Primeros pasos con Knime

Luis P. Guerra Velasco

May 7, 2008

Índice

Capítulo 1

Introducción

El presente documento consiste en una guía de primeros pasos y utilización de la herramienta Knime (www.knime.org). Knime es software libre y su versión inicial fue lanzada en Julio de 2006, la versión actual es la 1.3.3, pero se encuentra en continuo desarrollo por los creadores de la misma en Konstanz (Alemania) y toda la comunidad que quiere participar.

Knime es una plataforma modular de exploración de datos, que permite al usuario crear flujos de datos, o pipelines, de forma visual e intuitiva. Además, permite ejecutar de forma selectiva algunos de los pasos creados, así como ejecutar todo el flujo desarrollado. Tras la ejecución, los resultados se pueden investigar mediante vistas interactivas tanto de los datos como de los modelos.

Este documento utiliza Knime como herramienta eminentemente destinada a realizar minería de datos. Para ello, se va a utilizar un dataset de ejemplo muy sencillo, el cual se irá modificando en un primer preprocesado de datos, para posteriormente utilizarlo tanto en clasificación supervisada como en no supervisada.

El dataset se encuentra en formato ARFF y está formado inicialmente por 5 atributos y 14 instancias, y está destinado a averiguar si es posible jugar un partido de tenis al aire libre con unas condiciones meteorológicas determinadas.

Capítulo 2

Instalación y extensiones

La instalación de Knime es muy sencilla tanto en sistema operativo Windows como Linux. En http://www.knime.org/quickstartguide.html existe una guia rápida de instalación de la aplicación, así como del funcionamiento básico de cada elemento del sistema.

Una de las características principales de Knime es su posibilidad de ampliación y de conexión con otras herramientas para poder utilizar las funcionalidades de las mismas. En el ejemplo se utilizará su conexión con Weka, lo cual se consigue de manera muy sencilla siguiendo las instrucciones que se pueden encontrar en http://www.knime.org/download_extensions.html.

Capítulo 3

Preprocesado y manejo de datos

Cuando se ejecuta por primera vez Knime, aparece en la pantalla el entorno vacio sobre el que se va a trabajar (ver imgen 3.1).

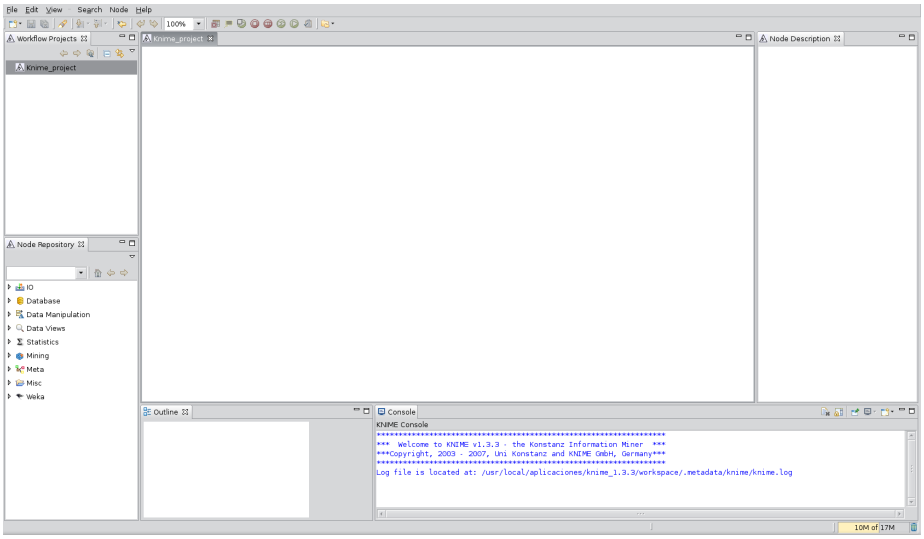

Figura 3.1: Pantalla de inicio

Lo primero que se necesita es obtener el conjunto de datos a partir del cual se va a realizar cualquier tarea. Para ello, hay que utilizar uno de los nodos de lectura existentes. En este caso, se elige el nodo de lectura de ficheros con formato ARFF, seleccionándolo en el repositorio de nodos y arrastrandolo al entorno de trabajo. Una vez allí, y seleccionando con el boton derecho del ratón, se abre la ventana de configuración donde se buscará la localización del fichero de datos en el disco duro, como se puede ver en la figura 3.2.

