Sigues avanzando en técnicas muy divertidas.

Ahora aplicará técnicas de reducción dimensional

Vamos a trabajar con la tabla de datos “abalone”.

IMPORTA LOS DATOS ABALONE

Para no complicarnos puedes leer la tabla de datos de esta forma:

```
data(abalone)
dataset <- abalone
# split input and output
x <- dataset[,1:8]
y <- dataset[,9]
```

Este es un problema de regresión donde las “x” son las variables de entrada y las “y” las variables de salida.

Lo primero que vamos a hacer es trabajar con la regresión lineal stepwise y calcularemos el rango de importancia de las variables.

Esto nos ayudará a entender la influencia de las entradas en la salida.

¡Después crearemos un modelo predictivo de regresión!

APLICANDO EL PAIRWISE REGRESSION

Aplica el algoritmo de pairwise en la regresión para identificar de las 8 variables de entrada cuáles son las más relevantes conjuntamente.

Copia el modelo inicial poniendo las 8 variables de entrada.
Y luego copia el modelo refinado después del pairwise

Todas las variables:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.89464    0.29157  13.358  < 2e-16 ***
SexI         -0.82488    0.10240  -8.056 1.02e-15 ***
SexM          0.05772    0.08335   0.692    0.489
Length       -0.45834    1.80912  -0.253    0.800
Diameter     11.07510    2.22728   4.972 6.88e-07 ***
Height       10.76154    1.53620   7.005 2.86e-12 ***
```

```
`Whole weight`     8.97544    0.72540  12.373  < 2e-16 ***
`Shucked weight` -19.78687    0.81735 -24.209  < 2e-16 ***
`Viscera weight` -10.58183    1.29375  -8.179 3.76e-16 ***
`Shell weight`     8.74181    1.12473   7.772 9.64e-15 ***


MSE = 4.802664
AIC = 6574.427
```

Modelo Refinado:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.87038    0.27536  14.056  < 2e-16 ***
SexI          -0.82644    0.10220  -8.087 7.96e-16 ***
SexM           0.05755    0.08333   0.691     0.49
Diameter      10.56951    0.98887  10.688  < 2e-16 ***
Height        10.74911    1.53525   7.002 2.94e-12 ***
`Whole weight`     8.97751    0.72528  12.378  < 2e-16 ***
`Shucked weight` -19.80258    0.81490 -24.301  < 2e-16 ***
`Viscera weight` -10.61279    1.28782  -8.241 2.27e-16 ***
`Shell weight`     8.75078    1.12405   7.785 8.73e-15 ***

MSE =  4.802738
AIC = 6572.491
```

AIC ambos modelos:

```
     Step Df  Deviance Resid. Df Resid. Dev      AIC
1        NA       NA      4167   20060.73 6574.427
2 - Length  1 0.3089962      4168   20061.04 6572.491
```

¿Qué variables te han quedado?

Sé libre de contar lo que ves ...

Me han quedado todos los datos salvo la variable Length, que tenía un P-valor de 0.8. El factor masculino de la variable Sex, tiene un P-valor alto, pero no fue eliminado, imagino que porque forma parte de la variable Sex. El factor femenino de la variable Sex no aparece reflejado en ninguno de los dos modelos, desconozco el por qué.

