ITESO

Universidad Jesuita
de Guadalajara

Proyecto 02

Boston Housing

Minería de Datos — Primavera 2022
ESI3466K
17 de junio de 2023

OLVERA HERNÁNDEZ, ROEBRTO — ib721045

Profesor: Dr. Juan Antonio Vega Fernández

Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO)

San Pedro Tlaquepaque, Jalisco, MX.

Índice

1. Introducción

1.1. El dataset *Boston Housing*

El conjunto de datos *Boston Housing* fue publicado originalmente por Harrison y Rubinfeld [1] en 1978, tomando datos publicados por el Servicio de Censo Poblacional de los Estados Unidos (U.S CS) de la ciudad de Boston, MA.

Inicialmente se publicó el artículo con el objetivo de evaluar el precio de la vivienda en la ciudad en función de la calidad del aire utilizando solo dos variables. Actualmente, la estructura del set constituye de 506 observaciones o *casos* (filas) con 14 variables (columnas), ver Cuadro 1, y es utilizando en gran medida como caso de estudio para *aprendizaje estadístico* [2]-[6].

1.2. Regresión lineal múltiple

Asumimos el siguiente modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{1}$$

donde las variables β_0 y β_1 son constantes desconocidas que representan el *intercepto* y la *pendiente*, respectivamente, también conocidos como *coeficientes* o *parámetros*, y ϵ es el error.

Dadas las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ para los coeficientes del modelo, podemos predecir valores futuros en una cantidad i de variables con:

$$\hat{y} = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 \hat{x}_2 + \ldots + \beta_i \hat{x}_i \tag{2}$$

donde $\hat{y}$ indica una predicción[1] Y respecto a $X = x$.

Sea $\hat{y}_i = \beta_0 + \beta_1 \hat{x}_i$ la predicción Y basado en el valor i-ésimo del valor X. Después, dado el error $e_i = y_i - \hat{y}_i$ podemos definir la *Suma Residual de Cuadrados* (RSS, por sus siglas en inglés):

$$\text{RSS} = \sum_{n}^{i=1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2 \tag{3}$$

La aproximación de mínimos cuadrados usa $\hat{\beta}_0$ y $\hat{\beta}_1$ para minimizar la RSS. Minimizando los valores se puede mostrar que:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

donde $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ son las medias de los datos.

Para evaluar la **precisión general** del modelo de regresión lineal se pueden calcular 3 medidas del error: 1) *Error Estándar de Residuales* (Ec. 5); 2) *R-cuadrada* (Ec. 6); y 3) *Suma Total de Cuadrados* (Ec. 7):

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=i}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

[1] El *circunflejo* representa valores estimados.

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

2. Metodología

Todo el análisis de datos fue realizado usando la plataforma `Orange` con el flujo de trabajo que se muestra en la Figura 1.

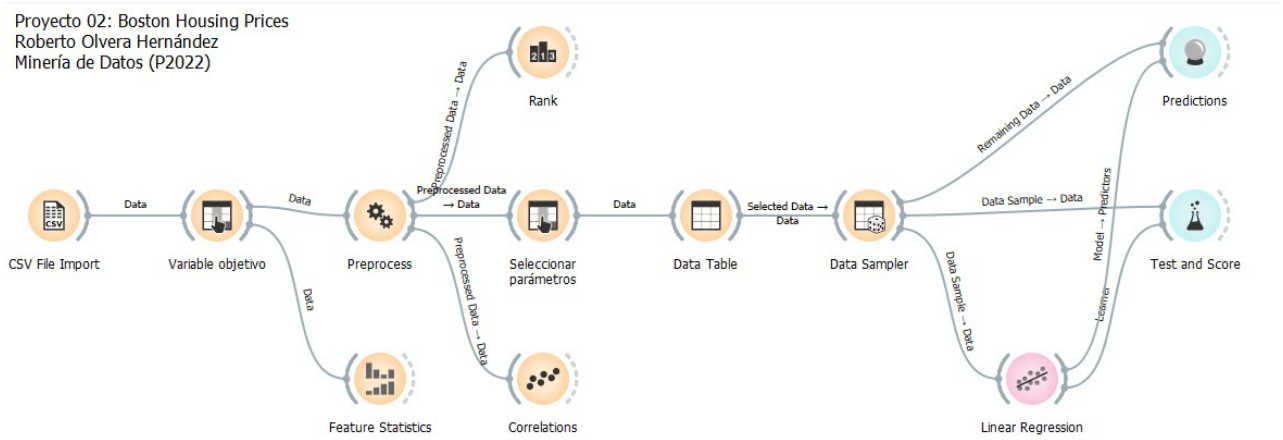


Figura 1: Flujo de trabajo de Modelo Lineal Mútltiple.