Una vez ejecutado el nodo de lectura, los datos estarán listos para poder utilizarse con la herramienta. Seleccionando con el boton derecho del ratón

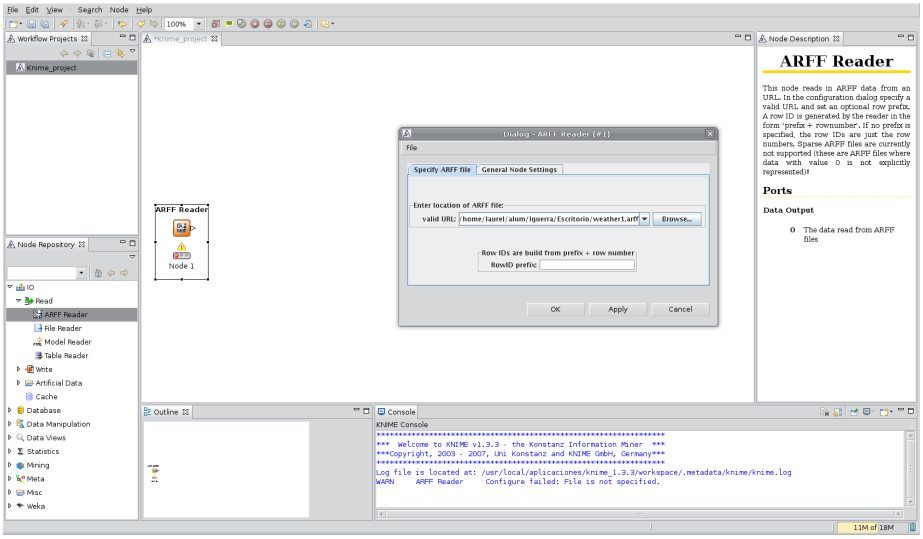

Figura 3.2: Nodo Reader

nuevamente en el nodo ejecutado, se pueden ver los datos cargados, así como las características de cada atributo (ver imgen 3.3).

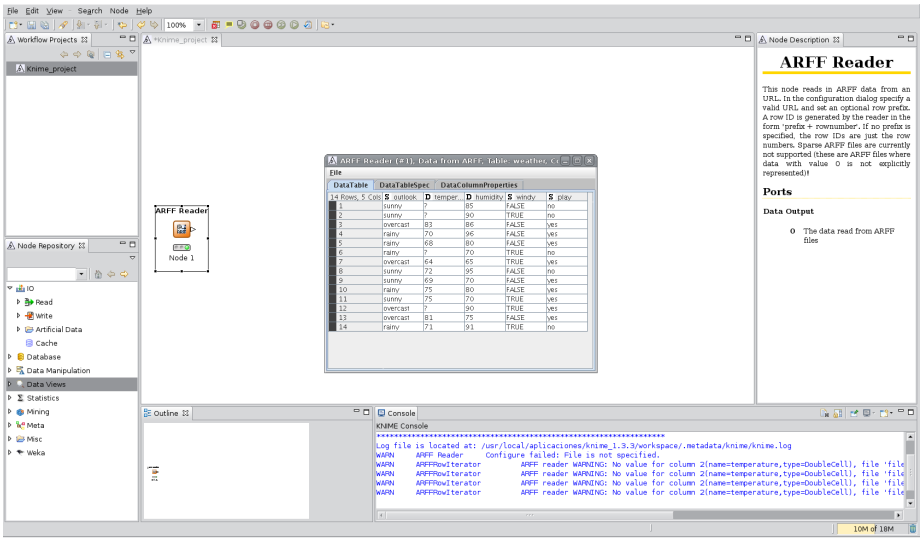

Figura 3.3: Vista de los datos iniciales

En la imágen 3.3 se puede apreciar como existen missing values en algunas instancias del atributo Temperature. Por lo tanto es necesario tratar esos datos de alguna manera para poder trabajar con el conjunto de datos. Para manejar los datos faltantes, existe un nodo llamado Missing Value, en la zona de manipulacion de datos del repositorio. Lo arrastramos al entorno de trabajo y unimos

su entrada con el nodo de lectura ya existente. Abriendo la configuración del nodo aparecen las diversas opciones para el manejo de los valores faltantes. En este caso, figura 3.4, se a nadirá el valor medio del resto de instancias en los valores que falten.

