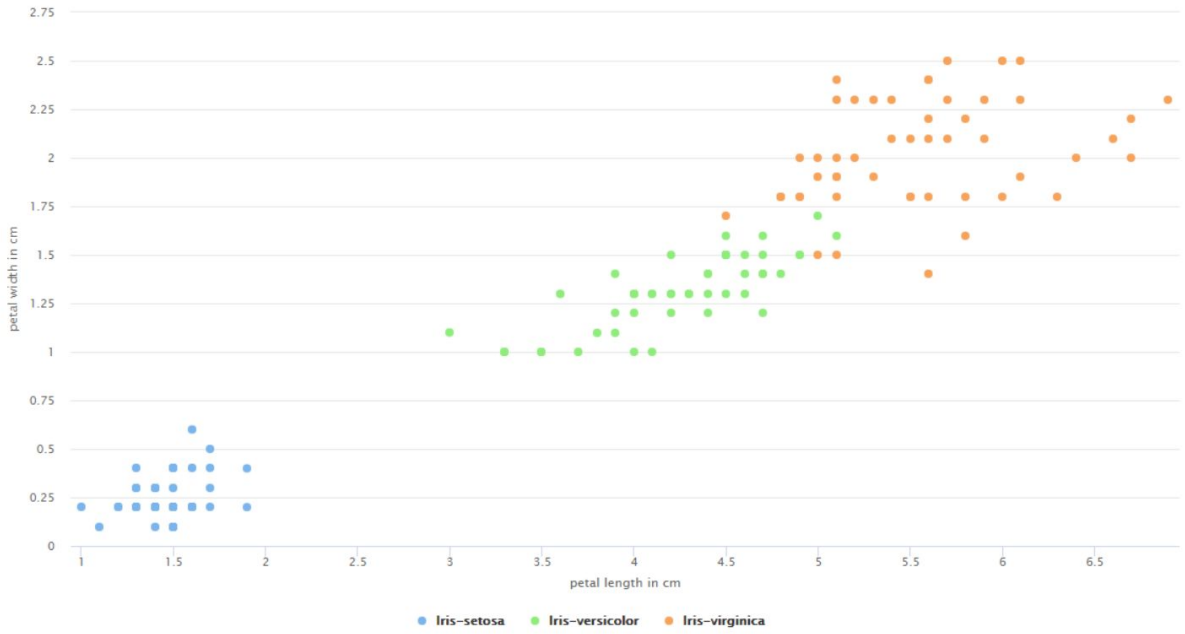
TA8 - Enzo Cozza - Agustín Fernández

Ejercicio 2

Preparación de datos

3)



Se puede observar que en cuanto al tamaño de los pétalos, una de las clases (setosa) se encuentra más alejada, mientras que las otras dos (versicolor y virgínica), si bien están separadas, son más cercanas entre sí.

A priori, se podría decir que los pétalos con ancho menor a 0.75cm y largo menor a 2cm pertenecen a la clase setosa. El límite entre las otras dos se encuentra un poco más confuso, sin embargo se podría estimar que el largo del pétalo para la clase versicolor se encuentra entre 3 y 5cm, y el ancho entre 1 y 1.5cm, y para la clase virgínica se podría tomar un largo mayor a los 5cm y un ancho mayor a 1.75cm.

Por último, se podría decir que con dos variables casi que se podría clasificar cualquier nuevo valor. Sin embargo, quizá utilizando una tercer variable, el límite entre versicolor y virgínica podría quedar aún más definido y ayudaría a una mejor clasificación de los datos.

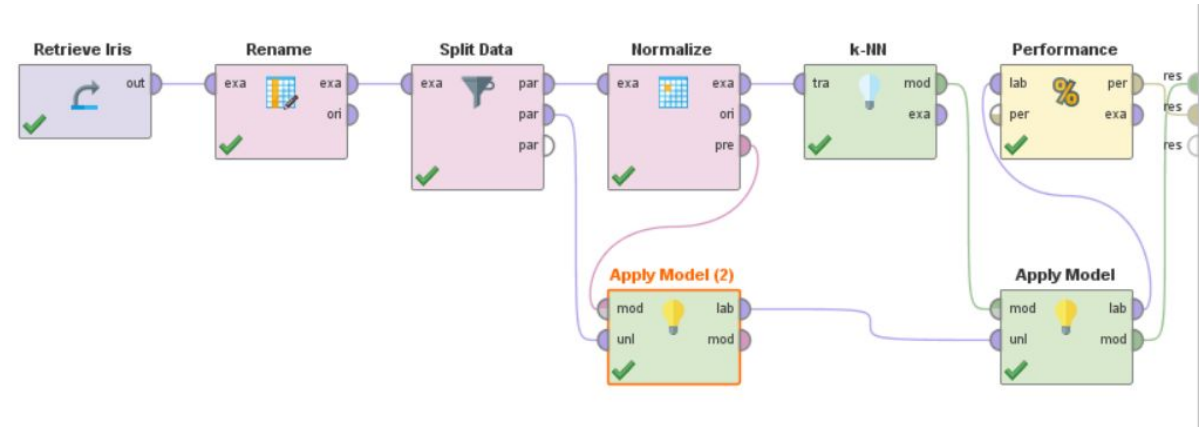
4) No presenta valores faltantes en los atributos ni tampoco se identificaron outliers, pero lo que sí se podría efectuar es una normalización en los datos de los atributos en cuestión, para reescalar los datos y asegurar que se encuentren en un mismo rango.

