TAREA - PLS PARA REGRESIÓN

PRESENTA:

ING. ALEJANDRO NOEL HERNÁNDEZ GUTÍERREZ

ITESO

INSTITUTO TECNOLÓGICO DE ESTUDIOS SUPERIORES DE OCCIDENTE

MAESTRÍA EN CIENCIA DE DATOS

MODELADO PREDICTIVO

IMPARTE: DR. RIEMANN RUÍZ CRUZ

AGOSTO 2023

Contenido

Introducción

El conjunto de datos que se usará en esta actividad recibe el nombre de “Abalone” y puede ser encontrado en el repositorio “UC Irvine Machine Learning Repository” (Abalone - UCI Machine Learning Repository).

El conjunto de datos contiene 4177 observaciones de 8 atributos para predecir la edad de la abulón. La edad de la abulón se determina cortando la concha a través del cono, teñiéndola y contando el número de anillos a través de un microscopio, una tarea aburrida y que requiere mucho tiempo. Existen otras medidas que son más fáciles de obtener para predecir la edad; aunque es posible que se requiera más información, como patrones climáticos y ubicación (por lo tanto, disponibilidad de alimentos), para resolver el problema.

Descripción del conjunto de datos

Nombre	Importancia variable	Tipo de variable	Descripción	Unidades
Sex	Dependiente	Categórica	M, F e I (Infantil)	
Length	Dependiente	Continua	Medida más larga de la concha	mm
Diameter	Dependiente	Continua	Medida perpendicular a la longitud	mm
Height	Dependiente	Continua	Altura con carne en la concha	mm
Whole_weight	Dependiente	Continua	Peso completo	grams
Meat_weight	Dependiente	Continua	Peso de la carne	grams
Víscera_weight	Dependiente	Continua	Peso de las vísceras después de desangrar	grams
Shell_wight	Dependiente	Continua	Peso de la concha	grams
rings	Objetivo	Integer	Anillos, +1.5 gives the age in years	

Desarrollo

En esta sección se abordará el estudio del conjunto de datos desarrollando 7 puntos clave enumerados.

1. Determine si faltan datos y decida si es apropiado utilizar alguna estrategia para completar los datos faltantes.

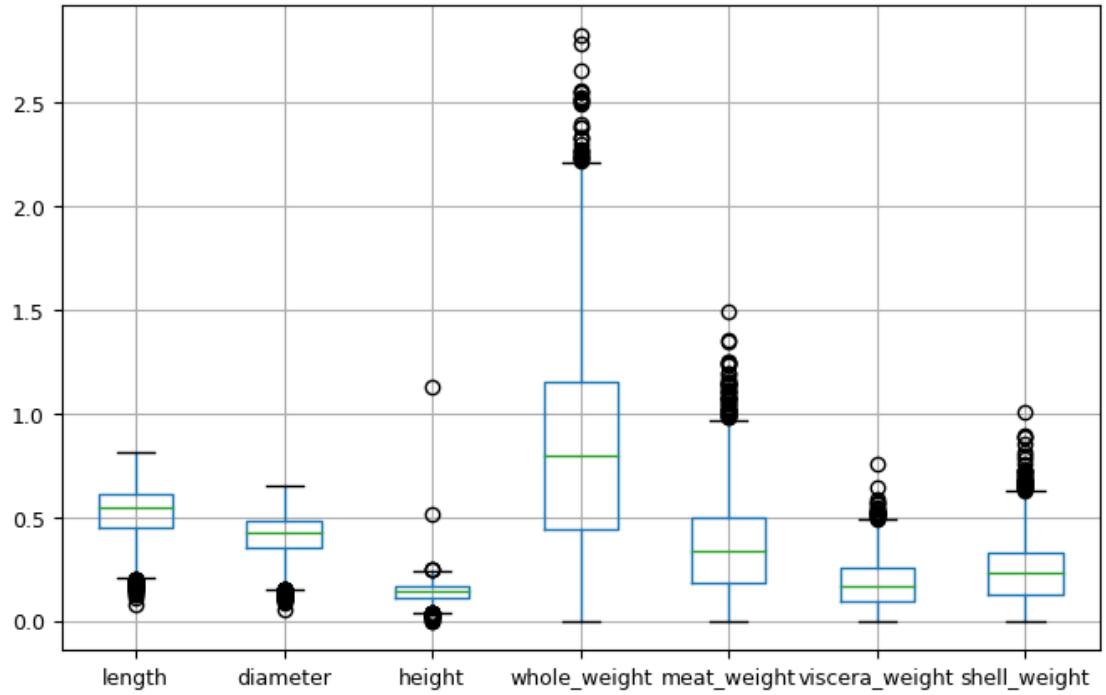
Se cuenta con 4177 observaciones, 8 características de entrada y una característica objetivo. En ningún caso se presentan valores nulos.

```
RangeIndex: 4177 entries, 0 to 4176
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   sex             4177 non-null   object
 1   length          4177 non-null   float64
 2   diameter        4177 non-null   float64
 3   height          4177 non-null   float64
 4   whole_weight    4177 non-null   float64
 5   meat_weight     4177 non-null   float64
 6   viscera_weight  4177 non-null   float64
 7   shell_weight    4177 non-null   float64
 8   rings           4177 non-null   int64
```