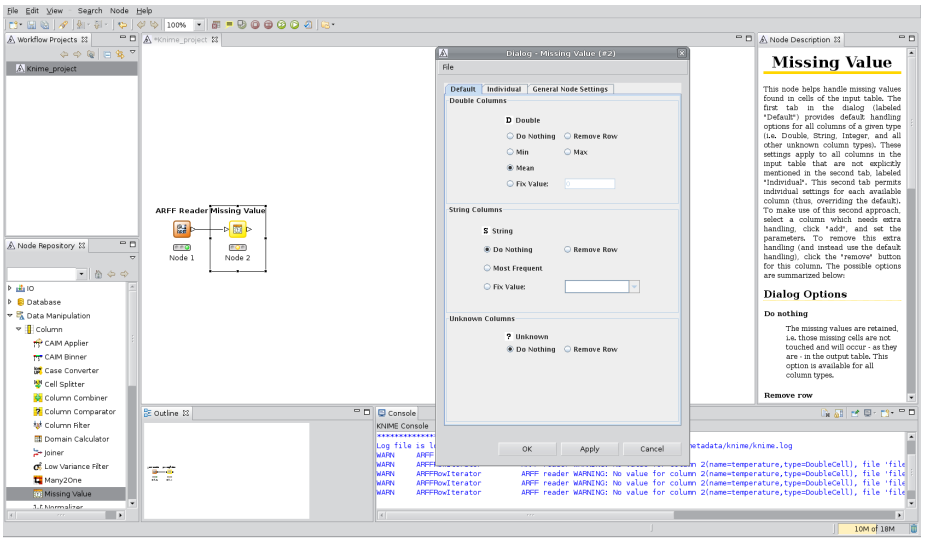

Figura 3.4: Nodo para solucionar missing values

Si se observa la salida del nodo de Missing Values, se ve como ya no hay instancias sin algunos valores, y que la media ha sido a nadido correctamente. Debido a que el tama no del conjunto de datos es bastante peque no, se va a ampliar el número de instancias del mismo con filas generadas por la herramienta. Para realizar esto se utiliza el nodo SMOTE que se encuentra de nuevo en la manipulación de datos del repositorio. Este nodo genera nuevos datos extrapolando un objeto real de una clase dada y uno de sus vecinos más cercanos, y escogiendo un punto aleatorio entre ambos objetos para generar el dato nuevo. Por lo tanto, se selecciona el nuevo nodo y se arrastra al entorno de trabajo, y nuevamente se une a la salida del último nodo utilizado para seguir con el flujo de datos creado. En las opciones del nodo seleccionamos el número de vecinos a utilizar para la creación, así como la clase que tiene que observar y el número de filas que se va a generar por cada una existente, figura 3.5.

Por último, para acabar con el preprocesado de los datos, se va a dividir el conjunto de datos obtenido en dos conjuntos diferentes. Así tendremos un conjunto de datos del cual se puede aprender un modelo (training set) y otro con el que validar el modelo obtenido (test set). Para ello se utiliza el nodo llamado Partitioning que se encuentra en el repositorio en Manipulación de datos. Dentro de las opciones de este nodo, haremos que la división se haga de forma aleatoria y que uno de los dos conjuntos tenga un 33% de los datos y el otro el resto, esto se puede ver en la figura 3.6.

Una vez obtenidos los dos conjuntos de datos, se puede continuar trabajando directamente con los datasets, o bien guardarlos en el disco duro en cualquiera de los formatos soportados por Knime utilizando un nodo Write.

