KNN Iris

Se crea un nuevo proceso con el dataset Iris de UCI. Luego, se grafican petal length y petal width en un gráfico bidimensional, diferenciando las diferentes especies por color.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Claramente la especie iris-setosa es muy distinta a las otras 2 y por lo tanto un modelo construido con knn debería reconocer perfectamente a todos los ejemplos de dicha clase. Para el caso de iris versicolor e iris virginica, la separación es más difusa, y es muy probable que nuestro modelo de knn cometa errores al clasificar algunos ejemplos de ellas.

En cuanto a las tareas de acondicionamiento necesarias para este dataset y este algoritmo, se deberán normalizar todos los atributos dado que KNN trabaja con distancias. Una diferencia en escala entre los atributos podría perjudicar substancialmente la performance del modelo.

Operador Knn de rapidminer

K: Refiere a la cantidad de vecinos más cercanos que utilizará el algoritmo para hacer una predicción sobre un ejemplo no visto.

Weighted vote: La aplicación de un voto ponderado refiere a que; dentro de los k vecinos escogidos, el más cercano a un nuevo ejemplo será el que tendrá el mayor peso al aplicar la predicción y el más lejano será el que tendrá el menor peso.

Measure types: El tipo de medida que se utilizará para encontrar los vecinos más cercanos. Aquí se podrá mencionar si se tendrán atributos categóricos, numéricos ambos, etc.

Función de medida: Se podrá seleccionar un tipo de distancia tal que sea acorde con el tipo de medida que se seleccionó anteriormente. Por ejemplo para datos numéricos, se puede utilizar la distancia euclídea, mientras que para datos nominales se puede utilizar la Jaccard similarity.

**Entrenamiento de modelo y resultados**

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamenteK=5

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamenteK=3

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamenteK=1

Interfaz de usuario gráfica

Descripción generada automáticamenteK=1 Distancia= Chebychev Distance

**Análisis de resultados**

Contrario a lo que hubiera pensado, los mejores resultados fueron obtenidos con k=1 para este Split de este dataset. Las distancias de manhattan y euclídea presentaron los mismos resultados para k = 1, mientras que la de chebychev presentó una exactitud levemente menor.

Mi hipótesis inicial para la clase iris-setosa resultó ser cierta para la mayoría de los modelos creados (menos para el modelo en el que se utiliza la distancia de chebychev). Esta clase obtuvo una precisión y un recall de 100%.