Ejercicio 1

Nombre	Significado	Tipos de datos	Rangos	Distribuciones	Outliers
CRIM	Tasa de crimen per cápita según ciudad	Real-numérica	0-9,967	Distribución sesgada a la derecha	
ZN	Proporción de zona residencial con lotes sobre 25 mil metros cuadrados	Real-numérica	0-100	No tiene distribución conocida	
INDUS	proporción de acres de negocios no minoristas por ciudad	Real-numérica	0-27,740	No tiene distribución conocida	
CHAS	Limita con el río Charles	binomial	0 y 1		
NOX	Concentración de óxido nítrico (partes por 10 millones)	Real-numérica	0,385-0,871	No tiene distribución conocida	
RM	Número promedio de cuartos de la vivienda	Real-numérica	3,561-8,780	No tiene distribución conocida	
AGE	Proporción de unidades ocupadas por sus dueños ocupadas antes de 1940	Real-numérica	2,9-100	Polinomial	
DIS	Distancia ponderada a 5 centros de empleo de Boston	Real-numérica	1,130-12,127	Binomial negativa	
RAD	Índice de accesibilidad a autopistas radiales	Real-numérica	1-24	No tiene distribución conocida	Valores 666 desaparecen al filtrar missing en atributo a estimar
TAX	Valor de impuestos de la vivienda cada $10.000	Real-numérica	187-711	No tiene distribución conocida	
PTRATIO	Ratio alumno/maestro por ciudad	Real-numérica	12,6-22	No tiene distribución conocida	
B	1000(Bk - 0.63)^2 donde Bk es la proporción de negros por ciudad	Real-numérica	0,320-396,9	Polinomial	
LSTAT	% de población de clase baja	Real-numérica	1,73-34,410	Similar a una distribución chi-cuadrado	
MEDV	Valor medio de las viviendas ocupadas en miles de dólares	Real-numérica	6,3-50	Sesgada a la derecha	

Se identifican 54 valores missing en la variable a predecir, se eliminan esos valores obteniendo la tabla anteriormente expuesta.

La variable de salida es MEDV, el modelo busca predecir el valor promedio de las viviendas.

¿Por qué aplicamos shuffle luego de recuperar el dataset?
Para aleatorizar los componentes evitando cambios abruptos en los coeficientes y valores que toman los parámetros de estimación.

¿Cómo funciona el operador "filter examples range"?
Filtra el rango de filas establecido, desde el mínimo número de fila al máximo.

¿Qué hace "feature selection"? ¿Cómo?
Selecciona atributos para evitar que mi modelo contenga una cantidad extensa de atributos que aumentan su complejidad, dificultan la comprensión y además que pueden estar correlacionados entre sí.

¿Cómo afectan "eliminate colinear features" y "use bias"?
Si se selecciona "eliminate colinear features" el algoritmo intenta eliminar durante la regresión atributos correlacionados.
Use bias identifica si debe calcularse un valor de intercepto (B0 en el análisis teórico).

Ejercicio 3

Tomar nota de los resultados obtenidos para cada predictor.

LinearRegression

```
  0.145 * CRIM
+ 0.055 * ZN
+ 0.029 * INDUS
+ 3.120 * CHAS
- 8.454 * NOX
+ 4.472 * RM
- 0.004 * AGE
- 1.513 * DIS
+ 0.164 * RAD
- 0.010 * TAX
- 0.767 * PTRATIO
+ 0.011 * B
- 0.692 * LSTAT
+ 25.685
```

¿Cuáles no parecen ser muy significativos?

Según valores de R^2 los que parecieran no tener tanta incidencia son:

ZN: Proporción de zona residencial con lotes sobre 25 mil metros cuadrados

INDUS: proporción de acres de negocios no minoristas por ciudad

AGE: Proporción de unidades ocupadas por sus dueños ocupadas antes de 1940

TAX: Valor de impuestos de la vivienda cada $10.000

B: 1000(Bk - 0.63)^2 donde Bk es la proporción de negros por ciudad

Según la significación del modelo considerando el p valor los menos significativos:

NOX

TAX

B

CHAS

Para el parámetro "feature selection, greedy":

LinearRegression

```
  0.058 * ZN
+ 3.090 * CHAS
+ 4.485 * RM
- 1.341 * DIS
+ 0.180 * RAD
- 0.011 * TAX
- 0.692 * PTRATIO
+ 0.012 * B
- 0.724 * LSTAT
+ 19.586
```

Predictores que se han eliminado del modelo:

CRIM

INDUS

NOX

AGE

Operador PERFORMANCE:

• Utilizar "squared correlation" – este es el indicador R^2 visto en clase

  o Tomar nota de los valores obtenidos sin y con feature selection

• Observar también los valores del error medio cuadrático

Con greedy y R^2:

LinearRegression

```
  0.047 * ZN
+ 1.808 * CHAS
- 9.155 * NOX
+ 4.896 * RM
- 1.387 * DIS
+ 0.202 * RAD
- 0.010 * TAX
- 0.722 * PTRATIO
+ 0.012 * B
- 0.608 * LSTAT
+ 20.242
```

root_mean_squared_error

```
root_mean_squared_error: 3.755 +/- 0.000
```

Sin greedy:

LinearRegression

```
  0.003 * CRIM
+ 0.046 * ZN
- 0.007 * INDUS
+ 1.820 * CHAS
- 8.690 * NOX
+ 4.927 * RM
- 0.004 * AGE
- 1.403 * DIS
+ 0.198 * RAD
- 0.010 * TAX
- 0.715 * PTRATIO
+ 0.012 * B
- 0.601 * LSTAT
+ 19.956
```

root_mean_squared_error

```
root_mean_squared_error: 3.682 +/- 0.000
```

Se observan cambios en los coeficientes de algunos atributos al aplicar greedy o no. Al ser más exigente el primero elimina algunos atributos del modelo en cuanto a su significación y eso repercute en el ajuste leve a la baja de algunos coeficientes mientras que otros suben levemente.

