Ejercicio 2

Utilizaremos k-nn para clasificar las plantas de la especie “Iris”

Preparación de datos

1. Crea un nuevo proceso en RapidMiner
2. Descarga de UCI el dataset “Iris” e impórtalo.
    - Observa que tiene 4 atributos (reales) y una variable de salida, polinomial, que clasifica los tipos de plantas en 3 clases diferentes
3. Realiza un gráfico bidimensional, tomando como ejes “petal_length” y “petal_width”, y como “color column” la clase. Observa la distribución de los ejemplos.
    - ¿qué consideraciones puedes hacer a priori, en base a esta observación? Remite los comentarios a la tarea.
4. ¿qué tareas de acondicionamiento / preparación de los datos deben efectuarse?
    - Registra en un documento de texto y aplícalas al dataset
5. Agrega un operador “Split Data” para particionar el conjunto original en 2 subconjuntos del mismo tamaño, en forma aleatoria. Uno se usará para entrenamiento y el otro para test

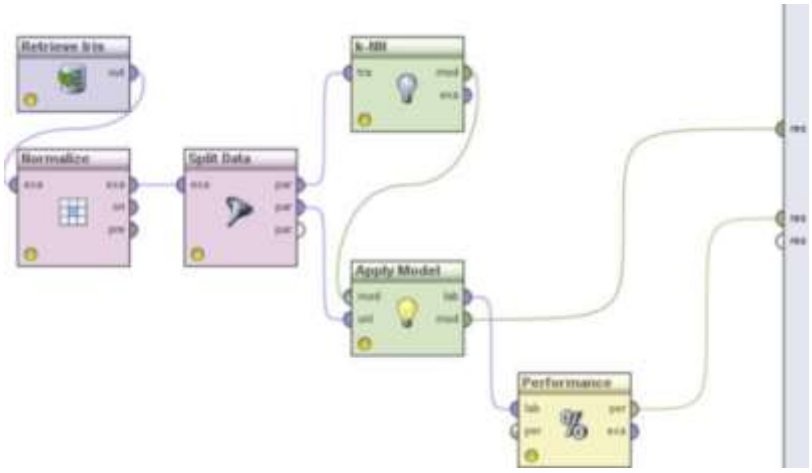
Operador de modelo y parámetros

El operador KNN de RapidMiner tiene algunos parámetros que se pueden configurar. Observa los mismos y resume en un documento de texto las principales características:

- k – tiene un valor por defecto de 1. Cámbialo a 3.
- Voto ponderado (Weighted Vote) - ¿cómo funciona? Toma nota de esto.
- Tipos de medición. RapidMiner tiene incluidas varias funciones para medición de distancia, que están agrupadas en Tipos de Medición
    - ¿Cuáles son estos tipos?
    - ¿qué características tiene cada uno?
- Funciones de medición. Observa las que están disponibles
    - Registra los nombres
    - ¿cómo funciona cada una?

Evaluación

Agrega un operador “Apply Model” y un “Performance (classification)”, y conecta los ports en forma apropiada para observar y comparar los resultados.



Ejecución e interpretación

1. Ejecutar el modelo y observar los resultados.
    a. Modelo k-nn: el modelo es simplemente todo el conjunto de entrenamiento.
    b. Vector de performance: matriz de confusión
2. Prueba con al menos 2 funciones de medición y valores de k diferentes. Realiza una matriz con estos datos, indicando los valores de exactitud de predicción alcanzados en cada caso.
3. En un POSTER, resume los hallazgos en función de los diferentes valores de k y de las funciones de distancia utilizadas. Explica estos resultados.

DISCUSION

Los equipos evaluarán los POSTERS de los demás equipos, y luego seguirá una discusión acerca de los mejores enfoques y resultados.