2.1. Preprocesamiento de datos

Antes de realizar cualquier arreglo en el conjunto de datos, se definió que la variable `CHAS` sería *categórica* en vez de una *numérica* como se detecta originalmente, ya que son números nominales que indican valores `TRUE` o `FALSE` con 0 y 1. Una vez redefinida, se utilizó el widget `Preprocess` para normalizar los datos en un intervalo [0,1] y eliminar valores *nulos*.

En seguida, se seleccionaron las 5 variables con mejor puntuación para obtener mayor información sobre `MEDV` según el widget `Rank`. Las variables fueron: 1) LSTAT; 2) RM; 3) PTRATIO; 4) AGE; 5) NOX. Estas se mandaron después a un `Data Table` y finalmente con el widget `Data Sampler` se seleccionaron el 70 % y 30 % de los datos para mandarlos al modelo de RLM y el predictor, respectivamente.

2.2. Regresión lineal múltiple

Una vez limpiados los datos, se corrieron las 6 variables en el widget de `Linear regression` sin regularizaciones con el 70 % de los datos obtenidos con el widget anterior y evaluados con `Test \& Score`. El resto de los datos se enviaron a `Predictions`.

3. Resultados y discusión

3.1. Exploración general de los datos

Un gráfico de cajas y bigotes (Figura 2(a)) muestra que la tasa de criminalidad (CRIM) es extremadamente baja, la mayoría cercanos al 0 %, también hay bastantes *outliers* que pueden ser identificados. También podemos observar que muy pocas zonas residenciales tienen lotes mayores a 25,000ft^2 (ZN).

Otra característica interesante que se encuentra en valores extremos es la proporción de personas negras (B). Hay que tomar en cuenta que la función $1000 \cdot (B_k - 0{,}63)^2 \propto B_k$, así, podemos observar que en la mayoría de las zonas residenciales cuentan con una población mayormente negra ($65{,}94 \pm 0{,}56$ %).

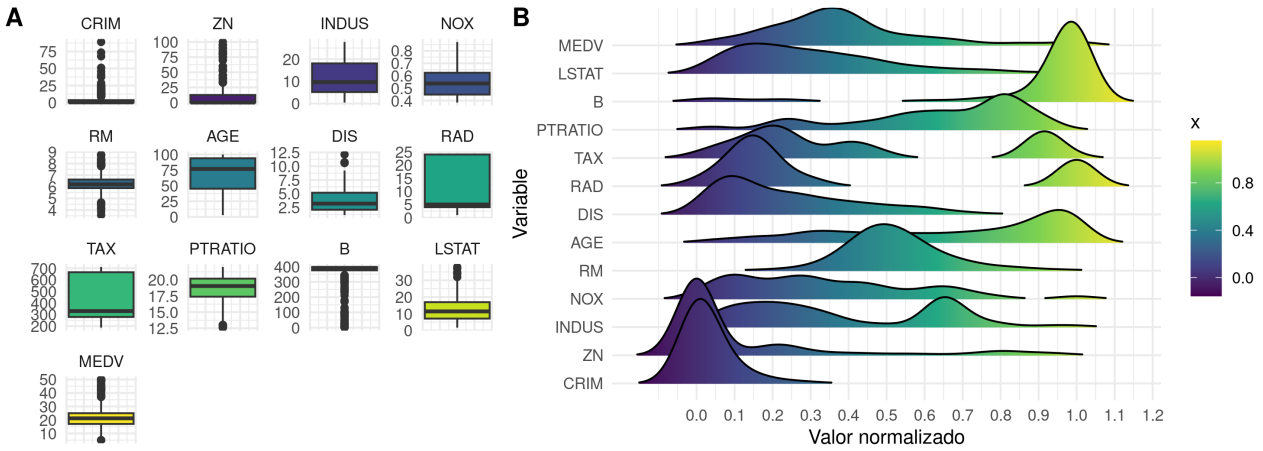


Figura 2: (A) Gráfico de cajas y bigotes de valores antes de normalización; (B) Gráfico de densidad después de normalización.