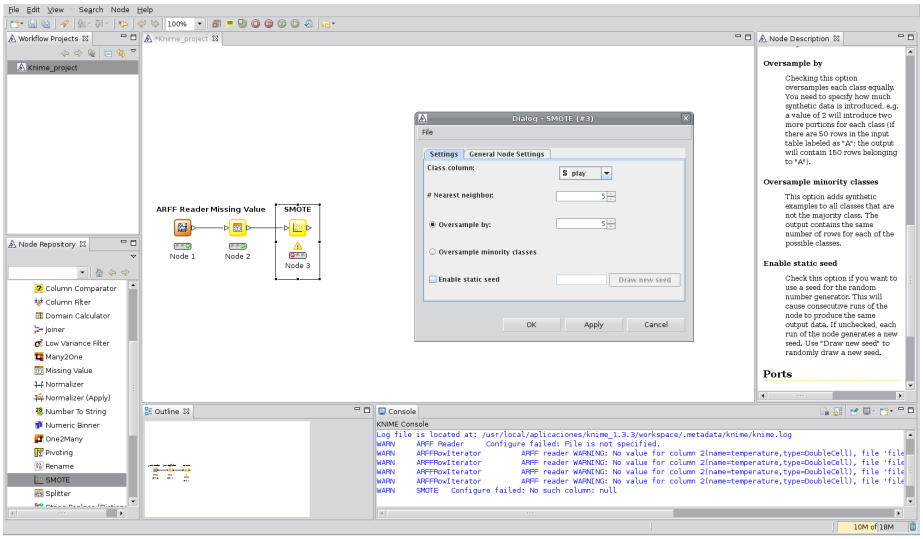

Figura 3.5: Generación de filas

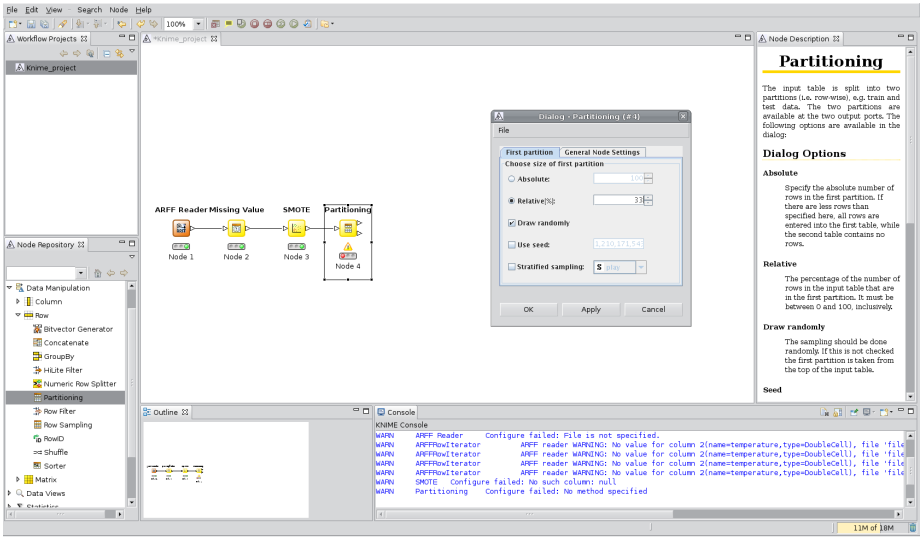

Figura 3.6: División de los datos

Capítulo 4

Clasificación no supervisada

Una vez se tengan los datos preparados para obtener resultados, se puede realizar la clasificación que se necesite. Para clasificación no supervisada existen varios nodos en Knime para los distintos tipos de clustering. Se encuentran en el apartado de Mining-Clustering del repositorio. En este ejemplo se va a utilizar K-medias para realizar el clustering, agrupando los datos de Temperature y Humidity en tres clusters como se aprecia en la figura 4.1.