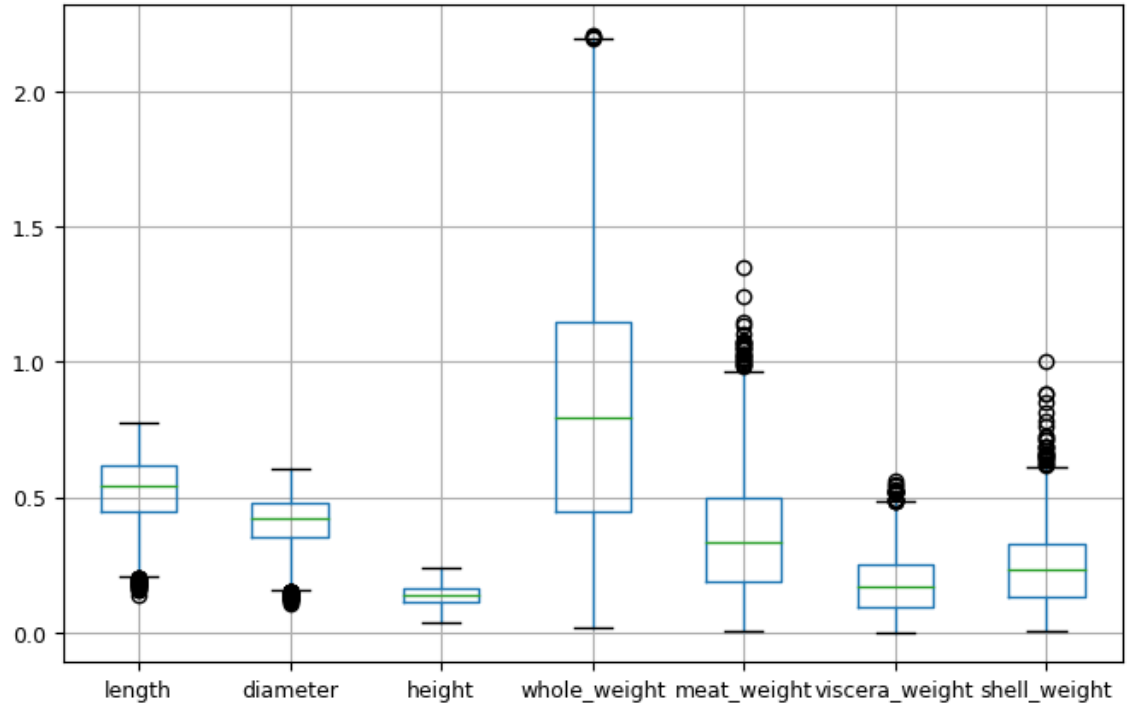
Sin embargo, para características de peso y altura no es válido el valor 0, en el análisis de datos se encuentran dos observaciones en la característica de altura.

```
sex               0
length            0
diameter          0
height            2
whole_weight      0
meat_weight       0
viscera_weight    0
shell_weight      0
rings             0
```

Antes de proceder se realiza un análisis de valores atípicos con el fin de eliminar observaciones que resulten desfavorables para el estudio.



Se decide eliminar las observaciones con height y whole_weight para observar el comportamiento al considerar que podrían eliminar consecuentemente algunos valores atípicos de otras características.