Observando el gráfico de densidad (Figura 2(b)) también podemos ver que el número de cuartos por hogar (RM) está justamente a la mitad de sus datos normalizados, en $6{,}27 \pm 0{,}70$ cuartos. Hay 64 residencias con un promedio mayor a 7 cuartos, y solamente 16 para mayores a 8 (Figura 3), esto quiere decir que el 3.16 % de las residencias en Boston podrían considerarse de clase alta.

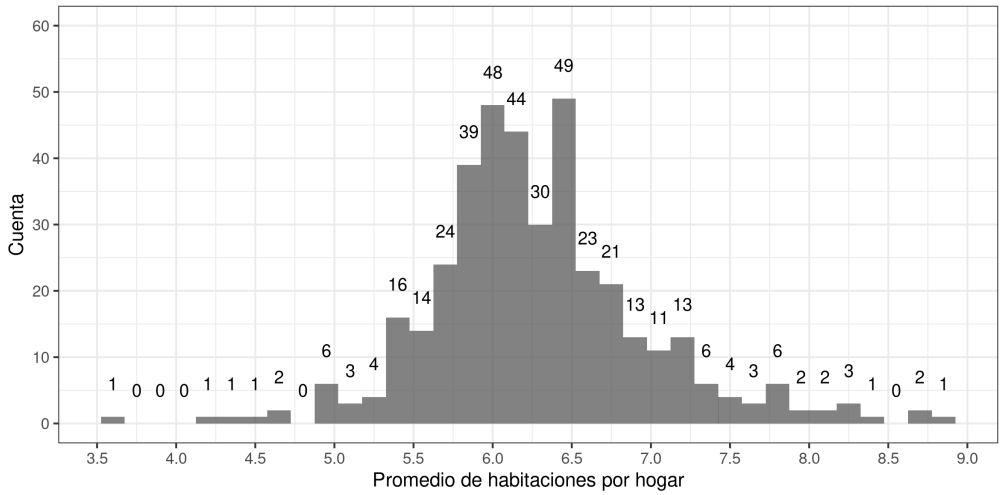
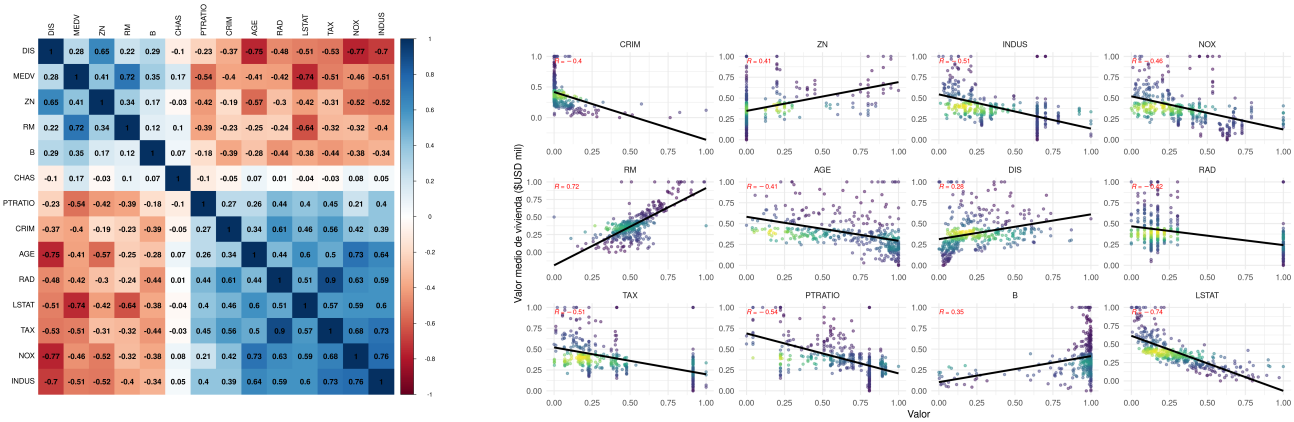


Figura 3: Histograma de cuenta para promedio de cuartos por residencia.

3.2. Correlaciones

Un cuadro de correlaciones (Figura 4(a)) mostró que las variables LSTAT y RM tienen una alta relación con MEDV, esto va de acuerdo con los resultados de Rank y en los gráficos de dispersión (Figura 4(b)) se ve claramente una tendencia lineal.

