Ejercicio 1

Contexto y significado de cada atributo del dataset:

1- CRIM: tasa de criminalidad per cápita por ciudad.
2- ZN: proporción de terrenos residenciales zonificados para lotes mayores a 25,000 pies cuadrados.
3- INDUS: proporción de acres de negocios no minoristas por ciudad.
4- CHAS: variable ficticia del río Charles (1 si la zona limita con el río; 0 en caso contrario).
5- NOX: concentración de óxidos nítricos (partes por cada 10 millones).
6- RM: número promedio de habitaciones por vivienda.
7- AGE: proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
8- DIS: distancias ponderadas a cinco centros de empleo de Boston.
9- RAD: índice de accesibilidad a carreteras radiales.
10- TAX: tasa de impuestos a la propiedad por cada $10,000.
11- PTRATIO: ratio de alumnos por maestro por ciudad.
12- B: 1000(Bk - 0.63)^2, donde Bk es la proporción de personas de raza negra por ciudad.
13- LSTAT: porcentaje de población de menor estatus socioeconómico.
14- MEDV: media de las viviendas ocupadas por sus propietarios en miles de dólares.

Tipos de datos y rangos:

Todos los tipos de datos son numéricos “reales regulares”.

Los rangos varían en todas las columnas, ejemplo

ZN: 0 – 100
INDUS: 0.460 – 27.740
TAX: 233 – 666

Distribuciones y outliers

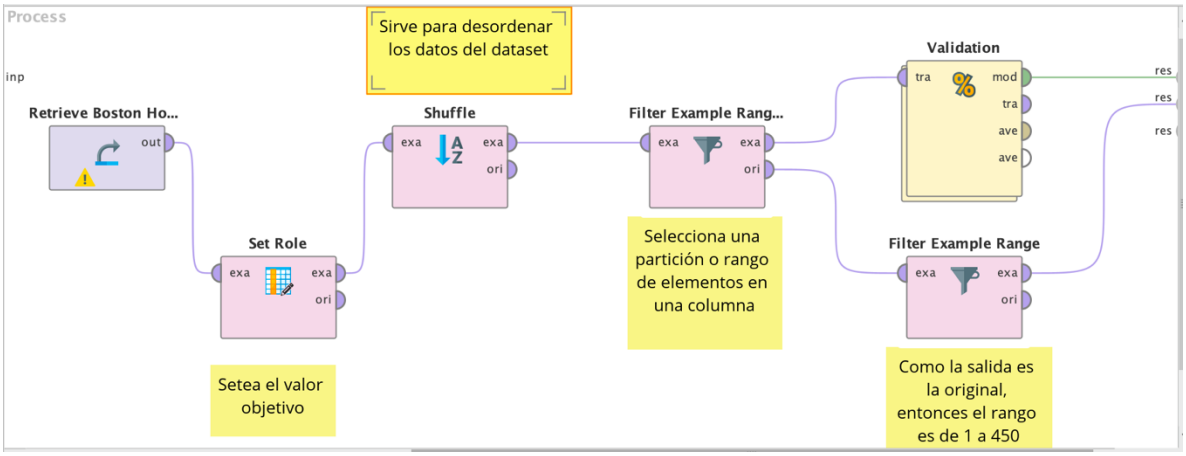
Hay varios otliers que pueden interferir en nuestros algrotimos, por ejemplo, en la columna de “Age” en algunos casos las edades están definidas con el siguiente formato: 98.800, 29.700, 83.400, etc.
También en la tabla “CRIM” la mayoría de los atributos se presentan entre cero y uno, pero hay valores fuera de rango de hasta 6 cifras.

¿Cuál es la variable de salida?