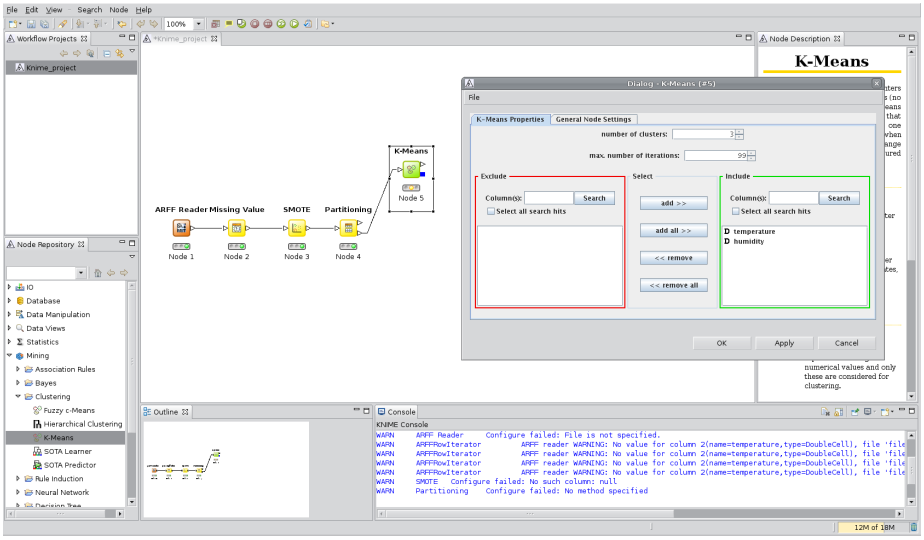

Figura 4.1: K-medias para realizar clustering

Una vez ejecutado el nodo, se obtiene el primer modelo creado en Knime. Hay multitud de opciones para visualizar los resultados obtenidos, en este caso se va a utilizar un gráfico de puntos (ScatterPlot). Para poder distinguir los puntos pertenecientes a cada cluster, se utiliza previamente el nodo Color Manager, el cual asigna un color distinto a todos los elementos de cada cluster generado. De esta forma la visualización es mucho más clara como se ve en la figura 4.2.

De esta forma se observa como los puntos quedan agrupados en cada cluster tal y como se esperaba que iba a realizar el algoritmo K-medias

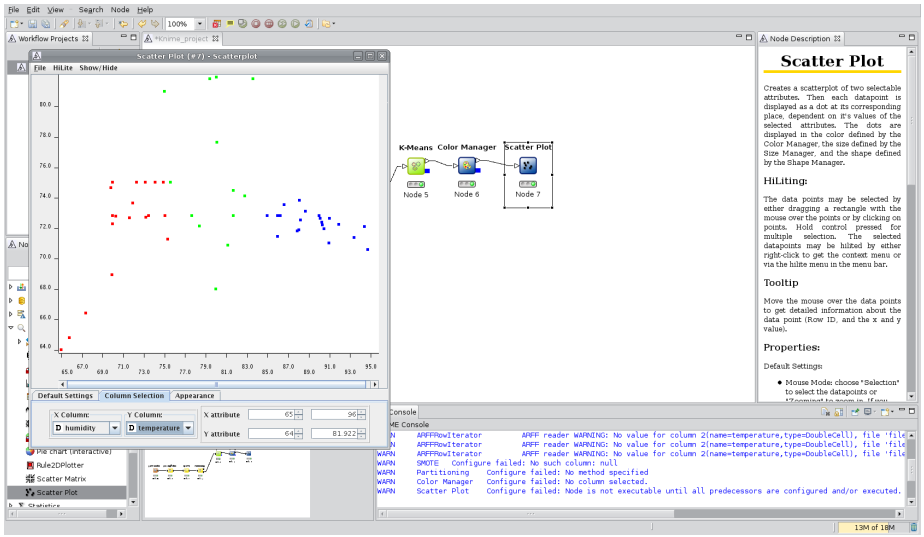

Figura 4.2: Visualizacin de los clusters creados

Capítulo 5

Clasificación supervisada

También existen multiples opciones para clasificación supervisada en Knime. Como ejemplo se va a realizar una clasificación con Naive Bayes. Los nodos que hay que utilizar se encuentran de nuevo en el apartado Mining del repositorio, en la sección de Bayes. Existen dos nodos, uno para aprender (Learner) el modelo y otro para validarlo (Predictor). Por esta razón es por la que el conjunto de datos ha sido separado en dos de forma previa. El conjunto de aprendizaje se utilizará como entrada en el nodo Learner, mientras que el modelo generado en este momento junto con el conjunto de test serán la entrada del nodo Predictor, el conjunto de esto se puede ver en la figura 5.1.

