Lab 4

Información del dataset

Los datos contienen información de las personas y, en base a esto, cuánto cobra la compañía de seguros para asegurarlos. El data set tiene las siguientes columnas:

- Age: la edad de la persona.
- Sex: el sexo de la persona.
- Bmi: Body mass index.
- Children: La cantidad de hijos.
- Smoker: Si es fumador o no.
- Region: El número de región de la persona.
- Charges: La cantidad que cobra la compañía de seguros para asegurar a la persona.

Información general de las columnas 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

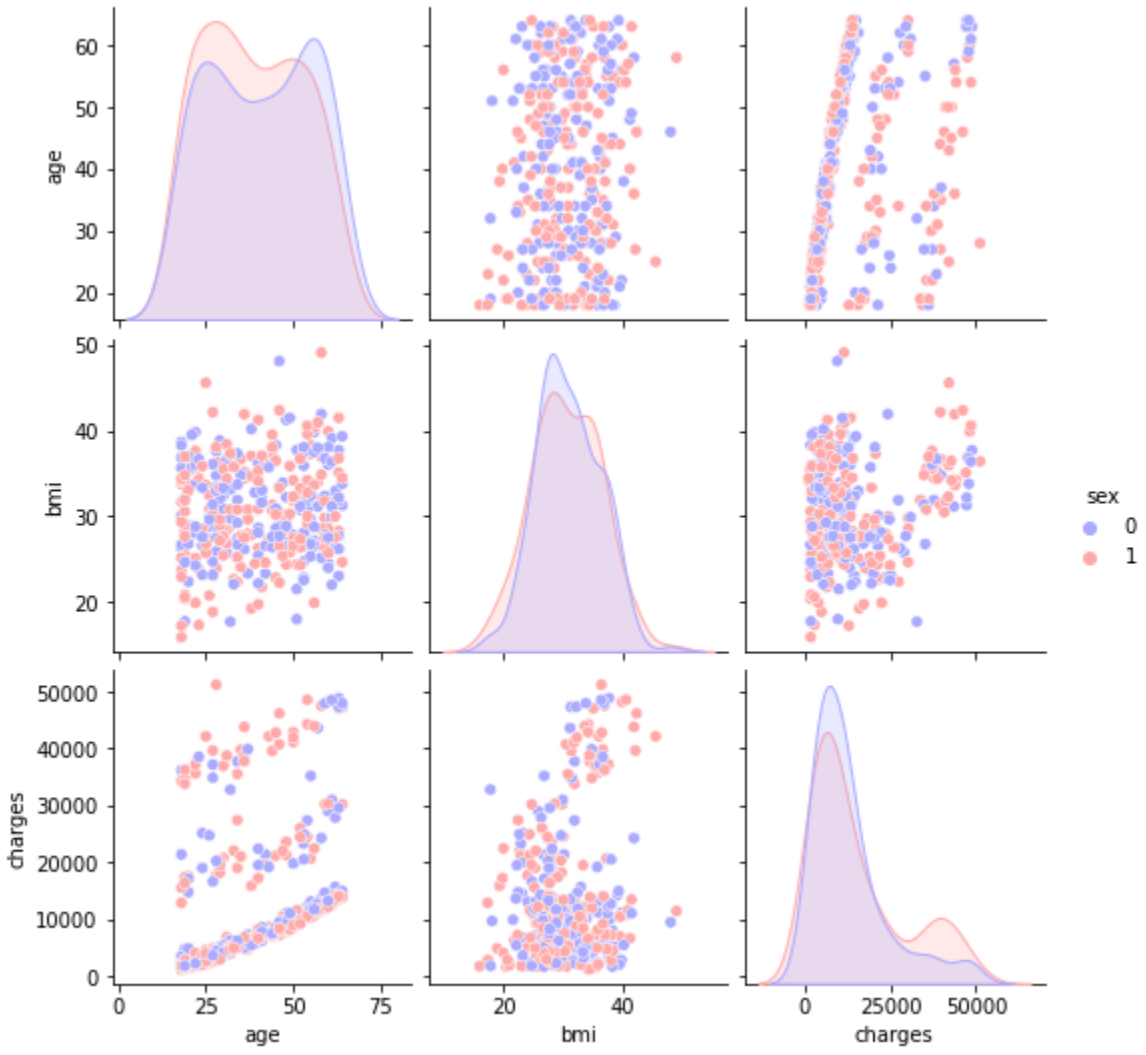
Hipótesis u objetivo

El objetivo de este laboratorio es predecir los cargos del seguro para las nuevas personas en función de la información que obtendremos de ellos. También como objetivo de este análisis es identificar la manera en la que es más efectivo el entrenamiento de la regresión lineal, explorando los datos, preparándolos antes de analizarlos realizando una normalización y estandarización de los features y escogiendo los features adecuados para que el modelo pueda predecir efectivamente ejemplos reales de la vida real.

