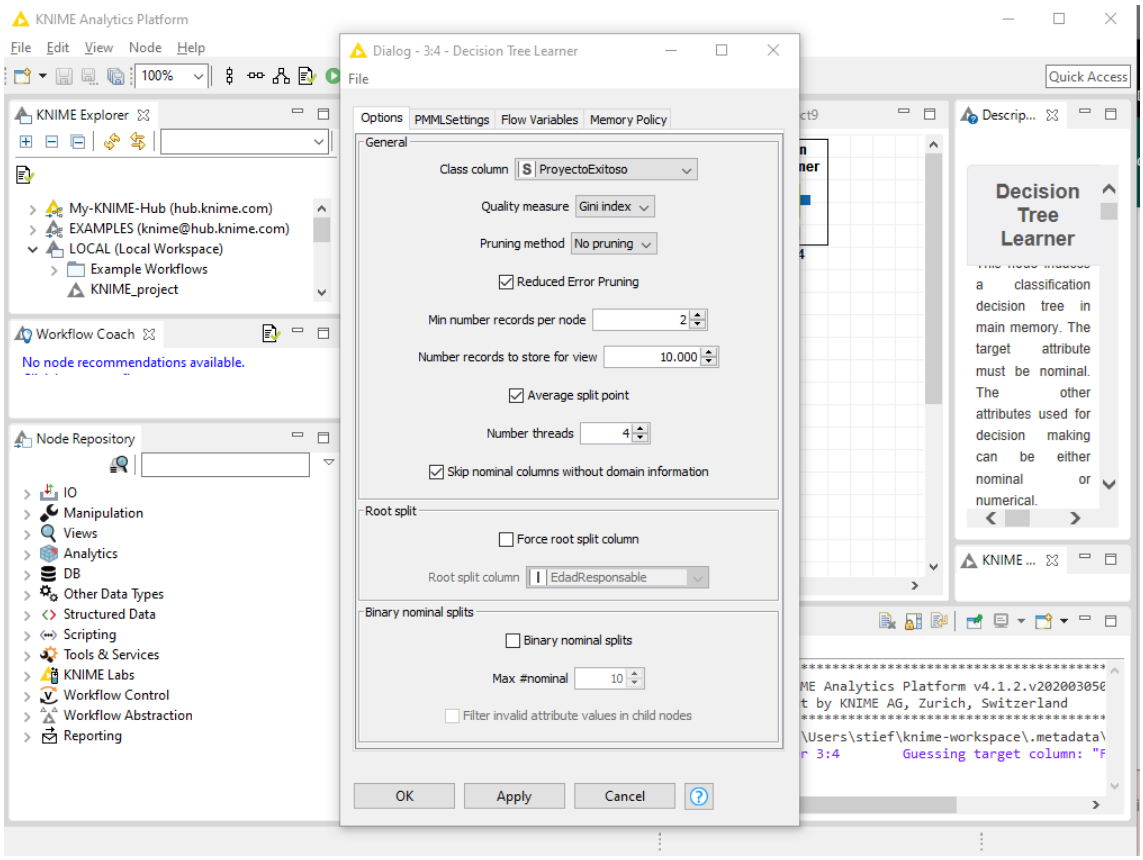
Con el nodo seleccionado, pulsamos F7 y de esta manera lo ejecutamos.

El próximo nodo que usaremos es el nodo Decision Tree Learner, por lo que lo buscamos dónde hemos buscado el resto de nodos y lo llevamos a nuestro espacio de trabajo.

Debemos unir la salida superior del nodo Partitioning con la entrada del nodo Decision Tree Learner. De esta manera estaremos trabajando con el 80% de los datos a partir de los cuales vamos a crear el modelo.

