Curso 2018-2019

APRENDIZAJE AUTOMÁTICO: ÁRBOLES DE REGRESIÓN Y ÁRBOLES DE MODELOS

mrodrigue212@alumno.uned.es | Manuel Rodríguez Sánchez

Contenido

Actividad 3

Tabla con el resumen de datos después de ejecutar cada uno de los tres experimentos

	Experimento 1	Experimento 2					Experimento 3
		1	5	10	15	20	
Coeficiente de correlación	0,9898	0,9206	0,9313	0,9465	0,9042	0,9840	0,8953
Error absoluto medio	8,5488	23,7948	18,3427	15,362	19,3339	11,9027	21,3458
Error cuadrático medio	19,6818	54,4715	41,9993	29,8377	52,6543	28,5842	66,983
Error absoluto relativo (%)	10,26%	28,38%	22,51%	21,29%	25,45%	13,58%	25,39%
Error cuadrático relativo (%)	14,24%	40,02%	36,42%	35,60%	44,36%	19,57%	48,07%
Número total de registros	188	64	64	64	64	64	188

Pregunta 1 - ¿Por qué el menor error se obtiene en el primer experimento?

Porque se manejan los 188 registros de la base de datos, es decir, todos los datos disponibles, esto hace que el porcentaje de error disminuya.

Pregunta 2 - ¿Por qué fluctúa el porcentaje de acierto en las distintas ejecuciones del segundo experimento?

Primeramente es por que estamos usando los 2/3 de la base de datos y al tener menos ejemplos el porcentaje de error es mayor. En segundo lugar, las distintas ejecuciones las hacemos con distintos valores de semillas, y esto hace que se particionen los datos; a distinto valor de semilla, la partición que se obtiene es también distinta.

Pregunta 3 – ¿Por qué se dice que el resultado más fiable de la evaluación de este modelo es obtenido en el tercer experimento, el siguiente más fiable es el obtenido en el segundo modelo, y el resultado menos fiable corresponde al obtenido en el primero?

Por que en el tercer experimento se utiliza la evaluación con validación cruzada, y se hacen 10 validaciones, que son las que se indican en el campo Folds; al ser una base de datos pequeña, esta es la forma de validación más aconsejable, y que mejores resultados da.

La fiabilidad del segundo experimento, se basa en que cuando se construye el clasificador, se extrae un subconjunto de datos del conjunto de datos original; el clasificador es evaluado en función del otro subconjunto que no fue seleccionado.