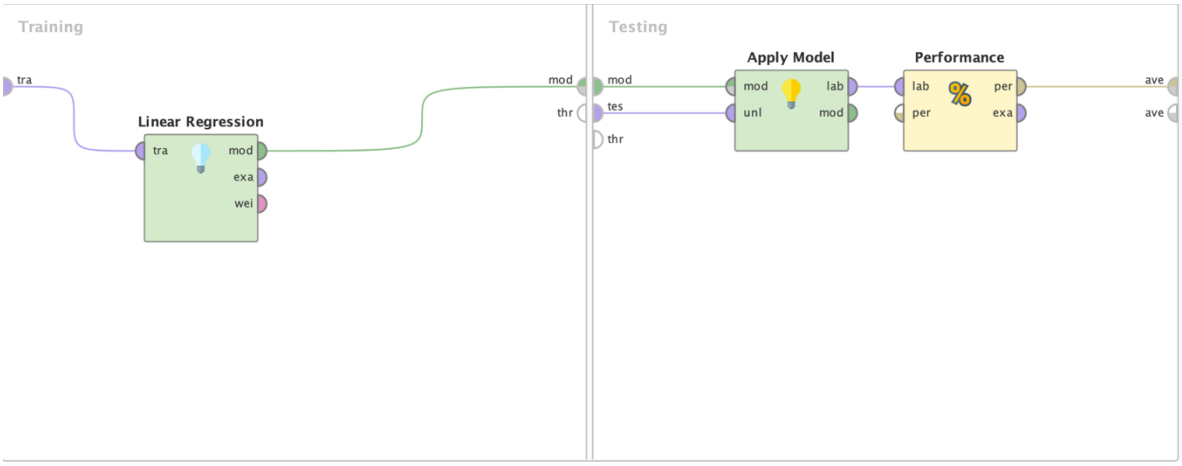
La variable de salida es “MEDV” ya que nos indica el valor medio de la casa.

Ejercicio 2

Construcción del modelo



Validation



Algunos de los resultados del modelo

Row No.	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
1	20.900	0.035	80	3.640	0	392	5876	19.100
2	25.300	0.141	0	6.910	0	448	6169	6.600
3	22.600	0.084	0	4.050	0	0.510	5859	68.700
4	20.700	0.091	20	6.960	1	464	5.920	61.500
5	26.200	0.191	22	5.860	0	431	6718	17.500
6	15.200	0.151	0	27.740	0	609	5454	92.700
7	24.300	537	0	6.200	0	504	5981	68.100
8	19.400	214918	0	19.580	0	871	5709	98.500
9	7.500	108342	0	18.100	0	679	6782	90.800
10	22	0.113	30	4.930	0	428	6897	54.300
11	22	0.110	0	11.930	0	573	6794	89.300
12	30.500	0.069	45	3.440	0	437	6739	30.800
13	19.900	383684	0	18.100	0	0.770	6251	91.100
14	50	489822	0	18.100	0	631	4.970	100
15	27.500	0.624	0	6.200	1	507	6879	77.700
16	21	0.475	0	9.900	0	544	6113	58.800
17	14.400	0.254	0	6.910	0	448	5399	95.300

Shuffle: Sirve para mezclar los datos

Filter Example Range: Sirve para tomar un rango de datos.

¿Qué parámetros podemos variar en el operador "Linear Regression"? Feature selection, eliminate colinear features, min tolerance, use bias, ridge

Feature selection: Este parámetro decide si se debe seleccionar automáticamente un subconjunto de características relevantes para el modelo. Esto ayuda a evitar el sobreajuste y a reducir la complejidad del modelo.

