

REPORTEREGRESSION

Paula Camila González Ortega

DATA
MINING

INFORMACIÓN DEL DATASET

En este proyecto se trabajará con el dataset insurance.csv, un conjunto de datos que indican ciertas características físicas y rutinarias de las personas. Dichos datos personales permitirán predecir los cargos de los usuarios. Entre estos datos se puede mencionar la edad, el sexo, el índice de masa corporal, región, cantidad de hijos, si fuma cigarrillos o no y los 'charges' de la persona.

HIPÓTESIS U OBJETIVO

En este laboratorio, se tiene como objetivo crear varios modelos de regresión lineal y polinomial. En base a dichos modelos, desarrollados con las características de distintos usuarios, será posible predecir los 'charges' de cada persona.

SOLUCIÓN Y EXPLORACIÓN

Se encontraron 348 registros de usuarios, con 7 columnas como se ve en la *Imagen 1*. Ningún registro cuenta con valores nulos en las respectivas columnas.

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520

Imagen 1

Tanto la columna de sexo como la de si fuma o no, presentan un valor binario. Mientras que los valores en el resto de columnas son números enteros o decimales.

Se obtuvieron datos estadísticos como la media , máximos y mínimos, entre otros (*Imagen 2*) para poder tener una visión más general del dataset.

	age	sex	bmi	children	smoker	region	charges
count	348.000000	348.000000	348.000000	348.000000	348.000000	348.000000	348.000000
mean	39.591954	0.508621	30.676552	1.091954	0.232759	1.497126	14016.426293
std	14.417015	0.500646	5.625850	1.192021	0.423198	1.104089	12638.887852
min	18.000000	0.000000	15.960000	0.000000	0.000000	0.000000	1137.011000
25%	27.000000	0.000000	26.782500	0.000000	0.000000	1.000000	4888.466125
50%	40.000000	1.000000	30.300000	1.000000	0.000000	2.000000	9719.305250
75%	53.000000	1.000000	34.777500	2.000000	0.000000	2.000000	19006.316150
max	64.000000	1.000000	49.060000	5.000000	1.000000	3.000000	51194.559140

Imagen 2

En la *Imagen 3* se puede ver la relación entre todas las variables del dataset,. Entre más intenso el color mayor relación entre las variables hay y como era de esperarse toda la diagonal es azul intenso. Sin embargo, se encontró relación entre la variable *smoker* y *charges*.

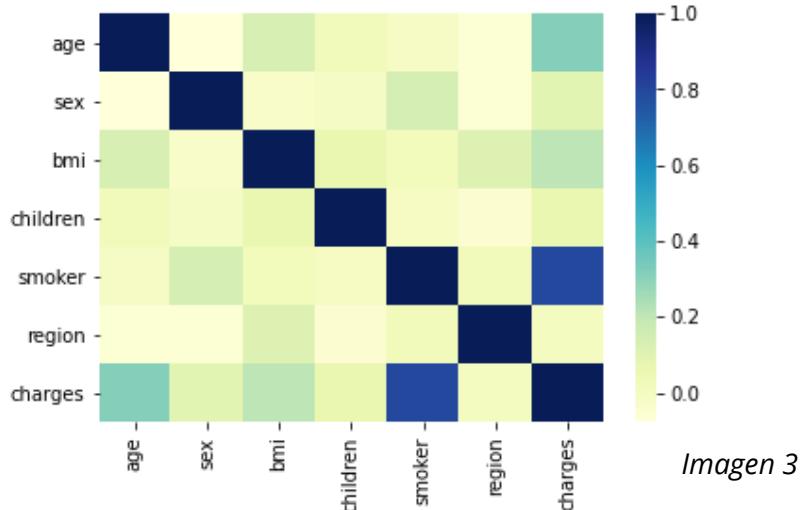


Imagen 3

Con la gráfica de la *Imagen 4* se descubrió que la edad no influye en los cargos, ya que se tienen registrados cargo de 10000 hasta 50000 para todas las edad. También que entre mayor sea sean los cargos, mayor dispersión entre ellos hay.