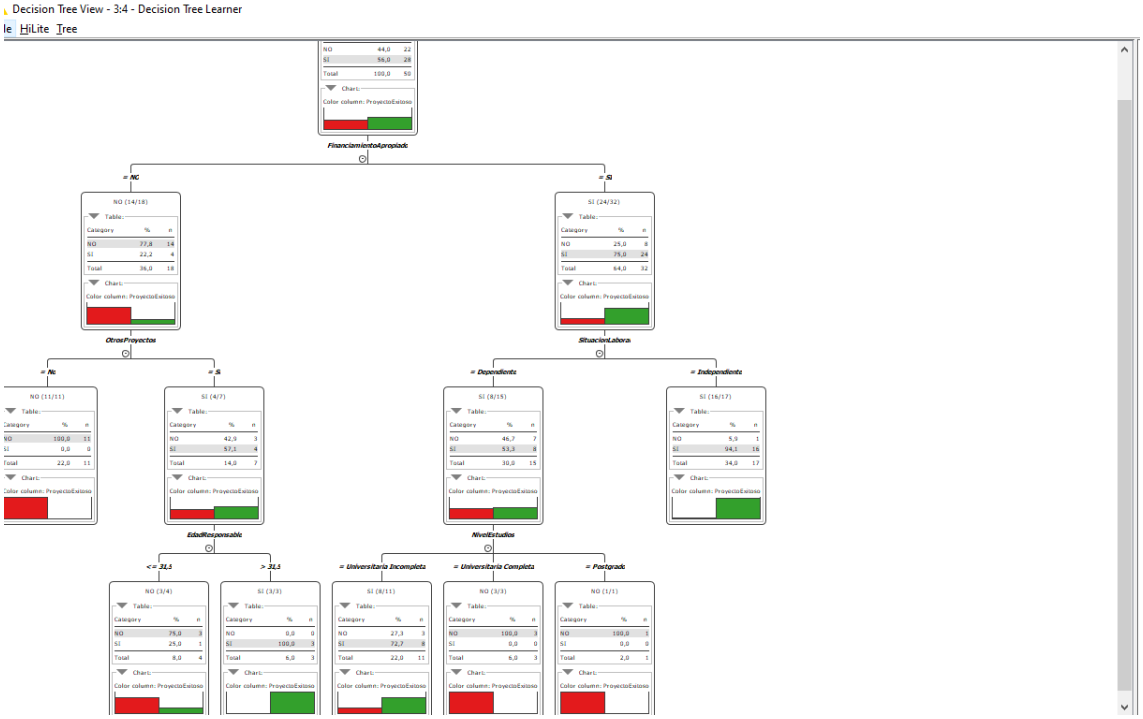
Con el nodo Decision Tree Learnes seleccionado, pulsamos F6 para configurarlo.

En esta ventana solo tenemos que modificar el Class Column y poner el atributo que queremos clasificar. En nuestro caso será ProyectoExitoso.



Tras haberlo configurado, pulsamos ok y ,si no teníamos ya conectada la salida superior del nodo Partitioning con la entrada del nodo de generación de árbol de decisión, lo conectamos y ejecutamos.

Haciendo click derecho en el nodo del árbol de decisión, podremos observar una opción que se llama View Decision Tree Learner. La pulsamos y llegamos a esta pantalla.



En nuestro árbol, podemos observar que la primera apertura, el atributo que más discrimina, es el FinanciamientoApropiado.

Podemos decir que en nuestro caso el atributo más discriminante es FinanciamientoApropiado, pero solo podemos decirlo para nuestro caso, ya que cuando hicimos la partición de los datos con el nodo Partitioning, dejamos marcada una opción dentro del menú de configuración que decía: Draw randomly.