Actividad 4

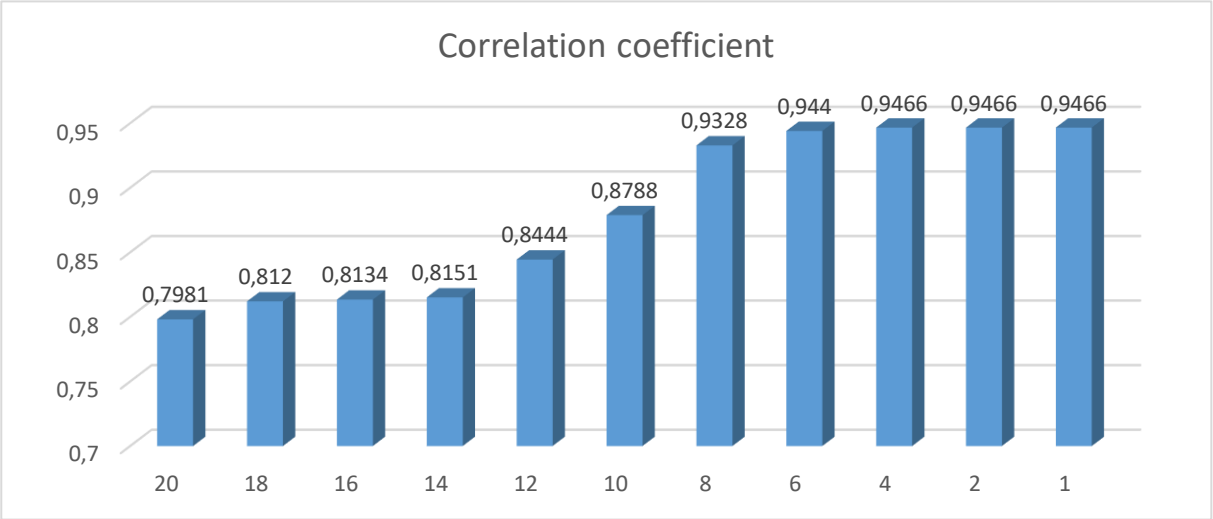


Ilustración 1 - Coeficiente de correlación

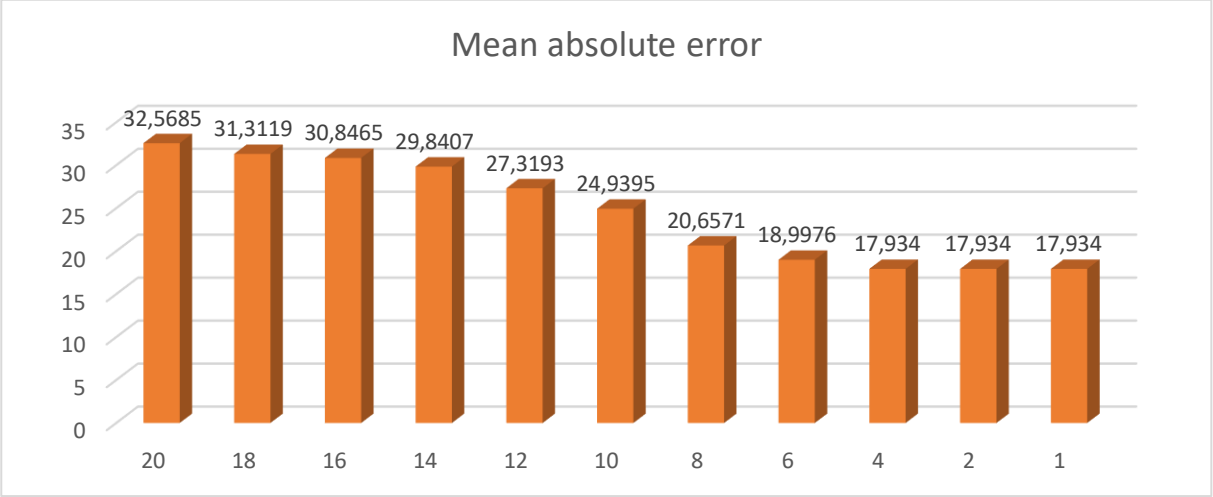


Ilustración 2 - Error absoluto medio

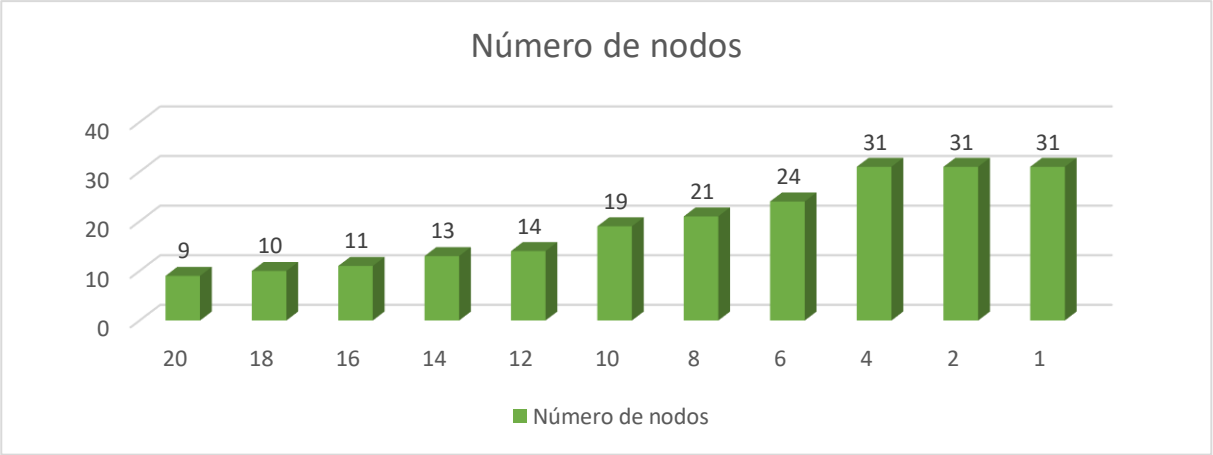


Ilustración 3 - Número de nodos generados.

Podemos observar como configurando el algoritmo para que el número de instancias sean 6 (como tope para que un nodo sea considerado como nodo hoja) el coeficiente de correlación es el más óptimo y no varía prácticamente nada con las instancias 4, 2 y 1; lo mismo ocurre

con el error absoluto medio; pero lo más destacable es que si ponemos un valor menor a 6, el número de nodos aumenta y con ello la complejidad del árbol, por lo tanto, el valor de la instancia más óptimo sería el 6, mantiene un buen coeficiente de correlación, el error absoluto medio no varía mucho con respecto a la instancia 4, y el árbol sería menos complejo al tener menos nodos.

	20	18	16	14	12	10	8	6	4	2	1
Correlation coefficient	0,7981	0,812	0,8134	0,8151	0,8444	0,8788	0,9328	0,944	0,9466	0,9466	0,9466
Mean absolute error	32,5685	31,3119	30,8465	29,8407	27,3193	24,9395	20,6571	18,9976	17,934	17,934	17,934
Root mean squared error	84,4914	81,7041	81,2804	80,8697	74,1899	67,2031	49,969	46,2894	45,3822	45,3822	45,3822
Relative absolute error (%)	39,0346	37,5285	36,9708	35,7653	32,7432	29,891	24,7583	22,7694	21,4946	21,4946	21,4946
Root relative squared error (%)	61,0149	59,0021	58,6961	58,3995	53,5757	48,5303	36,0848	33,4276	32,7725	32,7725	32,7725
Número de nodos	9	10	11	13	14	19	21	24	31	31	31
Linear Model (LM)	10	11	12	14	15	20	22	25	32	32	32
Total number of instances	188	188	188	188	188	188	188	188	188	188	188

Ilustración 4 - Tabla de datos obtenidos en el experimento 4

Actividad 5

Gráfica con las prestaciones del modelo: coeficiente de correlación y error medio absoluto.

