Diagram

Description automatically generated

3) El workflow utiliza el dataset Iris dividiéndolo con muestreo estratíficado en una parte de entrenamiento de un 80% y otra de prueba de un 20% para el caso de un modelo de árbol de regresión simple del cual se obtiene la performance y se la visualiza a través de una es gráfica.

Chart, waterfall chart

Description automatically generatedGraphical user interface, text, application

Description automatically generated Chart, line chart

Description automatically generated

4)

* 1. El dataset es el de Iris que es utilizado para predecir a partir del ancho y el largo del sépalo y el pétalo de una flor su clase entre Iris-setosa, Iris-versicolor y Iris virgínica.

Las opciones de importación del operador File Reader son configurar el Row Id, el delimitador de columna, si se leen los headers, si se ignoran los espacios en blanco o tabs, permite seleccionar tipos de tipos comentarios y configuraciones avanzadas.

* 1. El operador equivalente a FileReader en RapidMiner es Retrieve. Estos se diferencian en que el primero deja editar características del dataset luego de ser incluido en el proyecto como los tipos para las variables y en RM esto se configura solo al importar el dataset.
  2. En el dataset sepal length, sepal width, petal length y pethal width son doubles y class es una string. La variable de predicción se define en el Simple Regresión Tree Learner con el Target Column y en el Simple Regression Tree Predictor con el Prediction column name.

5)

* 1. El operador partitioning ofrece elegir el tamaño de la partición y la forma en la que se quiere separar determinando por ejemplo si se eligen los primeros valores o si se hace muestreo estartíficado, da la opción para usar una seed como en RM, permite alternativas acerca de las Flow variables que funcionan para a hacer variar ciertas configuraciones en el nodo de forma dinámica con cada ejecución y deja a elección políticas sobre el uso de la memoria.
  2. El operador equivalente a Patitioning en RapidMiner es Split. La principal diferencia es que mientras que en el componente de KNIME se hace una partición por cada unidad del componente, en RM se permite realizar varias a la vez.

6)

* 1. El operador Simple Regression Tree Learner soporta predictores de tipo numérico y categórico.
  2. Solo soporta target columns de tipo numérico.
  3. Utiliza un algoritmo base de árbol de regresión simple. El proceso sigue el algoritmo descripto en “Classification and Regression Tress” (Breiman et al, 1984) con algunas simplifaciones como no pruning, no necesariamente árboles binarios, tratar de encontrar la mejor dirección para los missing values, etc. En estos árboles de regesión el valor de predicción el valor para cada nodo hoja es la media de los registros dentro de ella y la predicción mejora cuanto menor sea la varianza de los valores dentro de una hoja. Por lo tanto, para armarlo en cada nodo se hacen splits que minimicen la suma de errores cuadráticos de los hijos.
  4. Los parámetros que se pueden configurar del algoritmo son usar splits binarios para los atributos nominales, la forma en la que se manejan los missing values siendo XGBoost un algoritmo que calcula la mejor dirección para los valores faltantes y Surrogate que calcula para cada Split otros alternativos que mejoran la aproximación, el límite para la profundidad del árbol, el mínimo de valores que puede tener un nodo para que el Split se intente y el mínimo de registros que un nodo hijo puede tener.

7)

* 1. El operador Simple Regression Tree Predictor recibe por un lado el modelo entrenado y por otro los datos de test, cómo salida tiene las predicciones realizadas.
  2. Permite modificar manualmente la columna de predicción.

8)

* 1. La gráfica compara los valores de predicción con la salida conocida para esos inputs del dataset de entrenamiento.
  2. Los parámetros que se pueden editar son el número de filas a mostrar, el límite del valor nominal a partir del cual una columna sea ignorada y la opción de colocar Flow Variables.

9)

* 1. Numeric Scorer funciona realizando los valores de coeficiente de terminación, media de error absoluto, error cuadrático medio y desviación media con signo de la predicción realizada.

10)