Estos resultados tienen sentido, ya que a medida que una casa con más cuartos costará más y menos es la cantidad de personas de clase baja que pueden pagarla.



(a) **(b)**

Figura 4: Gráficos descriptivos del modelo de Regresión Lineal Múltiple (RLM) (A) Matriz de correlación; (B) Gráfica de dispersión de todas las características numéricas del conjunto de datos en función de MEDV.

En el mismo cuadro podemos observar que también existe una relación entre LSTAT-RM (Figura 5), con esto y la información anterior podemos concluir que a mayor número de cuartos, mayor será el costo de vivienda, por lo que una menor cantidad de personas de clase baja podrá pagar una de estas casas.

Sobre las demás variables que fueron seleccionadas con Rank, las correlaciones según la matriz no superaron el valor de ±0,5, así que podemos predecir que no se tomarán en cuenta para el modelo posteriormente.

Hay buenas correlaciones con la tasa de criminalidad per capita (CRIM) con algunas variables (ver Cuadro 1), como el índice de accesibilidad a carreteras radiales (RAD, corr=0.61) y la tasa de impuesto (TAX, corr=0.56), pero, a pesar de ser buenos resultados, no tienen ninguna relación entre sí.

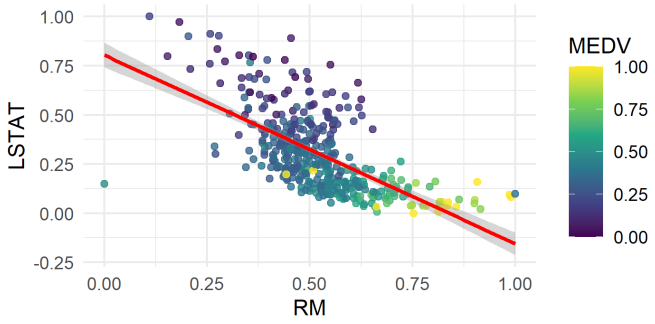


Figura 5: Relación normalizada entre el promedio de habitaciones con el porcentaje de población de bajos recursos.

3.3. Regresión Lineal Múltiple

A continuación se muestran los resultados presentados por el widget de `Test & Score` utilizando distintas combinaciones de parámetros, como *Best Subset Selection* usando el 70 % de los datos como método de entrenamiento. Primero haciendo una regresión lineal para cada parámetro ($n = 1$) en función de `MEDV`, y después usando todas ($n = 13$). Finalmente se corrieron 4 combinaciones distintas para $n = \{2, 3, 4\}$ variables, dando un total de 26 combinaciones diferentes, ver Cuadro 2.

Se utilizó el método de *Best Subset Selection*[2] para encontrar la combinación de variables que mejor describieran al modelo. Los incrementos de R^2 a partir de 4 parámetros fueron menores a 0.1, por lo que es seguro asumir que este es el número de variables necesarias para predecir el modelo. Haciendo referencia en el Cuadro 2 vemos que esa combinación de variables fueron `RM`+`PTRATIO`+`B`+`LSTAT` (ID = 22).