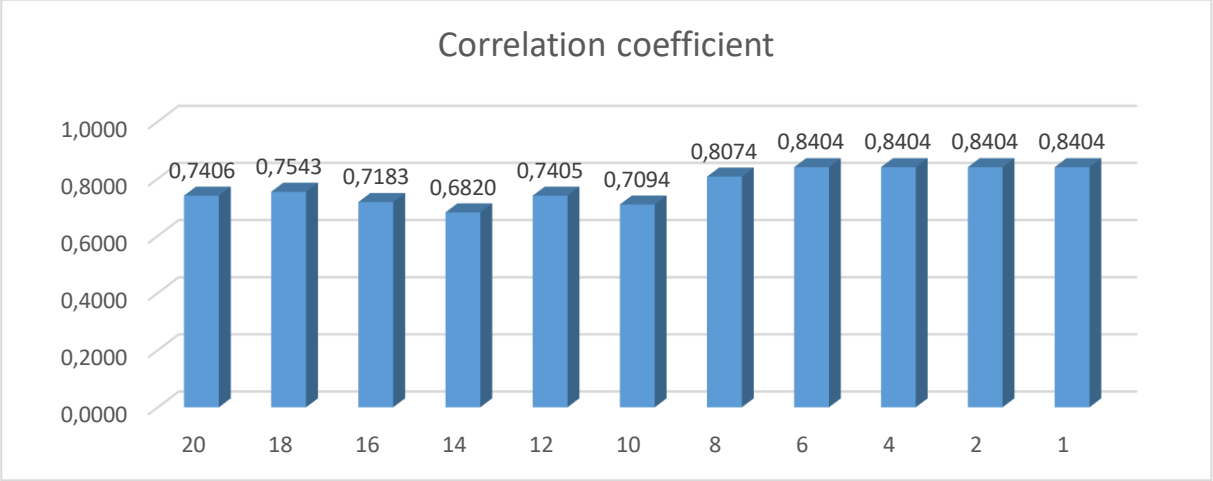


Ilustración 5 - Coeficiente de correlación Experimento 5

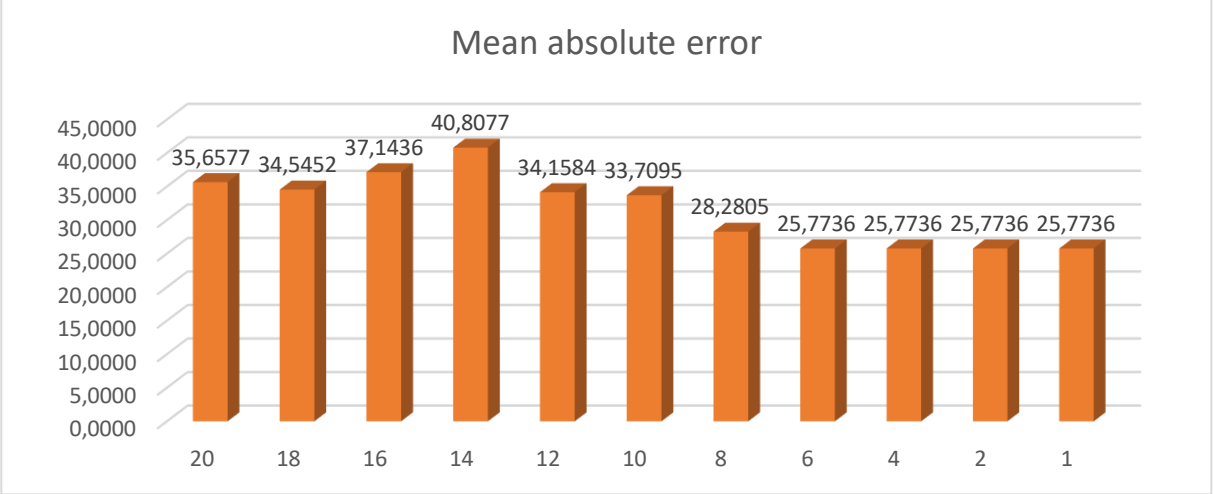


Ilustración 6 - Error absoluto medio Experimento 5

Gráfica con la complejidad del modelo en base al número de nodos.

Ilustración 7 - Número de nodos Experimento 5

Podemos ver que en el experimento 4, que no hemos podado el árbol, los valores del coeficiente de correlación y el error medio son mejores con respecto a los valores del experimento 5, donde sí se ha podado el árbol. Evidentemente el número de nodos del árbol del experimento 4 es mayor que el del experimento 5, tal y como se aprecia en las gráficas, esto provoca que su complejidad computacional aumente.

En resumen, los datos cuando no hay poda mejoran, pero aumenta la complejidad computacional y con ello su coste. Sin embargo, los datos cuando hay poda empeoran, pero al haber menos nodos, la complejidad computacional es menor. En base a esta información, pienso que el árbol más competitivo es el de mayor complejidad (Experimento 4, instancias 6 sin poda), ya que el dato que nos va a dar será más puro o fiable y equilibrado (depende también de la actividad que estemos analizando y de los recursos que dispongamos).