Operador de modelo y parámetros

k: Establece la cantidad de ejemplos de entrenamiento más cercanos que se considerarán para la predicción de clasificación.
Weighted vote: la distancia entre los ejemplos es tenida en cuenta a la hora de hacer una predicción. Los vecinos más cercanos tendrán un peso mayor sobre la decisión final que los que se encuentran más lejos.
Measure types: Selecciona el tipo de medición que será utilizado para encontrar los vecinos cercanos. Puede utilizar: MixedMeasures (utilizado para calcular distancias entre atributos con valores nominales y numéricos), NominalMeasures (se utiliza para calcular distancias entre atributos con valores nominales), NumericalMeasures (para calcular distancias entre atributos con valores numéricos) o BregmannDivergences (estas divergencias son tipos medidas de cercanía más genéricas).
Cada una de estas tiene distintas opciones para calcular la distancia:
MixedMeasures: MixedEuclideanDistance
NominalMeasures: NominalDistance, DiceSimmilarity, JaccardSimilarity, KulczynskiSimilarity, RogersTanimotoSimilarity, RussellRaoSimilarity, SimpleMatchingSimilarity.
NumericalMeasures: EuclideanDistance, CanberraDistance, ChebychevDistance, CorrelationSimilarity, CosineSimilarity, DiceSimilarity, DynamicTimeWarpingDistance, InnerProductSimilarity, JaccardSimilarity, KernelEuclideanDistance, ManhattanDistance, MaxProductSimilarity, OverlapSimilarity.
BregmanDivergence: GeneralizedDivergence, ItakuraSaitoDistance, KLDivergence, LogarithmicLoss, LogisticLoss, MahalanobisDistance, SquaredEuclideanDistance, SquaredLoss.

Evaluación

Ejecución e interpretación



k=3 - tipo de medición: numérica - medida: euclidean distance

accuracy: 93.33%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	3	88.46%
pred. Iris-virginica	0	2	22	91.67%
class recall	100.00%	92.00%	88.00%	

k=1 - tipo de medición: numérica - medida: euclidean distance

accuracy: 96.00%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	1	95.83%
pred. Iris-virginica	0	2	24	92.31%
class recall	100.00%	92.00%	96.00%	

k=5 - tipo de medición: numérica - medida: euclidean distance

accuracy: 93.33%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	3	88.46%
pred. Iris-virginica	0	2	22	91.67%
class recall	100.00%	92.00%	88.00%	

k=3 - tipo de medición: numérica - medida: manhattan distance

accuracy: 94.67%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	2	92.00%
pred. Iris-virginica	0	2	23	92.00%
class recall	100.00%	92.00%	92.00%	

k=1 - tipo de medición: numérica - medida: manhattan distance

accuracy: 96.00%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	1	95.83%
pred. Iris-virginica	0	2	24	92.31%
class recall	100.00%	92.00%	96.00%	

k=5 - tipo de medición: numérica - medida: manhattan distance

accuracy: 94.67%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	25	0	0	100.00%
pred. Iris-versicolor	0	23	2	92.00%
pred. Iris-virginica	0	2	23	92.00%
class recall	100.00%	92.00%	92.00%	

k=3 - tipo de medición: numérica - medida: chebychev distance

accuracy: 93.33%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	24	0	0	100.00%
pred. Iris-versicolor	1	23	2	88.46%
pred. Iris-virginica	0	2	23	92.00%
class recall	96.00%	92.00%	92.00%	

k=1 - tipo de medición: numérica - medida: chebychev distance

accuracy: 94.67%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	24	0	0	100.00%
pred. Iris-versicolor	1	23	1	92.00%
pred. Iris-virginica	0	2	24	92.31%
class recall	96.00%	92.00%	96.00%	

k=5 - tipo de medición: numérica - medida: chebychev distance

accuracy: 94.67%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	24	0	0	100.00%
pred. Iris-versicolor	1	23	1	92.00%
pred. Iris-virginica	0	2	24	92.31%
class recall	96.00%	92.00%	96.00%	

Se puede observar que utilizando un valor k=1, se obtiene la mayor precisión (96%) para los algoritmos de distancia euclidiana y distancia de manhattan.