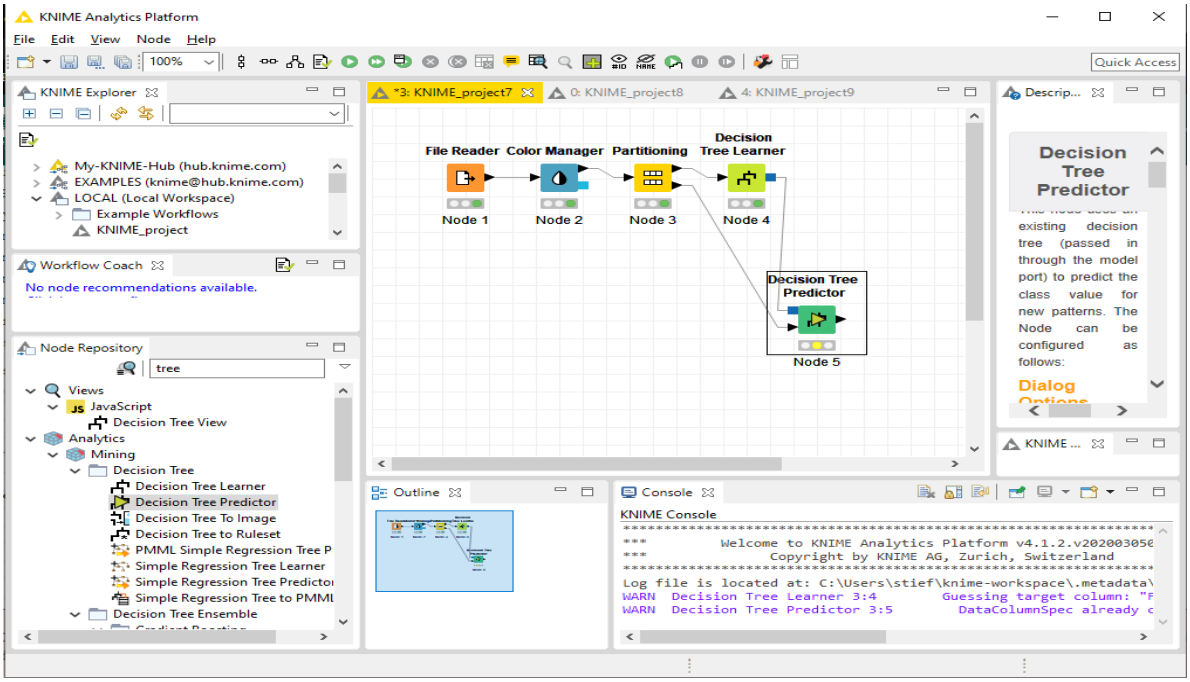
Esto quiere decir que al coger el 80% de los datos, lo hizo de forma aleatoria. Por poner un ejemplo podríamos decir que tenemos 10 datos. Solo queremos el 80% de ellos para crear el árbol de decisión, por lo que cogemos los 8 primeros datos. Coger los 8 primeros datos nos dará un árbol de decisión que podrá ser igual o no al árbol generado si decidimos coger los 8 últimos datos, o los datos 1, 2, 3 ,4, 7, 8, 9,10 y hubiésemos dejado los datos 5 y 6 para realizar el testeo.

Los arboles de decisiones son muy útiles, ya que si por ejemplo en nuestro caso, quisiera saber las condiciones que hicieron que un proyecto no fuese exitoso, me voy a la rama que tenga como consecuente proyecto no exitoso y la analizo. Nosotros vamos a analizar por ejemplo la rama de la izquierda.