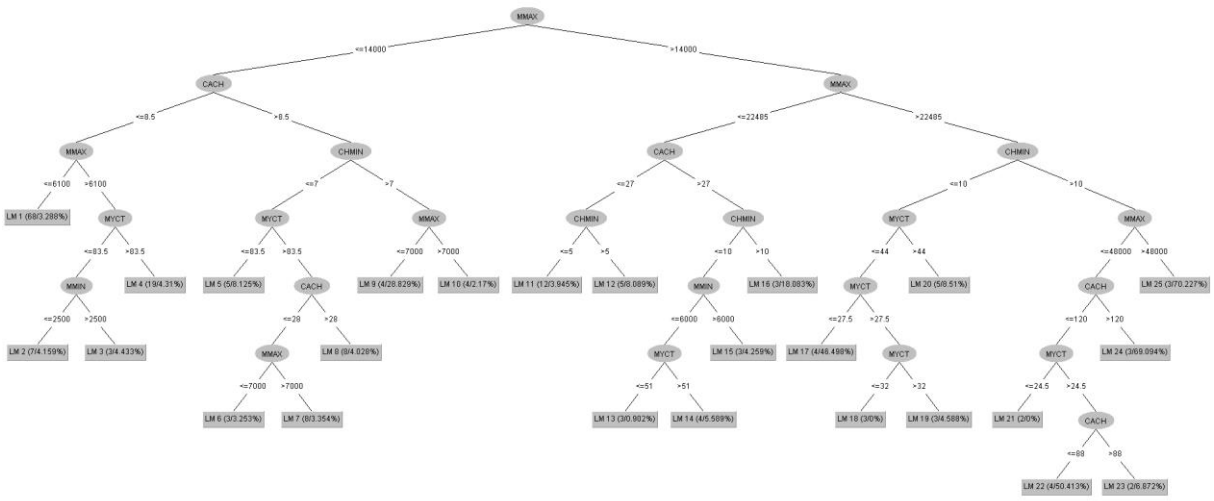


Ilustración 8 - Árbol más competitivo

	20	18	16	14	12	10	8	6	4	2	1
Correlation coefficient	0,7406	0,7543	0,7183	0,6820	0,7405	0,7094	0,8074	0,8404	0,8404	0,8404	0,8404
Mean absolute error	35,6577	34,5452	37,1436	40,8077	34,1584	33,7095	28,2805	25,7736	25,7736	25,7736	25,7736
Root mean squared error	94,9883	92,1521	98,2407	101,278	94,9473	97,9510	82,2086	75,6733	75,6733	75,6733	75,6733
Relative absolute error (%)	42,7372	41,4038	44,5181	48,9096	40,9401	40,4021	33,8953	30,8906	30,8906	30,8906	30,8906
Root relative squared error (%)	68,5952	66,5470	70,9438	73,1377	68,5656	70,7347	59,3664	54,6470	54,647	54,6470	54,6470
Número de nodos	7	7	5	11	11	11	11	11	11	11	11
Linear Model (LM)	8	8	6	12	12	12	12	12	12	12	12
Total numbre of instances	188	188	188	188	188	188	188	188	188	188	188

Ilustración 9 - Tabla con los resultados del experimento 5

LM2
44
44
44
59
41
47
46,5

Actividad 6

Traducción del árbol visualizado en el apartado anterior, a un sistema de reglas if – then.

```
if MMAX<=14000 then
        if CACH<=8.5 then
                if MMAX<=6100 then
                        LM1
                else
                        if MYCT<=83.5 then
                                if MMIN<=2500 then
                                        LM2
                                else MMIN>2500 then
                                        LM3
                else
                        if MYCT>83.5 then
                                LM4
        else
                if CACH>8.5 then
                        if CHMIN<=7 then
                                if MYCT<=83.5 then
                                        LM5
                                else
                                        if MYCT>83.5 then
                                                if CACH<=28 then
                                                        if MMAX<=7000 then
                                                                LM6
                                                        else
                                                                if MMAX>7000 then
                                                                        LM7
                                                else
                                                        if CACH>28 then
                                                                LM8
                        else
                                if CHMIN>7 then
                                        if MMAX<=7000 then
                                                LM9
                                        else
                                                if MMAX>7000 then
                                                        LM10
else
        if MMAX>14000 then
                if MMAX<=22485 then
                        if CACH<=27 then
                                if CHMIN<=5 then
                                        LM11
                                else
                                        if CHMIN>5 then
                                                LM12
                        else
                                if CACH>27 then
                                        if CHMIN<=10 then
                                                if MMIN<=6000 then
                                                        if MYCT <=51 then
                                                                LM13
                                                        else
                                                                if MYCT>51 then
                                                                        LM14
                                                else
                                                        if MMIN>6000 then
                                                                LM15
                                                else
                                                        if CHMIN>10 then
                                                                LM16
                else
                        if MMAX>22485 then
                                if CHMIN<=10 then
                                        if MYCT<=44 then
```

```
                    if MYCT<=27.5 then
                        LM17
                    else
                        if MYCT>27.5 then
                            if MYCT<=32 then
                                LM18
                            else
                                if MYCT>32 then
                                    LM19
            else
                    if MYCT>44 then
                        LM20
    else
            if CHMIN>10 then
                    if MMAX<=48000 then
                        if CACH<=120 then
                            if MYCT<=24.5 then
                                LM21
                            else
                                if MYCT>24.5 then
                                    if CACH<=88 then
                                        LM22
                                    else
                                        if CACH>88 then
                                            LM23
                        else
                            if CACH>120 then
                                LM24
                    else
                        if MMAX>48000 then
                            LM25
```

Actividad 8

Resultados de aplicar las instrucciones indicadas en la actividad:

=== Re-evaluation on test set ===

User supplied test set
Relation: cpu-new
Instances: unknown (yet). Reading incrementally
Attributes: 7