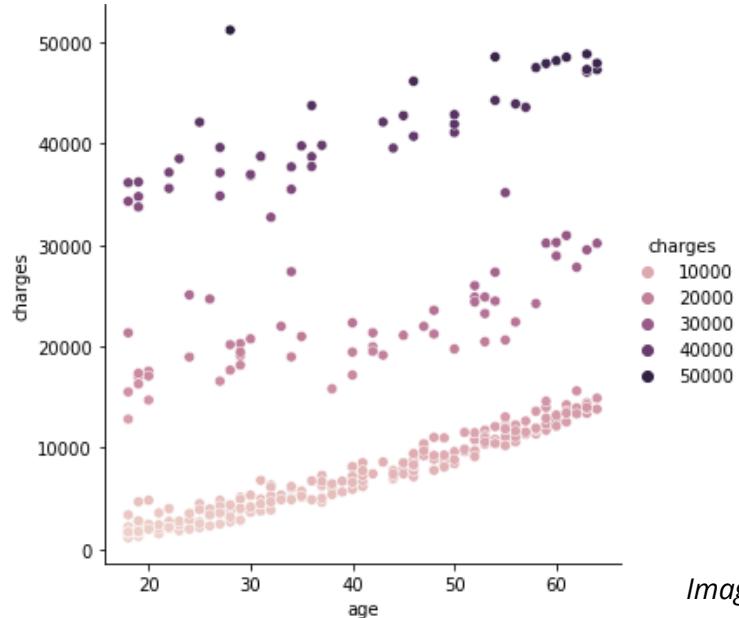
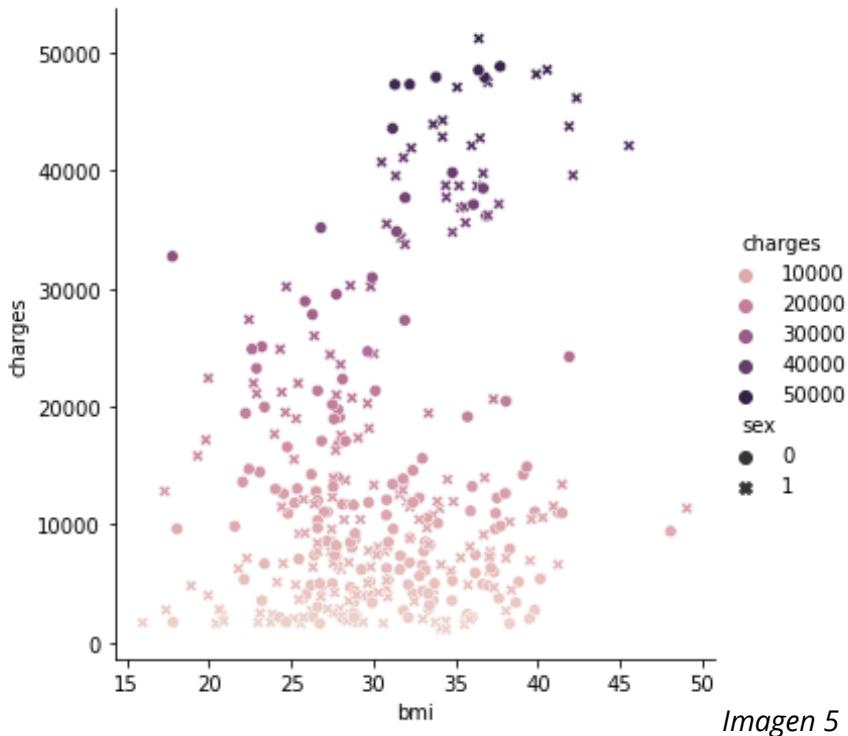


Imagen 4

Fue necesario conocer la relación entre BMI y cargos diferenciando el sexo de los usuarios. Con la gráfica de la imagen 5 se descubrió que tanto hombres como mujeres, tienden a tener un cargo mayor conforme su BMI se encuentra entre 35-40



La cantidad de registros es bastante equitativa entre hombres y mujeres como se puede observar en la imagen 6. Sin embargo hay ligeramente más hombres registrados en el dataset.