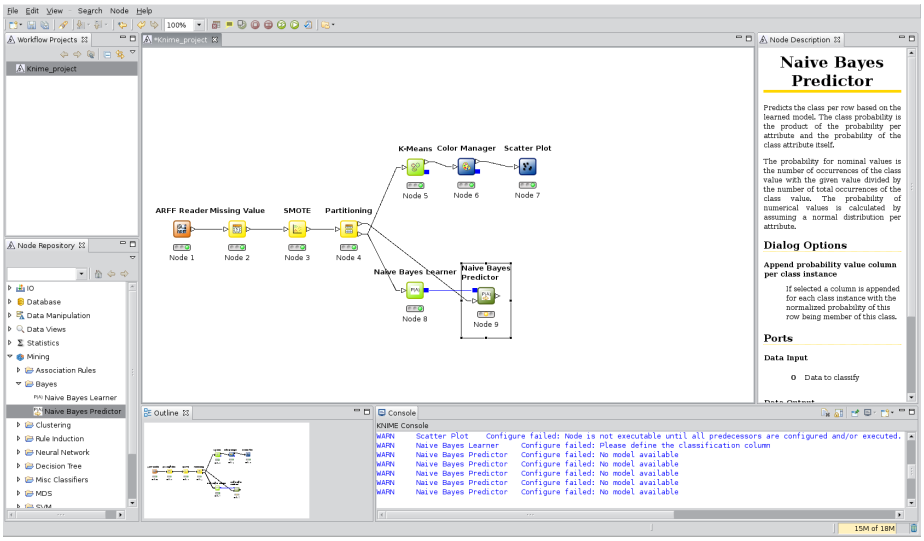

Figura 5.1: NaiveBayes para realizar clasificacin

Para obtener los resultados de la clasificación realizada con Naive Bayes, se puede utilizar el nodo Scorer. Este nodo de Knime, que también se encuentra en la zona de Mining-Scoring, se emplea para obtener la comparación entre dos columnas. En este caso, las columnas de Play que se tenían inicialmente, y las columnas generadas como estimación por el nodo Predictor de Naive Bayes.

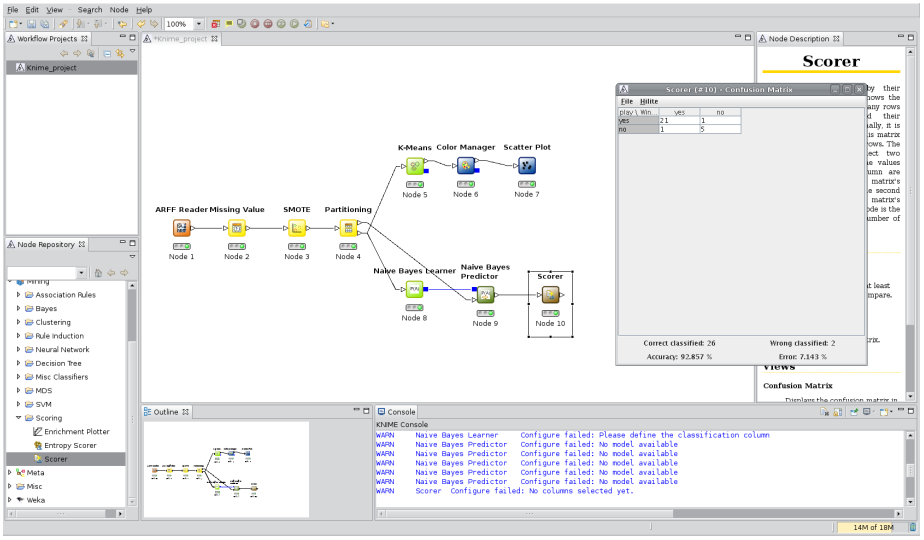

Figura 5.2: Obtención de los resultados de NaiveBayes

El resultado de ejecutar el nodo Scorer, como se puede ver en la figura 5.2, es la matriz de confusión de los datos, y por lo tanto, los valores de Accuracy y de Error que se han obtenido al realizar la estimación. En este caso, el valor resultante es muy positivo, debido a la similitud de los datos al haber sido generados la mayoría por un nodo de la herramienta basándose en el resto. Al ser datos muy parecidos, el aprendizaje es muy bueno y la evaluación por tanto, también lo es.