=== Predictions on user test set ===

inst#	*actual*	*predicted*	*error*
1	*?*	*174*	*?*
2	*?*	*470.333*	*?*
3	*?*	*24.074*	*?*
4	*?*	*47.5*	*?*
5	*?*	*24.074*	*?*
6	*?*	*24.074*	*?*
7	*?*	*24.074*	*?*
8	*?*	*131.75*	*?*
9	*?*	*24.074*	*?*
10	*?*	*40.368*	*?*
11	*?*	*40.368*	*?*
12	*?*	*350.5*	*?*
13	*?*	*174*	*?*
14	*?*	*24.074*	*?*
15	*?*	*75.917*	*?*
16	*?*	*77*	*?*
17	*?*	*24.074*	*?*
18	*?*	*45.714*	*?*
19	*?*	*75.917*	*?*
20	*?*	*94.4*	*?*
21	*?*	*260.75*	*?*

=== Summary ===

Total Number of Instances 0
Ignored Class Unknown Instances 21

Se nos muestran 21 instancias o registros de la base de datos que hemos cargado, donde se aprecian varias columnas: *ins#*: número de registro, *actual*: valor actual que pone una interrogación ya que no dispone de ningún valor, *predicted*: es el valor calculado esperado, *error*: pone interrogación ya que este no se ha podido calcular al faltar el valor de la clase de cada uno de los registros.

Actividad 9

Resultados de aplicar las instrucciones del ejercicio:

=== Re-evaluation on test set ===

User supplied test set
Relation: cpu-test
Instances: unknown (yet). Reading incrementally
Attributes: 7

=== Predictions on user test set ===

inst#	*actual*	*predicted*	*error*
1	*149*	*154.667*	*5.667*
2	*157*	*109.75*	*-47.25*
3	*460*	*882*	*422*
4	*28*	*24.074*	*-3.926*
5	*67*	*54.333*	*-12.667*
6	*34*	*40.368*	*6.368*
7	*31*	*24.074*	*-6.926*
8	*95*	*109.75*	*14.75*
9	*17*	*24.074*	*7.074*
10	*35*	*24.074*	*-10.926*
11	*41*	*47.5*	*6.5*
12	*34*	*40.368*	*6.368*
13	*80*	*75.917*	*-4.083*
14	*36*	*40.368*	*4.368*
15	*30*	*24.074*	*-5.926*
16	*126*	*134*	*8*
17	*267*	*272.333*	*5.333*
18	*25*	*24.074*	*-0.926*
19	*64*	*75.917*	*11.917*
20	*23*	*24.074*	*1.074*
21	*117*	*109.75*	*-7.25*

=== Summary ===

Correlation coefficient	*0.9518*
Mean absolute error	*28.5382*
Root mean squared error	*92.9523*
Total Number of Instances	*21*

	Validación cruzada	Generado a partir de un conjunto de datos
Correlation coefficient	0,8404	0,9518
Mean absolute error	25,7736	28,5382
Root mean squared error	75,6733	92,9523

Ilustración 10 - Tabla comparativa de los dos métodos

¿Se puede afirmar que la estimación del error mediante validación cruzada es comparable con el error obtenido al aplicar el modelo a un conjunto de datos no usado en el entrenamiento?

Considero que los datos no son comparables al no usar el mismo número de registros para la elaboración de los modelos. Es cierto que el coeficiente de correlación del generado a partir del conjunto, mejora con respecto a la validación cruzada, pero si nos fijamos en el error cuadrático medio y el error absoluto medio, estos aumentan de forma importante.

Actividad 10

	Modelo seleccionado en la actividad 5	Modelo generado en la actividad 10
Correlation coefficient	0,8404	0,8374
Mean absolute error	25,7736	26,2405
Root mean squared error	75,6733	76,1336
Relative absolute error (%)	30,8906	31,4502
Root relative squared error (%)	54,6470	54,9793
Número de nodos	11	11
Linear Model (LM)	12	12
Total numbre of instances	188	188

Ilustración 11 - Tabla comparativa de prestaciones con los modelos creados en el punto 5 y el punto 10.

Observamos que el hecho de meter ruido en el árbol (introduciendo el atributo "vendor"), no influye de forma importante en el resultado. Vemos en la tabla comparativa que el modelo seleccionado en la actividad 5 (recordemos que habíamos seleccionado al final uno de los generados en la actividad 4) no varía mucho con respecto a esos mismos datos, pero con el atributo "vendor" añadido. Corrobora pues, la tolerancia de los árboles de regresión al ruido.