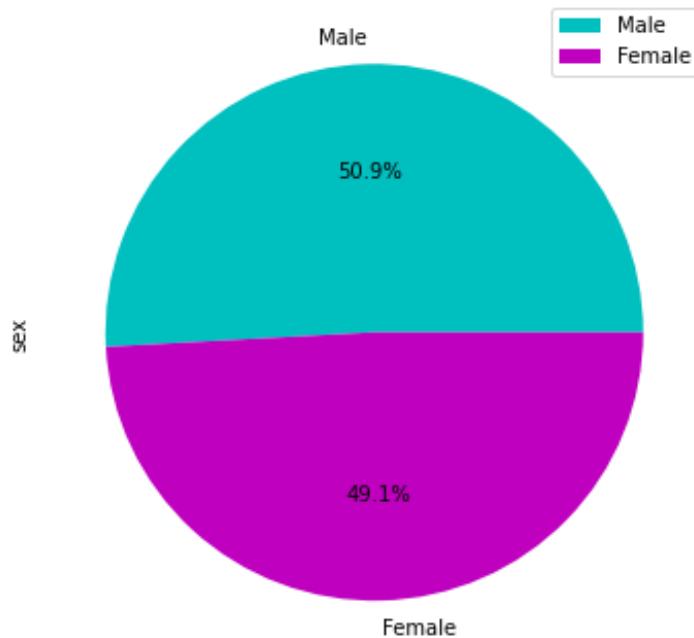


Imagen 6

Gracias a la gráfica de la imagen 7 se descubrió que más de 140 registros no tienen hijos y que solamente un porcentaje mínimo de los registros tienen 5 hijos

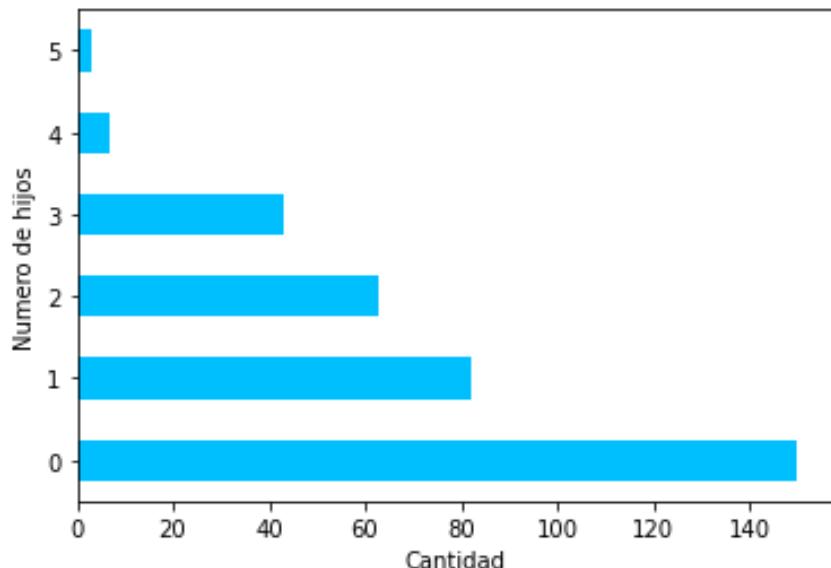


Imagen 7

El último hallazgo fue que el 76.7% de los usuarios registrados no son fumadores y el 23.3% si. Esto se respalda con la imagen 8 de este reporte