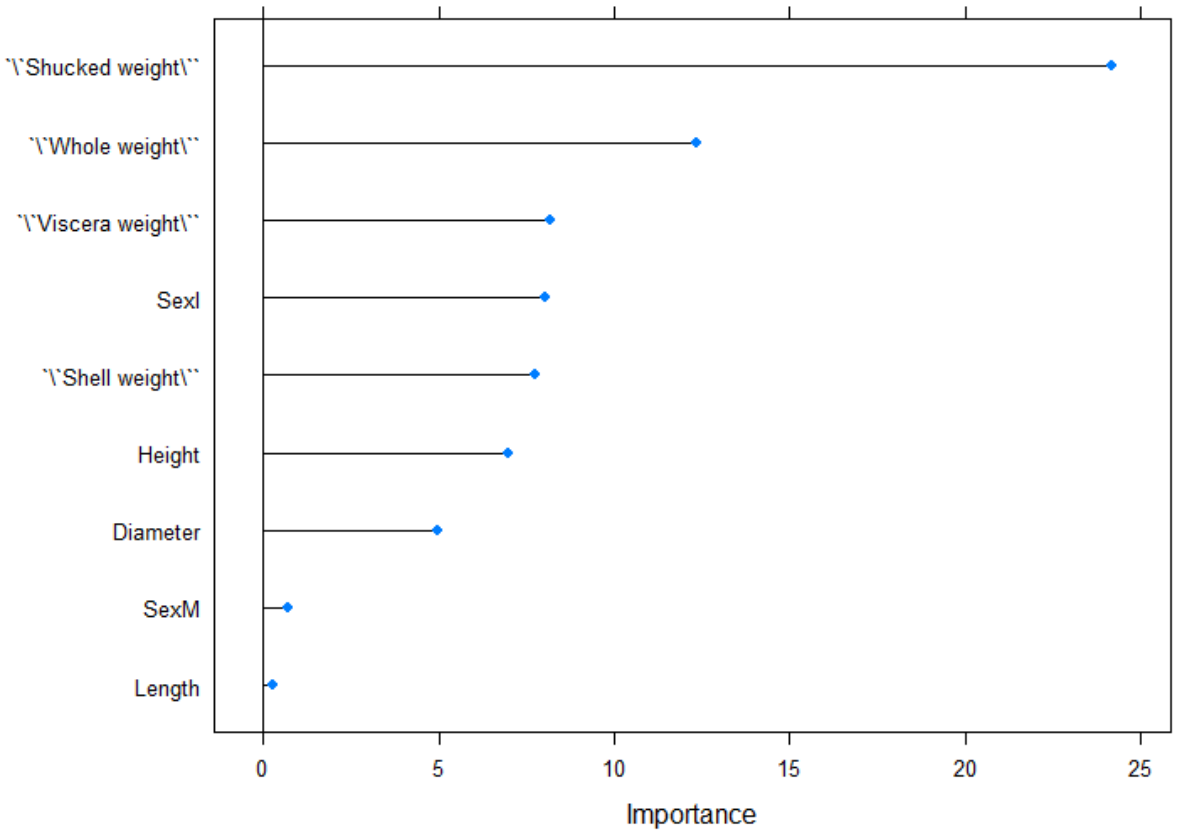
Al sacar la accuracy de ambos modelos, la mejor (menor) es la del primer modelo, por muy poco, pero se entiende porque tiene más variables que el segundo. En cambio según el AIC, el mejor modelo es el segundo, sin la variable Length.

CALCULA LA IMPORTANCIA DE LAS VARIABLES DE ENTRADA

Para seguir entendiendo las variables de entrada lo mejor es calcular el rango de importancia de las variables x.

Copia la importancia de todas ellas utilizando un modelo lineal “lm” y mira si coincide con el modelo que te ha quedado en el pairwise.

Copia la tabla de importancia de las variables …



La variable Length , tal y como lo indica el modelo pairwise, es la que tiene menor importancia.

CALCULA TU PRIMER MODELO PREDICTIVO DE REGRESIÓN

Parte los datos 70% en validation data.

Y 30% test data.