Como es posible observar, se reduce bastante los valores atípicos. Al tratarse de mediciones tan pequeñas, se decide no continuar con la eliminación de atípicos y revisar otra vez cuantas observaciones quedan.

	length	diameter	height	whole_weight	meat_weight	viscera_weight	shell_weight	rings
count	4120.000000	4120.000000	4120.000000	4120.000000	4120.000000	4120.000000	4120.000000	4120.000000
mean	0.524242	0.408028	0.139297	0.821193	0.355680	0.179205	0.237133	9.946602
std	0.116367	0.096246	0.037154	0.470582	0.212181	0.106062	0.134931	3.197028
min	0.135000	0.105000	0.040000	0.015500	0.005000	0.000500	0.005000	3.000000
25%	0.450000	0.350000	0.115000	0.444375	0.187000	0.093875	0.130000	8.000000
50%	0.545000	0.425000	0.140000	0.797500	0.335500	0.170000	0.232500	9.000000
75%	0.615000	0.480000	0.165000	1.145000	0.498500	0.250500	0.325000	11.000000
max	0.775000	0.605000	0.240000	2.210000	1.351000	0.564000	1.005000	29.000000

En la ilustración anterior se muestra que ya no hay observaciones con valor mínimo 0 (en los casos en que no es esperado). Además, la distribución de los datos no genera ninguna alerta de comportamiento extraño y la cantidad de datos es buena con respecto a los originales.

2. Cree dos subconjuntos de datos, donde el primero se utilizará para el proceso de entrenamiento y el segundo para el proceso de prueba. (“dividir conjuntos de datos de entrenamiento y prueba”).

Antes que nada, se aplica la codificación one-hot en la variable categórica para hacer posible el entrenamiento del modelo lineal.

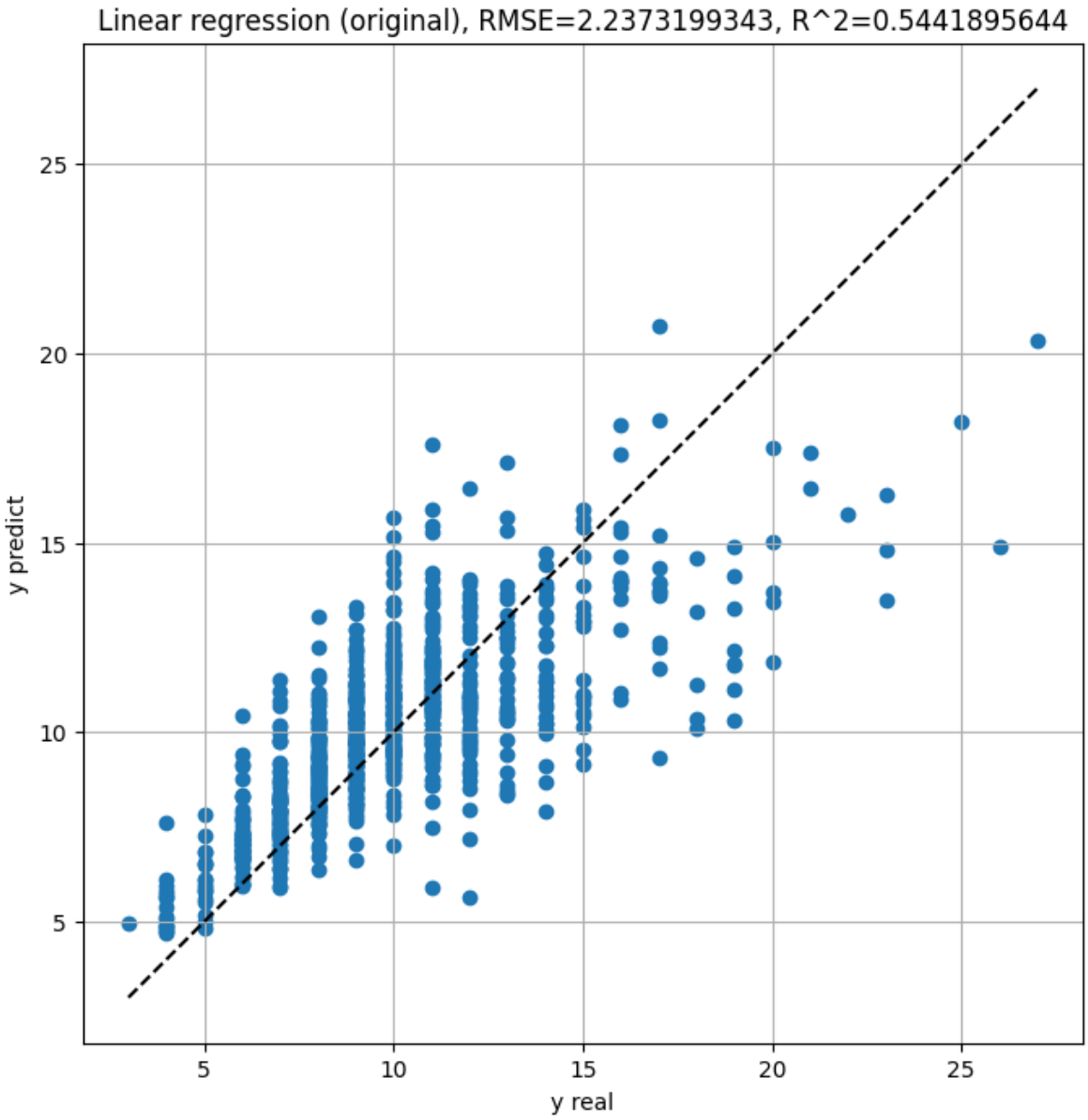
	length	diameter	height	whole_weight	meat_weight	viscera_weight	shell_weight	rings	sex_F	sex_I	sex_M
0	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15	0	0	1
1	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7	0	0	1
2	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9	1	0	0
3	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10	0	0	1
4	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
4172	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490	11	1	0	0
4173	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10	0	0	1
4174	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9	0	0	1
4175	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10	1	0	0
4176	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12	0	0	1