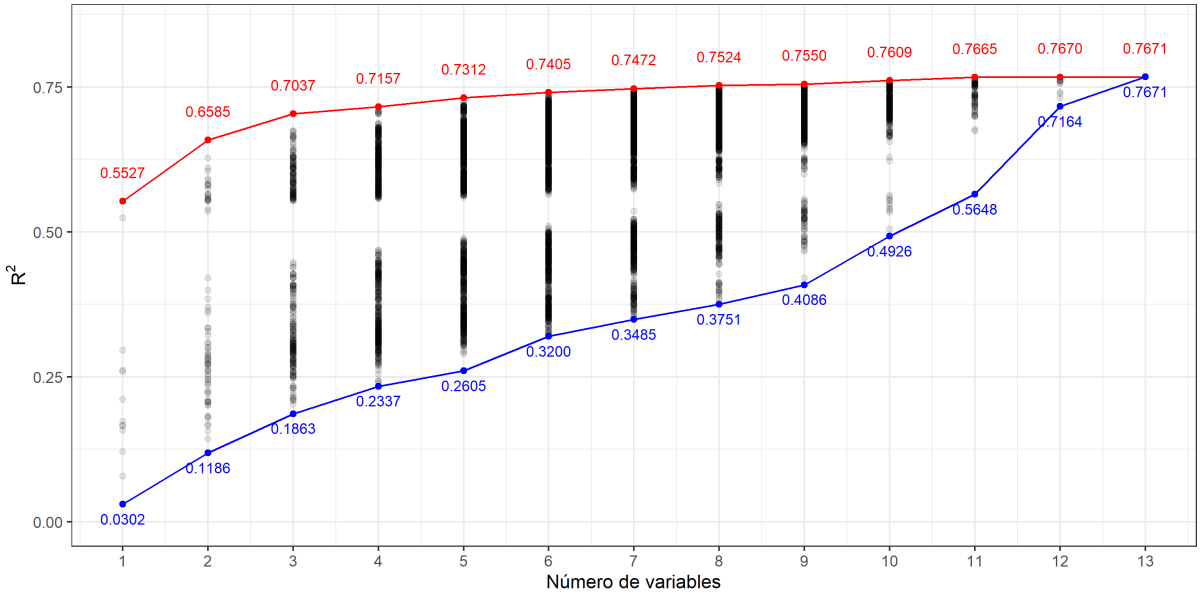


Figura 6: Best Subset Selection (BSS) de todas las variables en *Boston Housing*.

Vemos que `LSTAT` y `RM` están presentes en la combinación, esto va a acorde con lo que se vió en el cuadro de correlaciones (Figura 4(a)) y las tendencias en el gráfico de dispersión (Figura 4(b)), esto tiene sentido puesto que son variables que tienen una alta correlación con `MEDV`. La correlación negativa con `PTRATIO` indica que estas residencias se encuentran cerca de escuelas con una mayor cantidad de alumnos que de profesores; podríamos deducir que se trata de escuelas públicas donde existe una alta demanda de estudiantes y poca oferta de personal académico. Sobre la variable `B`, dado el contexto, parece ser solamente un parámetro *circunstancial*, ya que—como se mencionó anteriormente—la mayoría de la población en todas las ciudades es negra, es decir, que las casas de alto valor como de bajo valor están ocupadas por esta población.

4. Conclusiones

La regresión lineal es una herramienta de *statistical learning* sumamente útil para mostrar la relación de variables numéricas entre sí, y nos ayuda a poder formar una narrativa que nos ayude a explicar diversos

[2]Se realizó de forma independiente del proyecto en *R* con el paquete *olss* con el único fin de presentar la gráfica de la Figura 6.

fenómenos a partir de sus resultados.

En este proyecto se utilizó el software de `Orange` para jugar con los 13 parámetros del conjunto de datos de *Boston Housing* para encontrar la mejor combinación que conteste la pregunta: ¿Qué es lo que afecta el valor mediano de hogares en Boston, MA?

De acuerdo con un Best Subset Selection manual, una regresión lineal múltiple que toma en cuenta todas las variables es el modelo que mejor describe el comportamiento de `MEDV`, sin embargo, solamente 4 a 6 variables son necesarias para describirlo. El precio medio de vivienda en Boston está dictado principalmente por el tamaño de la casa (`RM`) y, a su vez, existe una relación directa con el estrato social que tiene acceso a pagarlas (`LSTAT`). Las casas con menor valor son las que se encuentran en zonas con una baja proporción de alumno-maestro (`PTRATIO`), y las personas negras (`B`) tienen acceso a casas de alto valor como de bajo valor.

Referencias

[1] D. Harrison y D. L. Rubinfeld, «Hedonic housing prices and the demand for clean air,» *Journal of Environmental Economics and Management*, vol. 5, págs. 81-102, 1 mar. de 1978, ISSN: 0095-0696. DOI: 10.1016/0095-0696(78)90006-2.

[2] A. Al Bataineh y D. Kaur, «A comparative study of different curve fitting algorithms in artificial neural network using housing dataset,» en *NAECON 2018-IEEE National Aerospace and Electronics Conference*, IEEE, 2018, págs. 174-178.

[3] M. Shahhosseini, G. Hu y H. Pham, «Optimizing ensemble weights for machine learning models: A case study for housing price prediction,» en *INFORMS international conference on service science*, Springer, 2019, págs. 87-97.