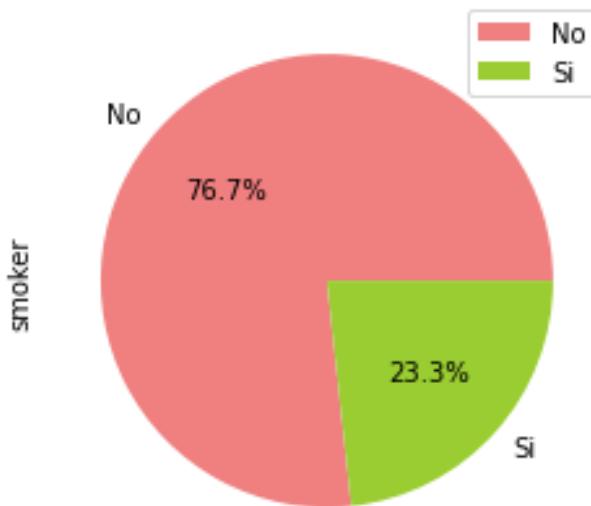


Imagen 8

Luego de explorar la data, se prosiguió establecer data de tipo categórico (sex , smoker y región). También se dividió la data de training y test para poder hacer las predicciones de cargos, las cuales serán desglosadas en el apartado de resultados.

RESULTADOS

Con la data de training y testing sí se lograron desarrollar los modelos lineales con y sin librerías, así como modelos polinomiales, En algunos se usó el BMI como feature (modelo 1 y 2) mientras que en el modelo 3 se utilizaron todas las variables como features.

A continuación desglozan los resultados de cada modelo de regresión lineal:

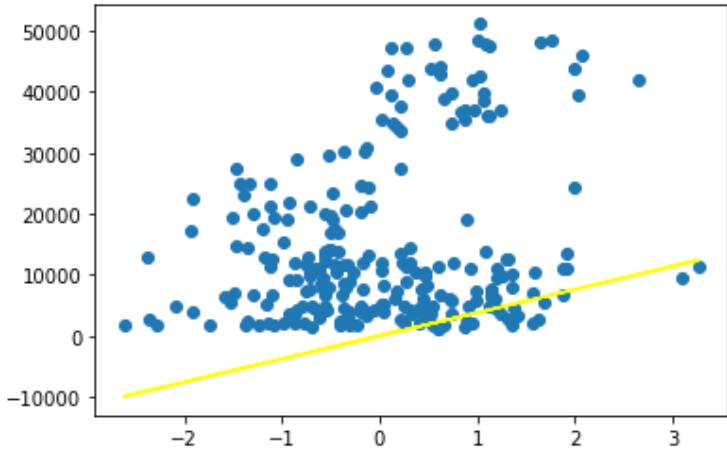


Imagen 9

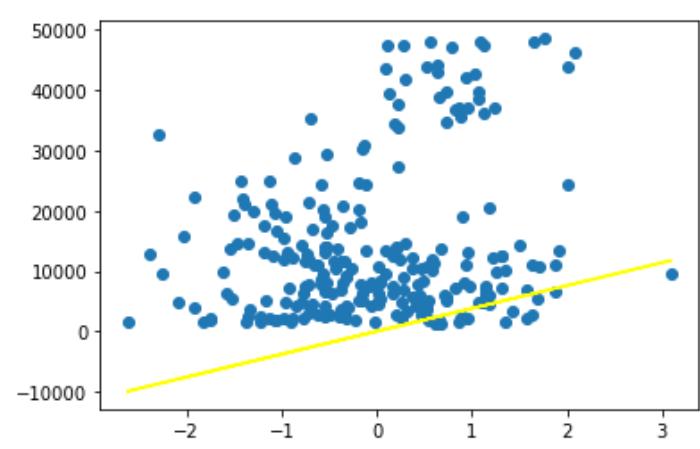


Imagen 10

En la imagen 9 se observa la regresión lineal de los datos en el modelo 1, realizado sin el uso de librerías. Mientras que en la imagen 10 se ven la regresión lineal del modelo 2, elaborado con librerías. Al ser similares se puede concluir que ambos modelos son exitosos sin importar si se utiliza o no alguna librería que facilite el proceso.