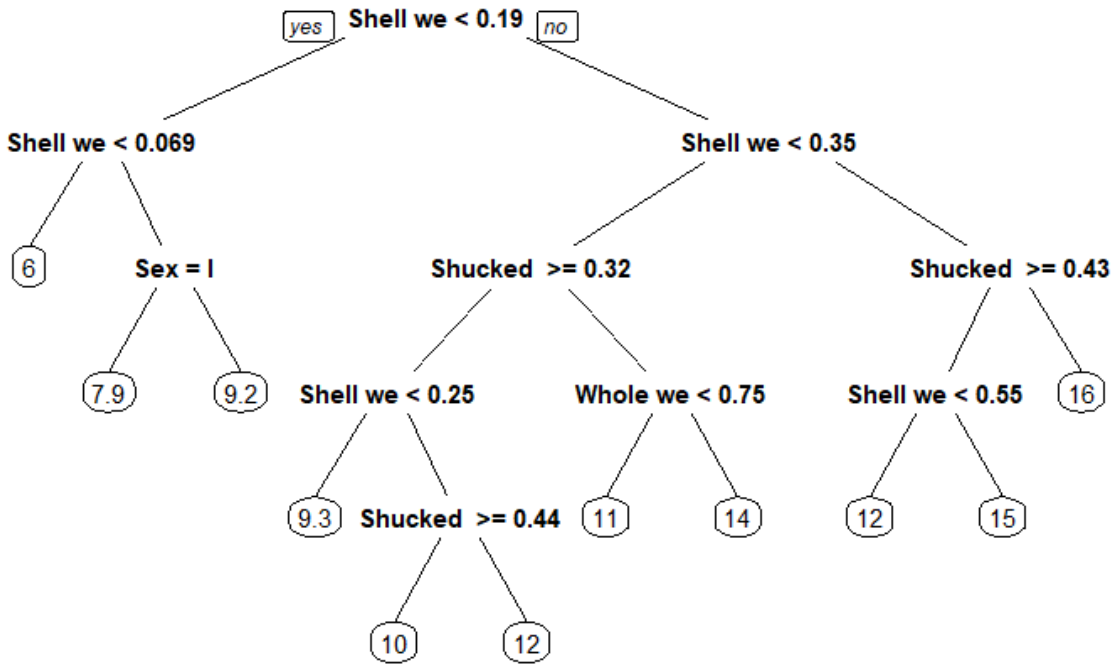
Con los datos de validation calcula un árbol de regresión (CART) y el modelo lineal que has creado antes (lm).

Compara estos modelos con el test data como ya sabes utilizando el RMSE.

¡A por ello!

¿Cuál de ellos está trabajando mejor?

CART:



RMSE = 15.78695

LM:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.06921	0.33486	12.152	< 2e-16	***
SexI	-0.85535	0.12350	-6.926	5.32e-12	***
SexM	0.02003	0.09946	0.201	0.84	
Diameter	10.36464	1.20027	8.635	< 2e-16	***
Height	9.30647	1.70510	5.458	5.22e-08	***
`Whole weight`	10.48693	0.88917	11.794	< 2e-16	***
`Shucked weight`	-21.28983	1.00856	-21.109	< 2e-16	***
`Viscera weight`	-13.13629	1.54776	-8.487	< 2e-16	***
`Shell weight`	8.13569	1.35822	5.990	2.36e-09	***

RMSE = 4.831927

Según el RMSE de ambos modelos, el que tiene menos error es el modelo lineal.

Interpretación del árbol:
La primera bifurcación viene a darse entre los que el peso del caparazón es menor a 0.19, que van hacia la izquierda, de este grupo los que son menores a 0.069 van a la izquierda con una probabilidad de 6% (imagino que es 6% pero la suma de todos los valores debajo de las ramas da un poco más de 100), de los que pesan entre 0.069 y menores a 0.19 van hacia la derecha, allí se dividen entre los que son de sexo indefinido con un 7.9% y masculinos el 9.2%.
Luego, los que el caparazón pesa más que 0.19, se dividen hacia la derecha del primer nodo y se subdividen en: los que el caparazón pesa menos que 0.35 van hacia la izquierda, de allí, los que el peso “shucked” (sin el caparazón) pesan menos que 0.35 pero pesan mayor o igual a 0.32 vuelven a la izquierda, de allí los que el peso del caparazón es menor a 0.25 son el 9.3%, el resto va a la derecha y se dividen en los que pesan sin caparazón más o igual a 0.44 son el 10% y el resto el 12%.
Entre los que sin caparazón pesan menos que 0.32, van a la derecha en ese nodo y se dividen en los que su peso total es menor a 0.75 (11%) y mayor o igual a 0.75 que van a la derecha con un 14%.
Finalmente está el grupo que el caparazón pesa menos que 0.35, que van a la derecha en ese nodo y se dividen en lo que su peso sin caparazón es mayor o igual a 0.43, de los que el caparazón les pesa menos de 0.55 son el 12%, los que les pesa más o igual a 0.55 son el 15% y finalmente los que su peso sin caparazón es de menos de 0.43 que representan un 16%.

¡Tu primer modelo de machine learning de regresión!

¡Muy muy bien!

A por más ☺