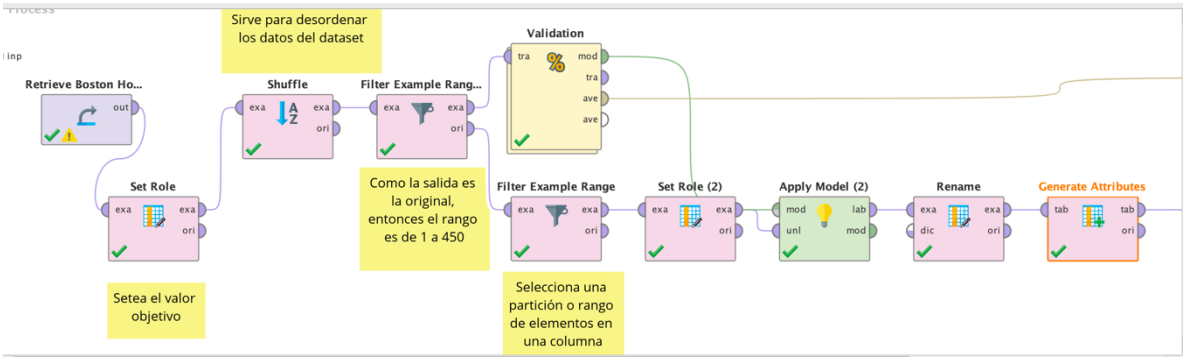
Ejercicio 3

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
CRIM	-0.000	0.000	-0.108	0.869	-2.423	0.016	**
ZN	0.038	0.015	0.097	0.846	2.576	0.010	**
INDUS	-0.012	0.069	-0.009	0.650	-0.170	0.865	
CHAS	3.392	1.071	0.091	0.984	3.166	0.002	***
NOX	-0.001	0.001	-0.017	0.982	-0.547	0.584	
RM	0.000	0.000	0.093	0.971	3.273	0.001	***
AGE	0.020	0.015	0.061	0.755	1.392	0.165	
DIS	-0.000	0.000	-0.160	0.882	-4.229	0.000	****
RAD	0.398	0.082	0.372	0.707	4.829	0.000	****
TAX	-0.017	0.004	-0.308	0.703	-3.790	0.000	****
PTRATIO	-1.060	0.146	-0.247	0.801	-7.258	0.000	****
B	0.008	0.003	0.082	0.899	2.576	0.010	**
LSTAT	-0.795	0.052	-0.617	0.608	-15.411	0	****
(Intercept)	50.972	3.361	?	?	15.167	0	****

LinearRegression

```
- 0.000 * CRIM
+ 0.038 * ZN
- 0.012 * INDUS
+ 3.392 * CHAS
- 0.001 * NOX
+ 0.000 * RM
+ 0.020 * AGE
- 0.000 * DIS
+ 0.398 * RAD
- 0.017 * TAX
- 1.060 * PTRATIO
+ 0.008 * B
- 0.795 * LSTAT
+ 50.972
```

Ejercicio 4



Row No.	MEDV	predictedM...	residuals	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS
1	20.900	24.376	3.476	0.035	80	3.640	0	392	5876	19.100	92203
2	25.300	26.603	1.303	0.141	0	6.910	0	448	6169	6.600	57209
3	22.600	28.166	5.566	0.084	0	4.050	0	0.510	5859	68.700	27019
4	20.700	23.553	2.853	0.091	20	6.960	1	464	5.920	61.500	39175
5	26.200	24.731	-1.469	0.191	22	5.860	0	431	6718	17.500	78265
6	15.200	10.490	-4.710	0.151	0	27.740	0	609	5454	92.700	18209
7	24.300	25.635	1.335	537	0	6.200	0	504	5981	68.100	36715
8	19.400	21.959	2.559	214918	0	19.580	0	871	5709	98.500	16232
9	7.500	10.009	2.509	108342	0	18.100	0	679	6782	90.800	18195
10	22	25.755	3.755	0.113	30	4.930	0	428	6897	54.300	63361
11	22	25.359	3.359	0.110	0	11.930	0	573	6794	89.300	23889
12	30.500	30.426	-0.074	0.069	45	3.440	0	437	6739	30.800	64798
13	19.900	20.832	0.932	383684	0	18.100	0	0.770	6251	91.100	22955
14	50	26.886	-23.114	489822	0	18.100	0	631	4.970	100	13325
15	27.500	31.349	3.849	0.624	0	6.200	1	507	6879	77.700	32721
16	21	21.915	0.915	0.475	0	9.900	0	544	6113	58.800	40019
17	14.400	12.012	-2.388	0.254	0	6.910	0	448	5399	95.300	5.870

Se destaca que los outliers afectan el resultado de nuestra predicción