También se obtuvieron p-values de las regresiones del modelo 1 y 2 (Imagen 11) y modelo 3 (Imagen 12)

OLS Regression Results						
Dep. Variable:	charges	R-squared:			1.000	
Model:	OLS	Adj. R-squared:			1.000	
Method:	Least Squares	F-statistic:			3.216e+31	
Date:	Sat, 17 Apr 2021	Prob (F-statistic):			0.00	
Time:	23:52:57	Log-Likelihood:			8212.5	
No. Observations:	348	AIC:			-1.641e+04	
Df Residuals:	338	BIC:			-1.637e+04	
Df Model:	9					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-3.183e-12	3.88e-12	-0.820	0.413	-1.08e-11	4.46e-12
age	-2.593e-13	6.17e-14	-4.202	0.000	-3.81e-13	-1.38e-13
sex	-4.547e-13	7.54e-13	-0.603	0.547	-1.94e-12	1.03e-12
bmi	4.192e-13	1.43e-13	2.936	0.004	1.38e-13	7e-13
children	-2.842e-13	6.29e-13	-0.452	0.652	-1.52e-12	9.54e-13
smoker	7.731e-12	1.72e-12	4.486	0.000	4.34e-12	1.11e-11
charges	1.0000	1.23e-16	8.11e+15	0.000	1.000	1.000
labeled_sex	-5.684e-13	7.54e-13	-0.753	0.452	-2.05e-12	9.16e-13
labeled_smoker	6.594e-12	1.72e-12	3.826	0.000	3.2e-12	9.98e-12
region_0	-1.137e-12	1.6e-12	-0.711	0.478	-4.28e-12	2.01e-12
region_1	1.961e-12	1.54e-12	1.270	0.205	-1.08e-12	5e-12
region_2	-3.411e-12	1.7e-12	-2.006	0.046	-6.76e-12	-6.58e-14
region_3	2.274e-13	1.64e-12	0.138	0.890	-3e-12	3.46e-12
Omnibus:	67.483	Durbin-Watson:			1.435	
Prob(Omnibus):	0.000	Jarque-Bera (JB):			102.646	
Skew:	-1.299	Prob(JB):			5.14e-23	
Kurtosis:	3.576	Cond. No.			2.37e+20	

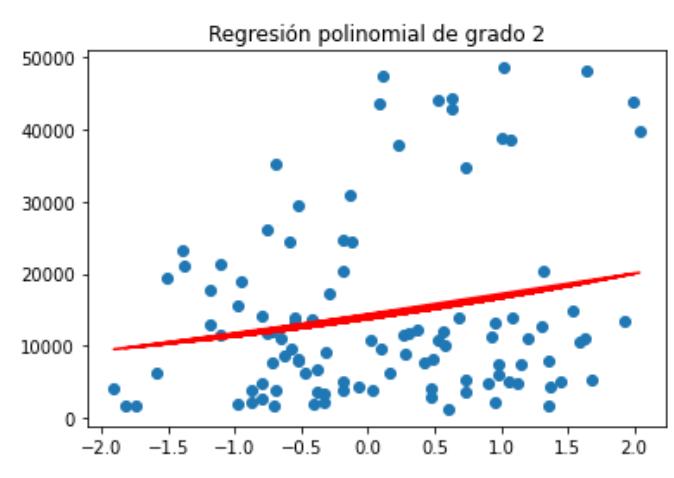
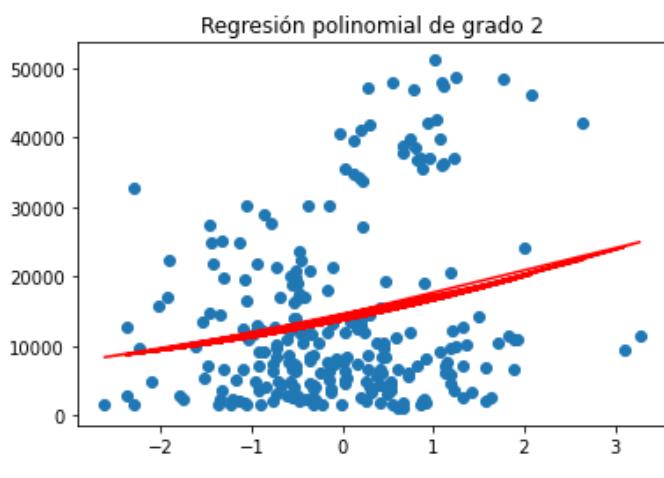
Imagen 11