A su vez aumenta el RSME en la predicción al ser más exigente.

El valor de R^2 es: 0.814, lo que implica que un 81,4% de la variación de la variable de salida y es explicada por el modelo de regresión.
Los atributos **más significativos** son:

RM

DIS

PTRATIO

LSTAT

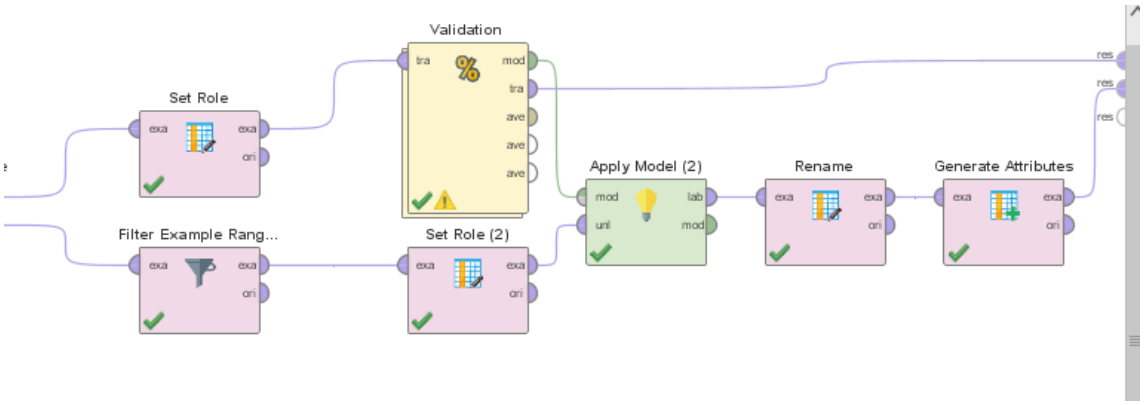
ZN

Que a su vez reflejan un p valor bajo.
Para mejorar el modelo habría que quitar del mismo las variables no significativas para el modelo o fuertemente correlacionadas entre sí.
Podría analizarse la **matriz de correlación**:

Attribut...	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
CRIM	1	-0.281	0.574	0.050	0.637	-0.142	0.448	-0.462	0.898	0.826	0.319	-0.413	0.425
ZN	-0.281	1	-0.514	-0.060	-0.501	0.307	-0.556	0.656	-0.267	-0.269	-0.364	0.150	-0.411
INDUS	0.574	-0.514	1	0.103	0.739	-0.365	0.606	-0.669	0.513	0.673	0.317	-0.317	0.565
CHAS	0.050	-0.060	0.103	1	0.134	0.077	0.123	-0.141	0.057	0.017	-0.100	0.013	-0.009
NOX	0.637	-0.501	0.739	0.134	1	-0.265	0.707	-0.746	0.542	0.615	0.103	-0.358	0.537
RM	-0.142	0.307	-0.365	0.077	-0.265	1	-0.188	0.139	-0.096	-0.215	-0.334	0.108	-0.607
AGE	0.448	-0.556	0.606	0.123	0.707	-0.188	1	-0.720	0.359	0.427	0.193	-0.224	0.573
DIS	-0.462	0.656	-0.669	-0.141	-0.746	0.139	-0.720	1	-0.388	-0.444	-0.152	0.234	-0.424
RAD	0.898	-0.267	0.513	0.057	0.542	-0.096	0.359	-0.388	1	0.873	0.387	-0.353	0.310
TAX	0.826	-0.269	0.673	0.017	0.615	-0.215	0.427	-0.444	0.873	1	0.385	-0.367	0.411
PTRATIO	0.319	-0.364	0.317	-0.100	0.103	-0.334	0.193	-0.152	0.387	0.385	1	-0.090	0.303
B	-0.413	0.150	-0.317	0.013	-0.358	0.108	-0.224	0.234	-0.353	-0.367	-0.090	1	-0.291
LSTAT	0.425	-0.411	0.565	-0.009	0.537	-0.607	0.573	-0.424	0.310	0.411	0.303	-0.291	1

Y decidir pruebas de feature selection seleccionando diferentes atributos para incorporar al modelo.
Se observan correlaciones entre atributos como ser NOX-AGE y NOX-DIS una en sentido positivo y otra en sentido inverso. También entre AGE y DIS lo que podría sugerir una buena opción eliminar alguna de estas tres del modelo y probar resultados. También hay fuerte asociación de CRIM con RAD y TAX respectivamente.

Ejercicio 4



Observar las estadísticas de los residuos. ¿Qué se destaca?

Se observan las distancias entre valores observados y esperados, se observa que el rango de variación de los residuos va de -16,8 a 11,6 es decir las predicciones para esas instancias se alejaron en casi 17 unidades por debajo de la realidad a un máximo de 12 unidades, esto podría analizarse en el contexto de que el rango de MED real va de 7 a 50, por lo que un 16,8 representa una cantidad importante de desvío del predictor. Por otro lado cuando observamos la media de los residuos la misma es 0,124 lo que resulta aceptable.