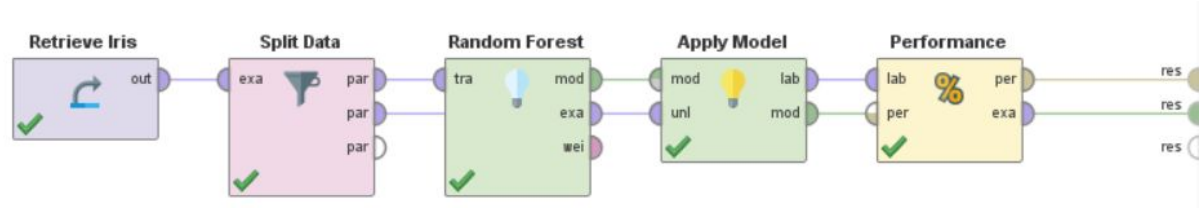
TA2 - Enzo Cozza - Agustín Fernández

Ejercicio 1



Se seleccionaron los siguientes valores para los parámetros:

maximal depth 3, gain_ratio, confidence vote.

Number of trees 10

accuracy: 95.56%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	15	0	0	100.00%
pred. Iris-versicolor	0	14	1	93.33%
pred. Iris-virginica	0	1	14	93.33%
class recall	100.00%	93.33%	93.33%	

Number of trees 100

accuracy: 95.56%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	15	0	0	100.00%
pred. Iris-versicolor	0	14	1	93.33%
pred. Iris-virginica	0	1	14	93.33%
class recall	100.00%	93.33%	93.33%	

La mayor parte de los árboles obtenidos, en ambos casos, cuentan en sus divisiones solamente con los atributos a3 y a4.
Aumentando la profundidad o la cantidad de árboles, la performance siempre se mantiene igual para el dataset dado.

Comparando con la TA1:

Votación

accuracy: 95.56%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	16	0	0	100.00%
pred. Iris-versicolor	0	16	1	94.12%
pred. Iris-virginica	0	1	11	91.67%
class recall	100.00%	94.12%	91.67%	

Bagging

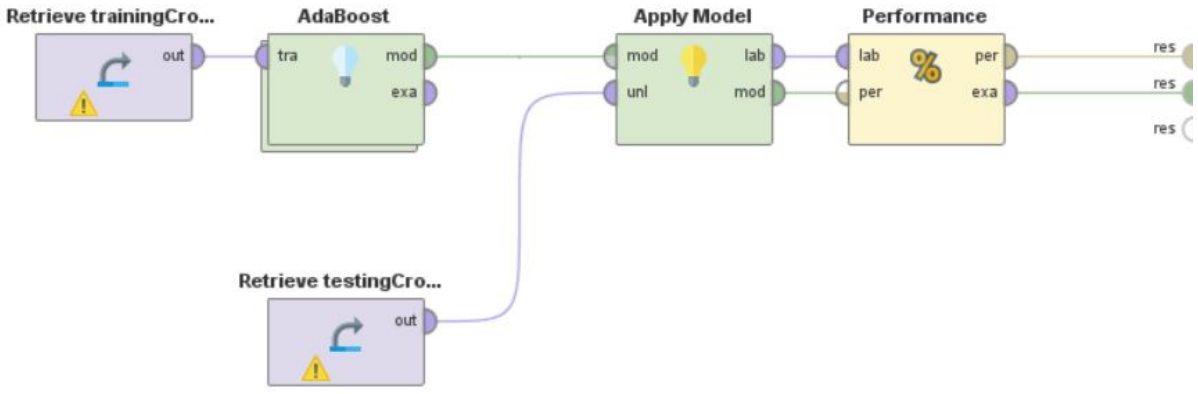
accuracy: 95.56%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	16	0	0	100.00%
pred. Iris-versicolor	0	16	1	94.12%
pred. Iris-virginica	0	1	11	91.67%
class recall	100.00%	94.12%	91.67%	

Se puede observar que se obtuvo la misma performance en los tres casos. Esto probablemente esté dado por el dataset con el que fueron evaluados, el cual tiene sus tres clases bastante marcadas.

Ejercicio 2

Parte 1



El problema que presenta este conjunto de datos es automatizar la clasificación de imágenes de satélite de diferentes clases de suelos. Las clases son: *impervious, farm, forest, grass, orchard, water.*

Los demás atributos son *max_ndvi* y *20150720_N - 20140101_N*. El primero es la máxima diferencia normalizada del índice de vegetación. Los demás son valores de NDVI extraídos (entre Enero 2014 y Julio 2015) en orden cronológico inverso con el formato *yyyymmdd*.