Actividad 11

Resultados después de aplicar las instrucciones del enunciado de la actividad:

	20	18	16	14	12	10	8	6	4	2	1
Correlation coefficient	0,9713	0,9711	0,9738	0,9733	0,9770	0,9767	0,9728	0,9742	0,9925	0,9925	0,9925
Mean absolute error	16,8359	18,1163	16,0927	16,3793	13,8325	12,9392	13,7491	12,0473	7,3255	7,3255	7,3255
Root mean squared error	32,8764	33,0762	31,4511	31,7761	29,6228	30,1842	32,2196	31,2466	17,3694	17,3694	17,3694
Relative absolute error (%)	20,1784	21,7130	19,2877	19,6312	16,5788	15,5082	16,4788	14,4392	8,7800	8,7800	8,7800
Root relative squared error (%)	23,7415	23,8858	22,7122	22,9469	21,3919	21,7973	23,2672	22,5646	12,5432	12,5432	12,5432
Número de nodos	1	1	1	1	1	3	3	3	3	3	3
Linear Model (LM) Nodos hoja	2	2	2	2	2	4	4	4	4	4	4
Total number of instances	188	188	188	188	188	188	188	188	188	188	188

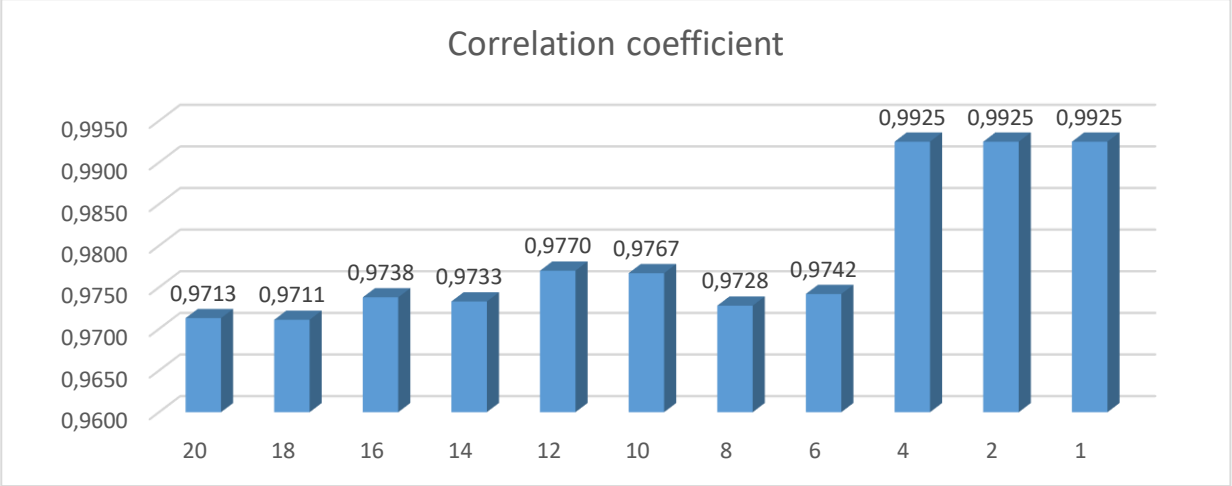


Ilustración 12 - Coeficiente de correlación para el árbol de modelos

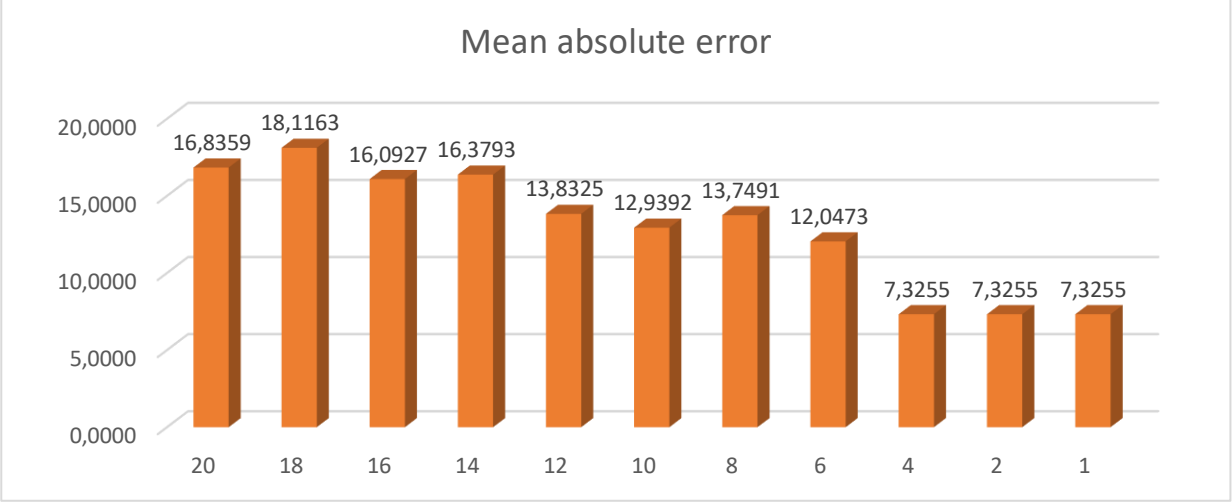
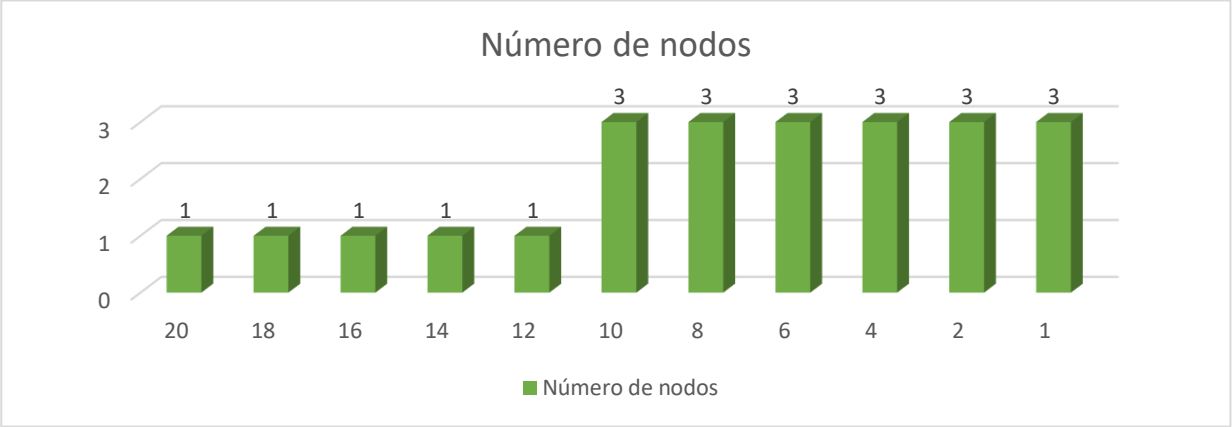