Con este código se divide el conjunto de datos en entrenamiento y prueba.

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import (mean_squared_error,r2_score)
x_train, x_test, y_train, y_test = train_test_split(data[x_column_names],
data.rings, _size=0.2, random_state=42)
```

3. Entrene un modelo lineal para estimar la característica *"Rings"* utilizando como entradas para el modelo todas las demás variables. Obtener los valores de RMSE y R2 para evaluar el rendimiento del modelo tanto en entrenamiento como en pruebas.

Estos son los resultados:



4. Considerando el mismo conjunto de datos utilizados en el punto 3; realizar una selección de características utilizando criterios de varianza o correlación. Con las variables resultantes del proceso de eliminación, entrenar un nuevo modelo lineal y calcular las métricas RMSE y R2 correspondientes al entrenamiento y testing.

Estas son las varianzas de las variables:

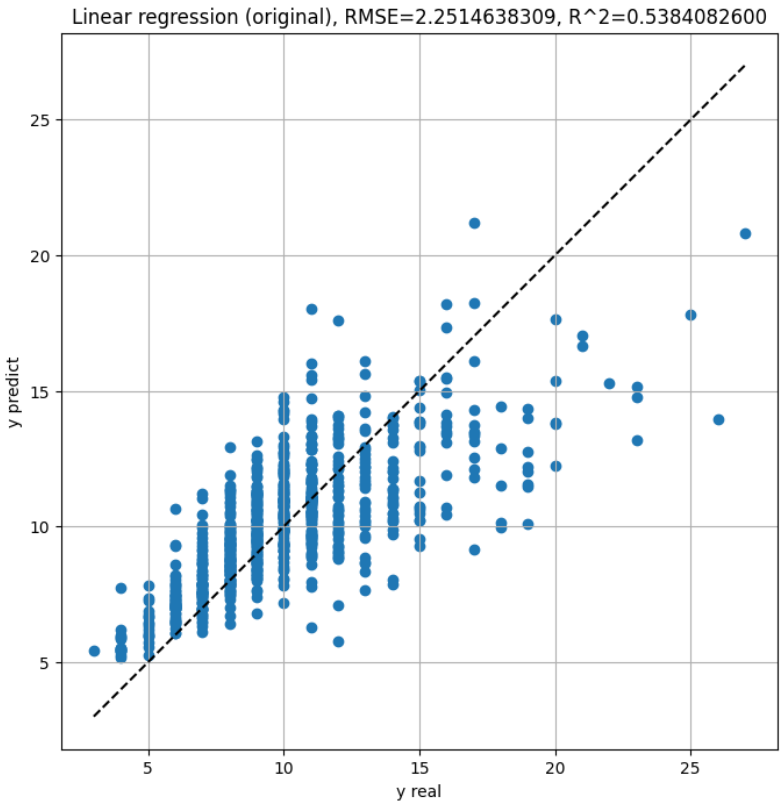
```
sex_F           0.215395
sex_I           0.217793
sex_M           0.232040
length          0.013541
diameter        0.009263
height          0.001380