[4] S.-W. Lin, K.-C. Ying, S.-C. Chen y Z.-J. Lee, «Particle swarm optimization for parameter determination and feature selection of support vector machines,» *Expert systems with applications*, vol. 35, n.º 4, págs. 1817-1824, 2008.

[5] T. Ryffel, A. Trask, M. Dahl y col., «A generic framework for privacy preserving deep learning,» *arXiv preprint arXiv:1811.04017*, 2018.

[6] R. Timofeev, «Classification and regression trees (CART) theory and applications,» *Humboldt University, Berlin*, vol. 54, 2004.

A. Descripción de variables

ID	Parámetro	Tipo de variable	Descripción.
1	CRIM	Numérica	Tasa de crimen *per capita* por ciudad.
2	ZN	Numérica	Proporción de áreas residenciales con lotes mayores a 25mil ft^2.
3	INDUS	Numérica	Proporción de acres comerciales no minoristas por ciudad.
4	CHAS	Numérica / Categórica	Variable artificial del Río Charles (0 = No limita; 1 = Limita).
5	NOX	Numérica	Concentración de óxidos nítricos (partes por 10millones).
6	RM	Numérica	Número promedio de habitaciones por vivienda.
7	AGE	Numérica	Proporción de unidades habitadas construidas antes de 1940.
8	DIS	Numérica	Distancias ponderadas de 5 centros de empleo en Boston.
9	RAD	Numérica	Índice de accesibilidad a carreteras radiales.
10	TAX	Numérica	Tasa de impuestos de propiedad completa por $10,000 USD.
11	PTRATIO	Numérica	Proporción alumno-maestro por ciudad.
12	B	Numérica	$1000(B_k - 0{,}63)^2$; donde B_k es la proporción de población negra en la ciudad.
13	LSTAT	Numérica	Percentaje de personas que pertenecen a la clase baja en la ciudad.
14	MEDV	Numérica	Valor mediano de propiedad ocupada en miles de USD ($USDmil).

Cuadro 1: Tabla de descripción de variables en el conjunto de datos *Boston Housing*.

B. Tablas de resultados

ID	n	Predictores	MSE	RMSE	MAE	R^2
1	1	LSTAT	40.4230	6.3580	4.5910	0.5527
2	1	RM	46.0536	6.7863	4.5399	0.5241
3	1	PTRATIO	66.9565	8.1827	5.9830	0.2957
4	1	INDUS	72.1585	8.4946	6.0492	0.2609
5	1	TAX	65.3017	8.0809	5.7910	0.2589
6	1	NOX	68.3234	8.2658	5.9955	0.2107
7	1	RAD	73.6673	8.5830	6.1894	0.1736
8	1	AGE	67.8550	8.2374	5.6922	0.1660
9	1	ZN	70.5212	8.3977	5.9814	0.1655
10	1	CRIM	70.6842	8.4074	6.0218	0.1578
11	1	B	76.9584	8.7726	6.2886	0.1206
12	1	DIS	79.9044	8.9389	6.3977	0.0781
13	1	CHAS	82.2370	9.0685	6.7172	0.0302
14	2	RM LSTAT	33.3145	5.7719	4.1470	0.6585
15	2	PTRATIO LSTAT	36.2053	6.0171	4.2816	0.6269
16	2	RM TAX	38.7602	6.2258	4.0774	0.6095
17	2	RM PTRATIO	40.3046	6.3486	4.2659	0.6045
18	3	RM PTRATIO LSTAT	30.6197	5.5335	3.9051	0.7037
19	3	RM TAX LSTAT	33.3924	5.7786	4.0956	0.6740
20	3	RM B LSTAT	32.4562	5.6970	4.0444	0.6739
21	3	CHAS RM LSTAT	29.3583	5.4183	3.8614	0.6721
22	4	RM PTRATIO B LSTAT	30.0252	5.4795	3.7939	0.7157
23	4	CHAS RM PTRATIO LSTAT	24.6675	4.9666	3.4114	0.7135
24	4	RM DIS PTRATIO LSTAT	29.8113	5.4600	3.8567	0.7095
25	4	CRIM RM PTRATIO LSTAT	22.8367	4.7788	3.3647	0.7082
26	13	CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT	24.5508	4.9549	3.4464	0.7671

Cuadro 2: Resultados de *Test & Score* para distintas combinaciones de parámetros del conjunto de datos.