Ilustración 13 - Error absoluto medio para el árbol de modelos



Podemos observar como los resultados son sumamente buenos con respecto a la regresión. El coeficiente de correlación no baja del 0,97 y hay un error bastante pequeño. Podemos ver que a partir del modelo con 4 instancias, aumenta el coeficiente hasta el 0,99 y disminuye el error hasta el 7,3.

Comparativa con las prestaciones del ejercicio 5:

	Árbol de Modelos	Árbol de regresión
	Modelo de la actividad 11 con 4 instancias	Modelo de la actividad 5 con 6 instancias
Correlation coefficient	0,9925	0,9440
Mean absolute error	7,3255	18,9976
Root mean squared error	17,3694	46,2894
Relative absolute error (%)	8,7800	22,7694
Root relative squared error (%)	12,5432	33,4276
Número de nodos	3	24
Linear Model (LM) Nodos hoja	4	25
Total number of instances	188	188

Ilustración 14 - Tabla comparativa donde se aprecia la mejora del árbol de modelo con respecto al árbol de regresión.

Si hemos de quedarnos con un modelo, lo haríamos con el que dispone de 4 instancias, su coeficiente de correlación es el mejor, junto con el error absoluto que es bastante bajo, además no es muy complejo computacionalmente al tener solo 3 nodos.

Actividad 12

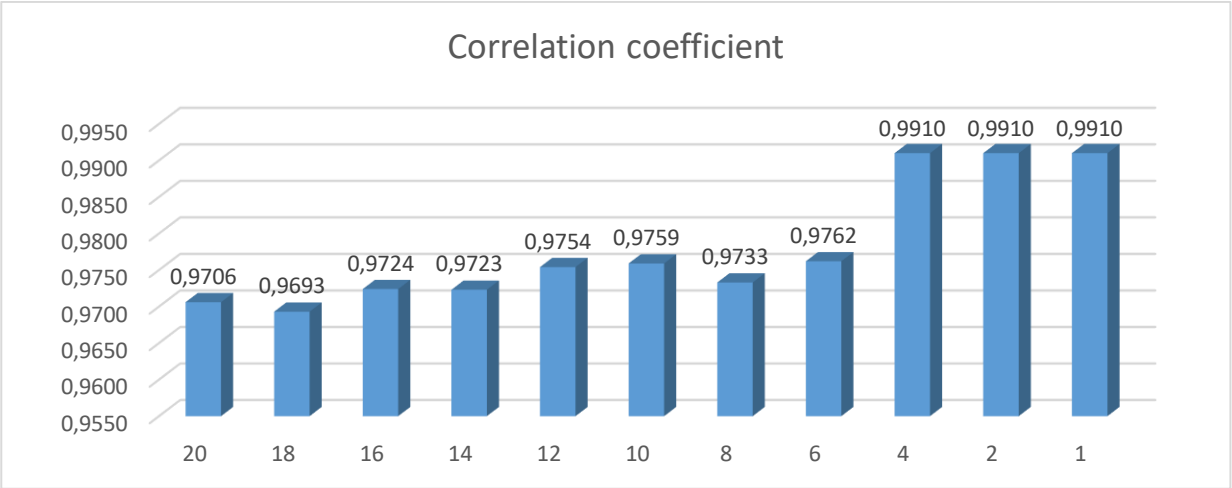
	Árbol de Modelos	
	Modelo con 4 instancias Mediante validación cruzada	Modelo reevaluado
Correlation coefficient	0,9925	0,9671
Mean absolute error	7,3255	24,1465
Root mean squared error	17,3694	87,0336
Total number of instances	188	21