whole_weight    0.221447
meat_weight     0.045021
viscera_weight  0.011249
shell_weight    0.018206
```

Se define un umbral como criterio para eliminar las varianzas más bajas. En este caso es el 1% de la varianza total.

Con este criterio se eliminan las siguientes variables:

`'diameter', 'height'`

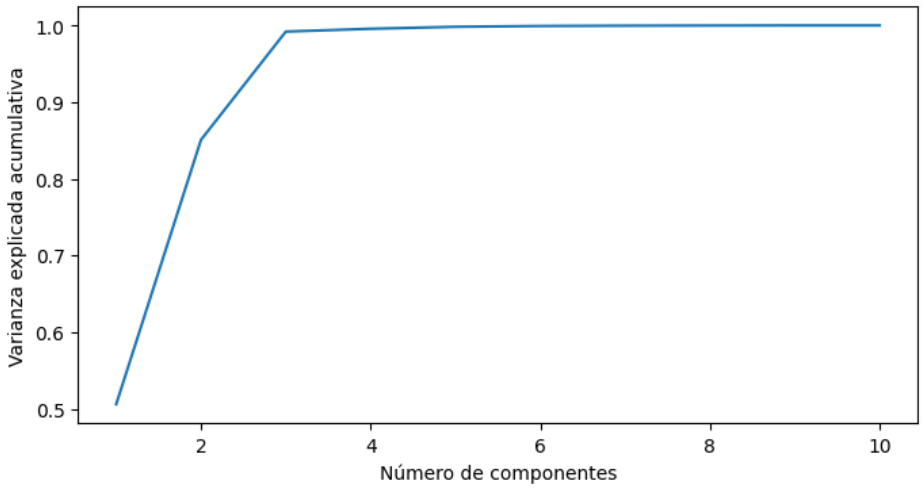
Los resultados obtenidos son:



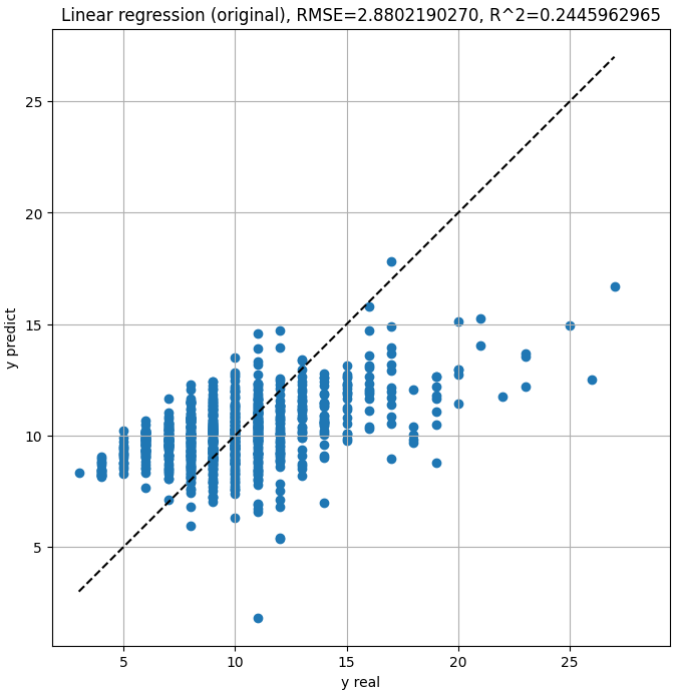
5. Nuevamente, considere el mismo conjunto de datos utilizado en el punto 3, realice una reducción de variables mediante análisis de componentes principales (PCA). Con las variables resultantes del proceso de reducción, entrenar un nuevo modelo lineal y calcular las métricas RMSE y R2, correspondientes al entrenamiento y testing.

Luego del PCA se obtienen 9 componentes principales.

Al revisar la varianza acumulada, se deciden dejar 3 componentes porque contienen más del 96% de la varianza acumulada.

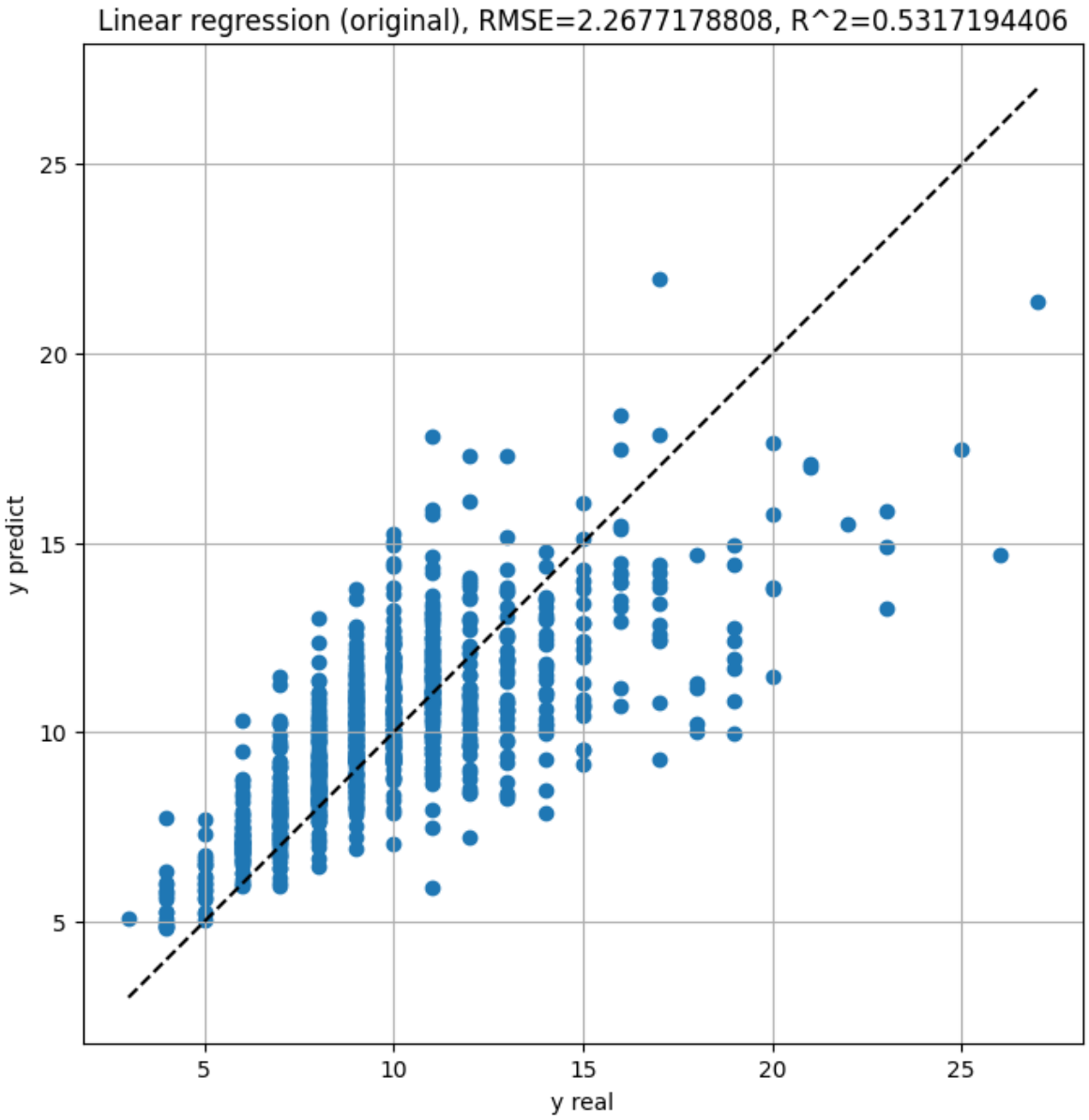


Y este es el resultado:



6. Considerando nuevamente el mismo conjunto de datos utilizado en el punto 3, entrene un nuevo modelo lineal usando la técnica PLS y calcule las métricas RMSE y R2 correspondientes al entrenamiento y prueba.

Usando 5 componentes en el modelo se obtiene



7. Debido a los pasos anteriores, tienes cuatro modelos lineales diferentes para resolver el problema propuesto. Realiza una tabla con las métricas de cada modelo para hacer una comparativa de los modelos.

Modelo	RMSE	R^2
Modelo simple LM	2.237319	0.544189
Eliminación por varianza y luego LM	2.251463	0.538408
PCA y luego LM 3 componentes	2.880219	0.244596
PLS y luego LM 5 componentes	2.267717	0.531719

Conclusiones

En el caso de este conjunto de datos una regresión lineal simple nos entregó mejores resultados con un error minimizado y una r2 maximizada con respecto a los demás modelos. Dependiendo la aplicación, se podría optar por alguna de las alternativas. Algo que se puede precisar es que, si el fin último de estos modelos es predecir la edad del Abalone y dependiendo de la precisión que se requiera, podría recomendarse hacer un modelo lineal simple eliminando las variables cuya varianza es menor porque se tiene la ventaja de no requerir que los científicos hagan esa medición.