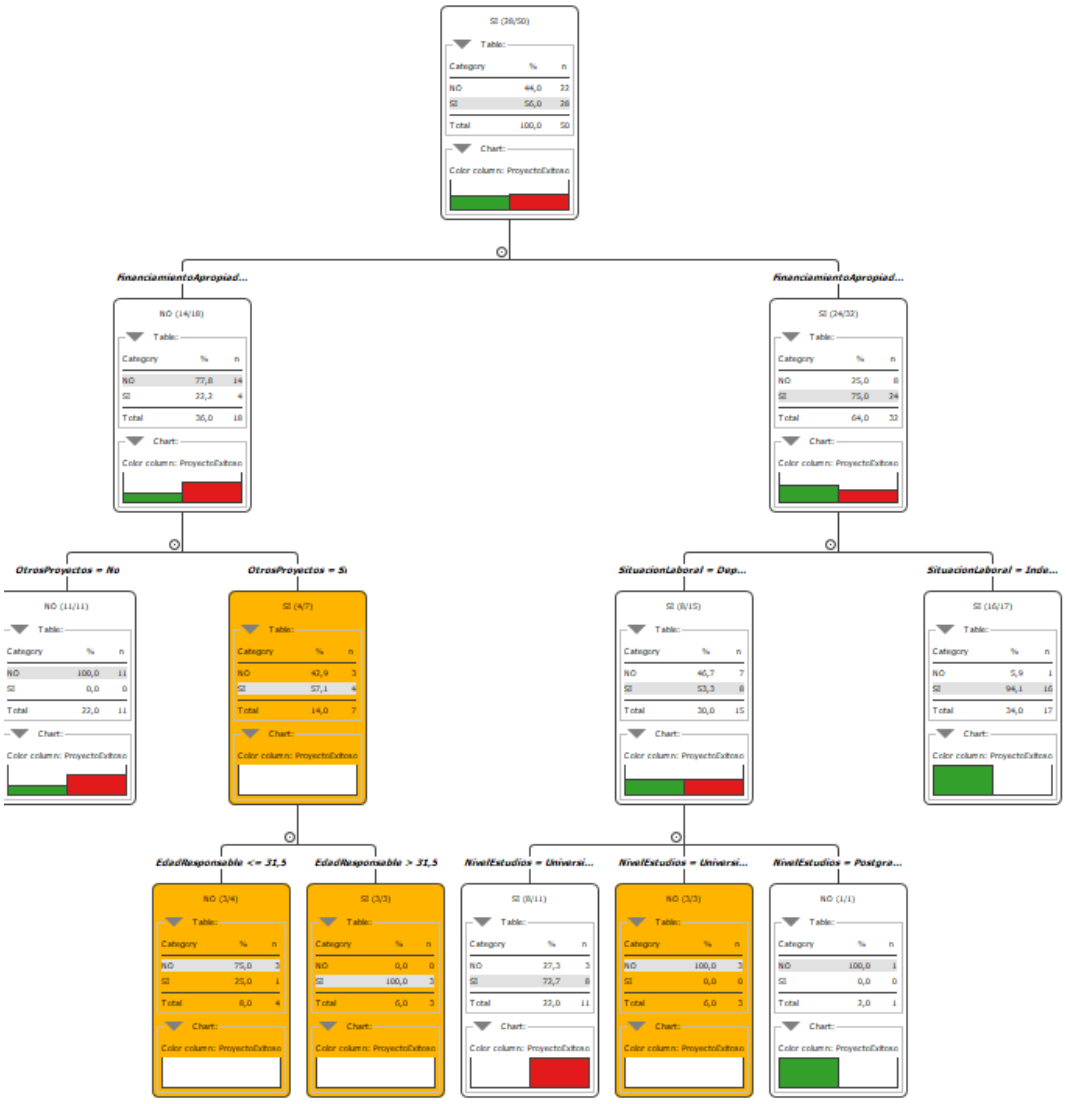
Podemos observar que si el financiamiento no fue apropiado, y no tenia proyectos previos, el consecuente es que en el 100% de los casos el proyecto no fue exitoso.

Una vez realizado esto, procederemos a crear un nodo Decision Tree Predictor. Este nodo se encargará de testear el modelo con el 20% restante de los datos que no usamos para crear el modelo. De esta manera podemos comprobar la eficacia del modelo que hemos creado.

Primero buscaremos el nodo en Node Repository y lo incluiremos al espacio de trabajo. Conectamos la salida del Decision Tree Learner a la entrada superior del nodo, y a la entrada inferior le conectaremos la salida del nodo Partitioning, la que contiene el 20% de datos que usaremos para el testeo del modelo.