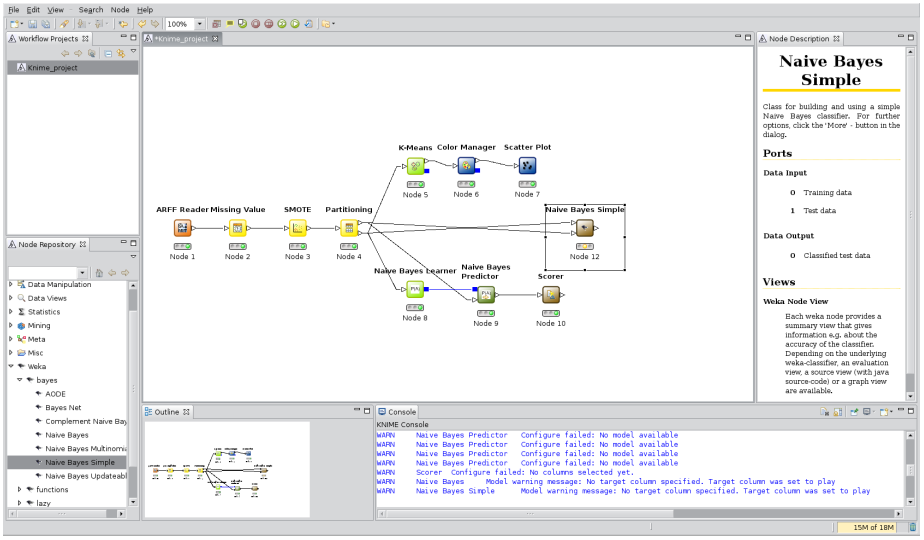

Figura 5.3: NaiveBayes utilizando Weka

Como ya se había adelantado, se puede utilizar la herramienta Knime junto con su conexión con otras aplicaciones. En este caso, vamos a utilizar el nodo de Naive Bayes correspondiente a la herramienta Weka, para comprobar que los resultados que se obtienen son los mismos como cabría esperar. Al haber realizado la conexión de forma previa, aparece el apartado Weka en el repositorio de nodos. Dentro de los múltiples algoritmos ofrecidos por Weka, seleccionamos el nodo correspondiente a Naive Bayes Simple. En este caso, tenemos dos conjuntos de datos de entrada, directamente uno de aprendizaje y otro de test, por lo que pasamos los datos de manera similar a lo realizado hasta el momento para obtener los resultados y teniendo cuidado de qué puerto es el correcto para cada conjunto de datos, todo esto se puede ver en la figura 5.3 [1]

Al ejecutar el nodo de Weka, aparece la salida que se obtendría al utilizar la propia herramienta Weka. Se puede ver en la figura 5.4 como los resultados obtenidos son los mismo que al utilizar los nodos de Knime.

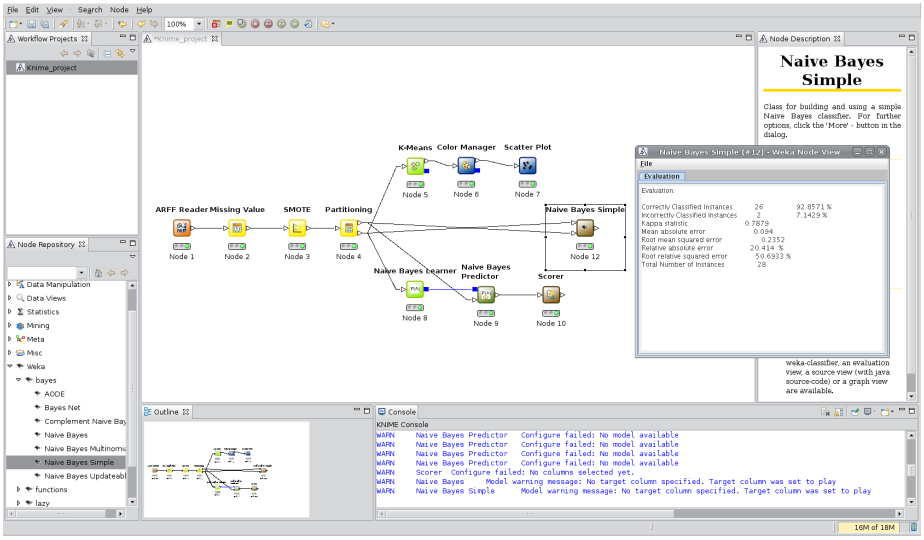

Figura 5.4: Resultados obtenidos con NaiveBayes de Weka

Por lo tanto de esta manera, se podría utilizar cualquier nodo de Knime, de las herramientas que se pueden conectar con Knime o incluso, los nodos creados por cada usuario que realizen la función que se requiera en cada caso.

[1] Para estar seguro de qué realiza cada nodo y cada puerto de entrada o salida de cada nodo, existe la ventana de ayuda en la parte derecha de la pantalla, la cual está relacionada con el nodo que esté seleccionado en ese momento