OLS Regression Results						
Dep. Variable:	charges	R-squared:	1.000			
Model:	OLS	Adj. R-squared:	1.000			
Method:	Least Squares	F-statistic:	3.216e+31			
Date:	Sun, 18 Apr 2021	Prob (F-statistic):	0.00			
Time:	00:03:07	Log-Likelihood:	8212.5			
No. Observations:	348	AIC:	-1.641e+04			
Df Residuals:	338	BIC:	-1.637e+04			
Df Model:	9					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	-3.183e-12	3.88e-12	-0.820	0.413	-1.08e-11	4.46e-12
age	-2.593e-13	6.17e-14	-4.202	0.000	-3.81e-13	-1.38e-13
sex	-4.547e-13	7.54e-13	-0.603	0.547	-1.94e-12	1.03e-12
bmi	4.192e-13	1.43e-13	2.936	0.004	1.38e-13	7e-13
children	-2.842e-13	6.29e-13	-0.452	0.652	-1.52e-12	9.54e-13
smoker	7.731e-12	1.72e-12	4.486	0.000	4.34e-12	1.11e-11
charges	1.0000	1.23e-16	8.11e+15	0.000	1.000	1.000
labeled_sex	-5.684e-13	7.54e-13	-0.753	0.452	-2.05e-12	9.16e-13
labeled_smoker	6.594e-12	1.72e-12	3.826	0.000	3.2e-12	9.98e-12
region_0	-1.137e-12	1.6e-12	-0.711	0.478	-4.28e-12	2.01e-12
region_1	1.961e-12	1.54e-12	1.270	0.205	-1.08e-12	5e-12
region_2	-3.411e-12	1.7e-12	-2.006	0.046	-6.76e-12	-6.58e-14
region_3	2.274e-13	1.64e-12	0.138	0.890	-3e-12	3.46e-12
Omnibus:	67.483	Durbin-Watson:	1.435			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	102.646			
Skew:	-1.299	Prob(JB):	5.14e-23			
Kurtosis:	3.576	Cond. No.	2.37e+20			

Imagen 12

En la imagen 13 se observa la regresión polinomial de los datos de training del modelo 2, que utiliza BMI como feature. Mientras que en la imagen 14 se ven la regresión polinomial con sus datos de testing.

Dichas regresiones son similares por lo que son correctos. Aunque se recomienda prestar atención en próximas iteraciones de testing porque se observó mayor dispersión entre los datos de testing.



```
1 #train data
2 print(mean_squared_error(polyreg3.predict(x_trainCp),y_trainCp))
3 print(r2_score(polyreg3.predict(x_trainCp),y_trainCp))
4 print(mean_absolute_error(polyreg3.predict(x_trainCp), y_trainCp))
```

```
20410446.439592738
0.874196326212864
2579.5735338683126
```

```
1 #test data
2 print(mean_squared_error(polyreg3.predict(x_testCp),y_testCp))
3 print(r2_score(polyreg3.predict(x_testCp),y_testCp))
4 print(mean_absolute_error(polyreg3.predict(x_testCp), y_testCp))
```

```
23072669.07394339
0.7875220543392588
2657.909713619047
```

Imagen 15

Finalmente con la imagen 15 se concluye que para el modelo 3, que utiliza todas las variables como feature, las regresiones polinomiales son similares entre la data de training y de testing. Por lo que el resultado es fue exitoso.