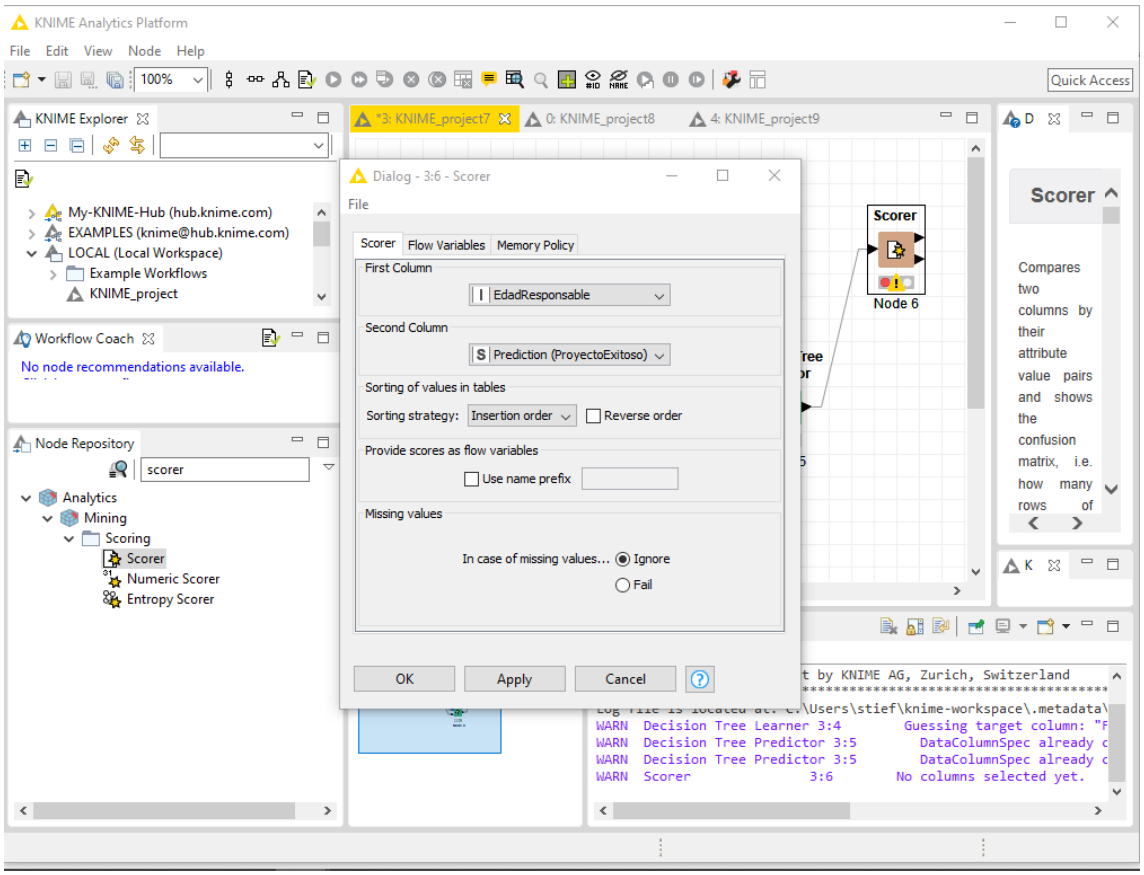
Una vez conectado, ejecutamos el nodo y abrimos el árbol.



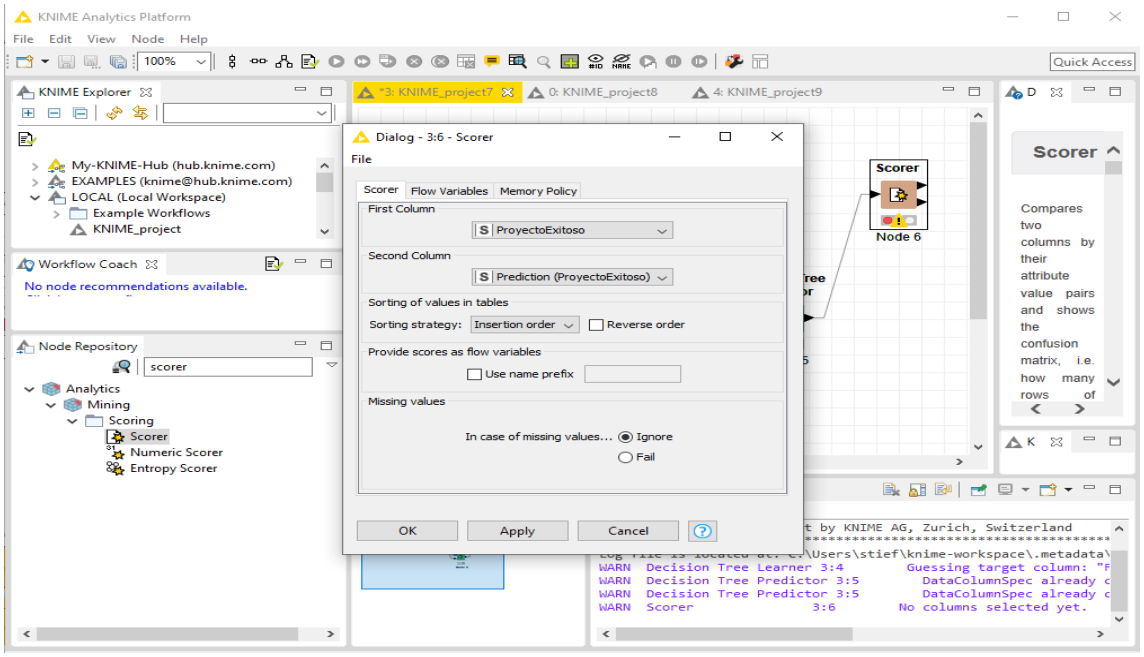
Este es el árbol de decisión de los datos que hemos usado para armar este caso. Pero esto no es lo que nos interesa, lo que realmente nos interesa es saber cuantos

Para ello usamos el nodo Scorer. Este nodo estará a la salida del Decision Tree Prediction. Este nodo nos generará una matriz de confusión. Buscamos en el Node Repository el nodo Scorer y lo incorporamos al espacio de trabajo. Lo seleccionamos, y pulsamos F6 para configurarlo.