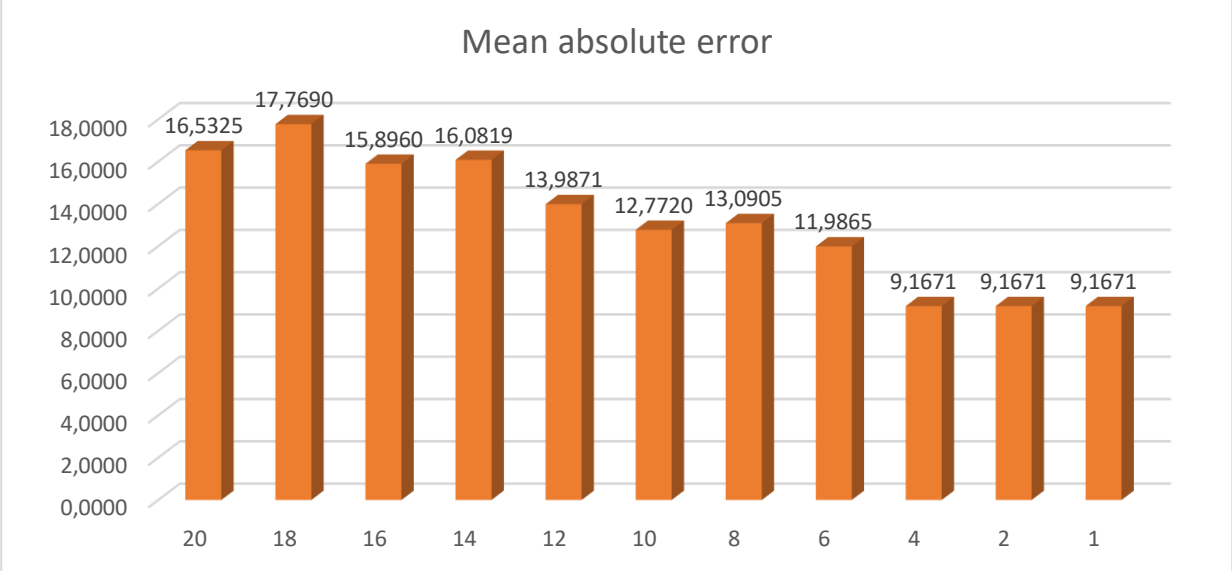
¿Se puede afirmar que la estimación del error mediante validación cruzada es comparable con el error obtenido al aplicar el modelo a un conjunto de datos no usado en el entrenamiento?

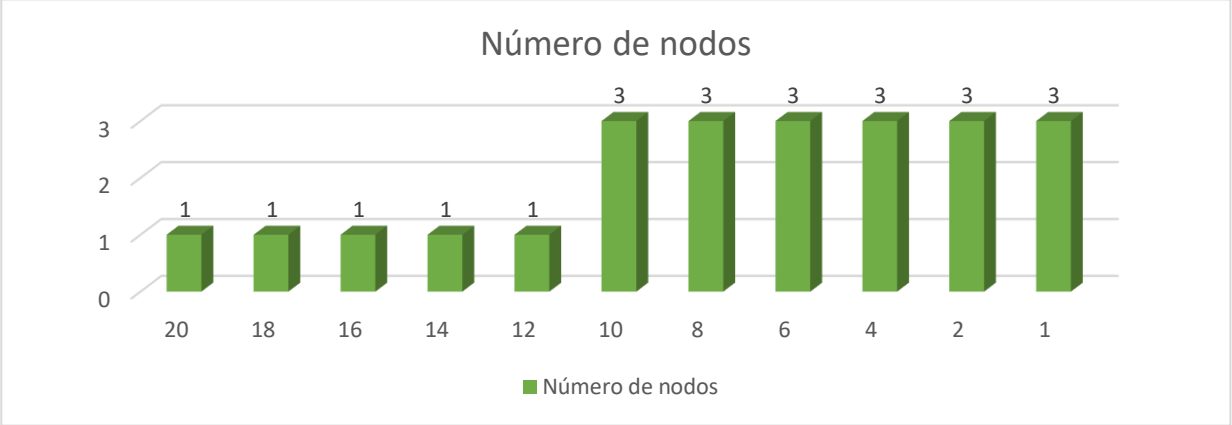
No es comparable. Se puede ver como hay diferencias significativas en el error absoluto medio y en el error cuadrático medio; estos aumentan considerablemente con el modelo reevaluado y también vemos que no hay mucha diferencia entre el coeficiente de correlación de cada modelo, pero al ser el error tan alto, éste no es un modelo fiable y por lo tanto no comparable.

Actividad 13

	20	18	16	14	12	10	8	6	4	2	1
Correlation coefficient	0,9706	0,9693	0,9724	0,9723	0,9754	0,9759	0,9733	0,9762	0,9910	0,9910	0,9910
Mean absolute error	16,5325	17,7690	15,8960	16,0819	13,9871	12,7720	13,0905	11,9865	9,1671	9,1671	9,1671
Root mean squared error	33,3520	34,0135	32,3297	32,3841	30,7462	30,7256	32,1064	30,2816	18,5544	18,5544	18,5544
Relative absolute error (%)	19,8148	21,2969	19,0520	19,2748	16,7640	15,3077	15,6894	14,3663	10,9871	10,9871	10,9871
Root relative squared error (%)	24,0849	24,5626	23,3467	23,3860	22,2032	22,1883	23,1854	21,8676	13,3989	13,3989	13,3989
Número de nodos	1	1	1	1	1	3	3	3	3	3	3
Linear Model (LM) Nodos hoja	2	2	2	2	2	4	4	4	4	4	4
Total number of instances	188	188	188	188	188	188	188	188	188	188	188







Vemos que el modelo con 4 instancias es el más competitivo, al tener un nivel de error bajo, un coeficiente de correlación alto, y el número de nodos que hace que su complejidad computacional sea bajo. A continuación, podemos ver una tabla comparativa del modelo más competitivo seleccionado en la actividad 11 con este modelo seleccionado.

Sin suavizado	Con suavizado
0,9925	0,9910
7,3255	9,1671
17,3694	18,5544
8,7800	10,9871
12,5432	13,3989
3	3
4	4
188	188