Árbol de decisión

accuracy: 54.00%

	true water	true forest	true grass	true farm	true orchard	true impervious	class precision
pred. water	37	0	0	0	0	1	97.37%
pred. forest	3	71	22	37	45	1	39.66%
pred. grass	0	3	6	1	0	0	60.00%
pred. farm	3	2	6	13	1	4	44.83%
pred. orchard	0	0	0	1	1	0	50.00%
pred. impervious	3	2	2	1	0	34	80.95%
class recall	80.43%	91.03%	16.67%	24.53%	2.13%	85.00%	

Iteraciones	Precisión
5	54%
10	54%
100	54%

k-NN

Se elige utilizar este método ya que es un método sencillo de utilizar, y ya que se utilizan valores reales para los atributos (con tres dígitos después de la coma), se decide no utilizar naive bayes.

accuracy: 62.00%

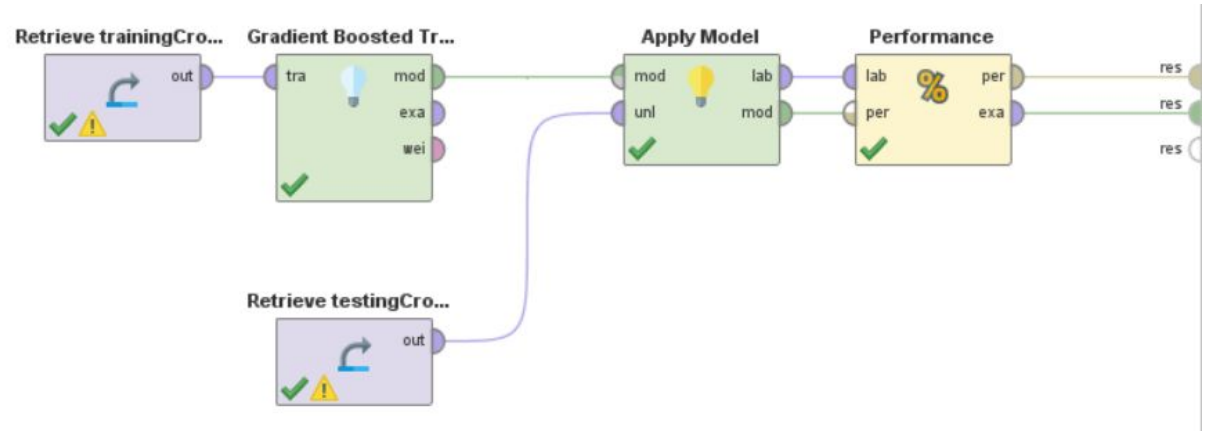
	true water	true forest	true grass	true farm	true orchard	true impervious	class precision
pred. water	32	0	1	0	0	0	96.97%
pred. forest	1	39	5	7	13	0	60.00%
pred. grass	5	26	16	1	3	2	30.19%
pred. farm	0	10	6	44	13	1	59.46%
pred. orchard	0	0	0	0	18	0	100.00%
pred. impervious	8	3	8	1	0	37	64.91%
class recall	69.57%	50.00%	44.44%	83.02%	38.30%	92.50%	

Iteraciones	Precisión
5	62%
10	62%
100	62%

Se puede concluir que al aplicar AdaBoost la precisión no sufre mayores cambios ya que analizando los modelos obtenidos, no se producen mejoras luego de las primeras 3 iteraciones.

Parte 2

Gradient Boosted Trees



accuracy: 60.33%

	true water	true forest	true grass	true farm	true orchard	true impervious	class precision
pred. water	36	0	1	0	0	1	94.74%
pred. forest	2	65	16	14	41	1	46.76%
pred. grass	1	7	7	1	1	1	38.89%
pred. farm	3	5	10	37	5	1	60.66%
pred. orchard	0	0	0	0	0	0	0.00%
pred. impervious	4	1	2	1	0	36	81.82%
class recall	78.26%	83.33%	19.44%	69.81%	0.00%	90.00%	

Un modelo *gradient boosted* es un ensamble de árboles tanto de regresión como de clasificación. Ambos métodos son métodos de ensamble que obtienen valores de predicción a partir de estimaciones gradualmente incrementales. *Boosting* es un método no lineal de regresión que permite mejorar la exactitud de los árboles. Aplicando secuencialmente algoritmos de clasificación débiles a los datos cambiados incrementalmente, una serie de árboles de decisión son creados que producen un ensamble de modelos de predicción débiles. Mientras los *boosting trees* incrementan la precisión, también decrementan velocidad e interpretabilidad. Este método *gradient boosting* generaliza el *tree boosting* para minimizar estos asuntos.