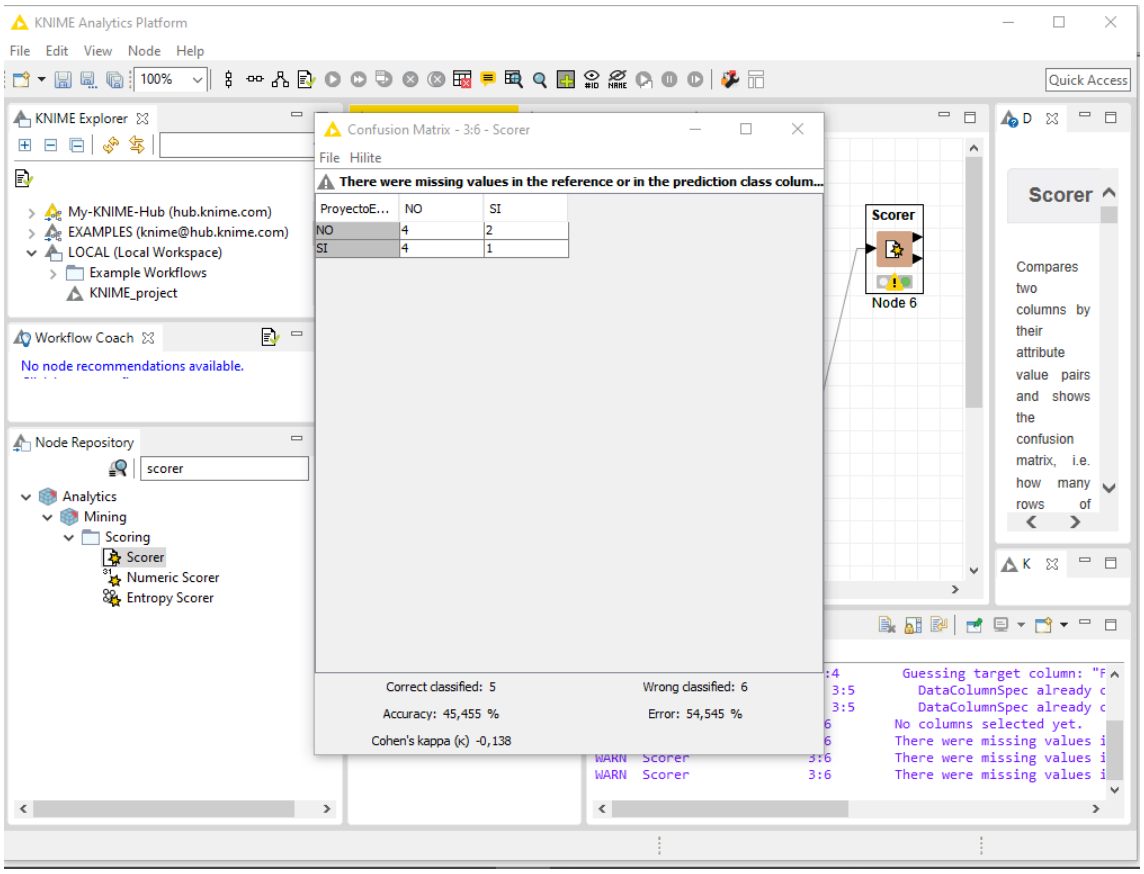
La pantalla que nos debe aparecer es la siguiente:



En este menú elegiremos que queremos comparar. En nuestro caso queremos comparar los datos que aparecen sobre si el proyecto fue exitoso con la predicción que genera el modelo sobre si el proyecto fue exitoso. Por tanto el nodo configurado quedaría de la siguiente manera:



Ahora lo aplicamos y ejecutamos el nodo pulsando F7. Luego pulsamos con el botón derecho el nodo y elegimos la opción Confusion Matrix. Esta matriz compara lo real con lo que predice el modelo. En nuestro caso:



Podemos observar que tenemos bien clasificado 5 (no, no →4, si, si→1) y tenemos 6 mal clasificado.