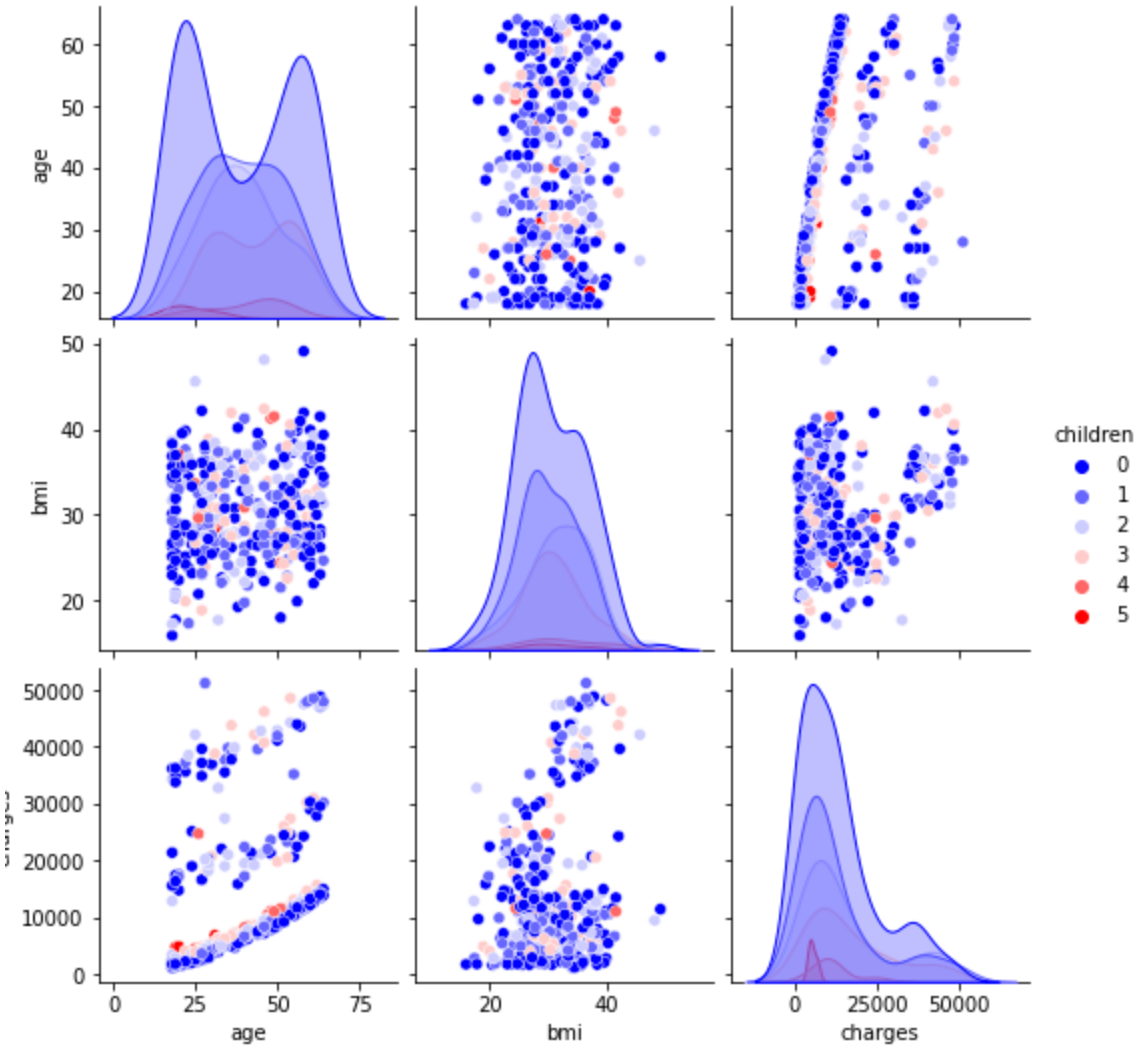
Solución y exploración

Primero se explora la data y se ve el tamaño que tiene y las características que se tiene de cada persona. Luego agrupados los datos para generar el pairewise y así visualizar los campos numéricos en una gráfica donde se pueda apreciar la distribución de estos. Estos se agruparon o dividieron en base a los atributos o campos no numéricos.

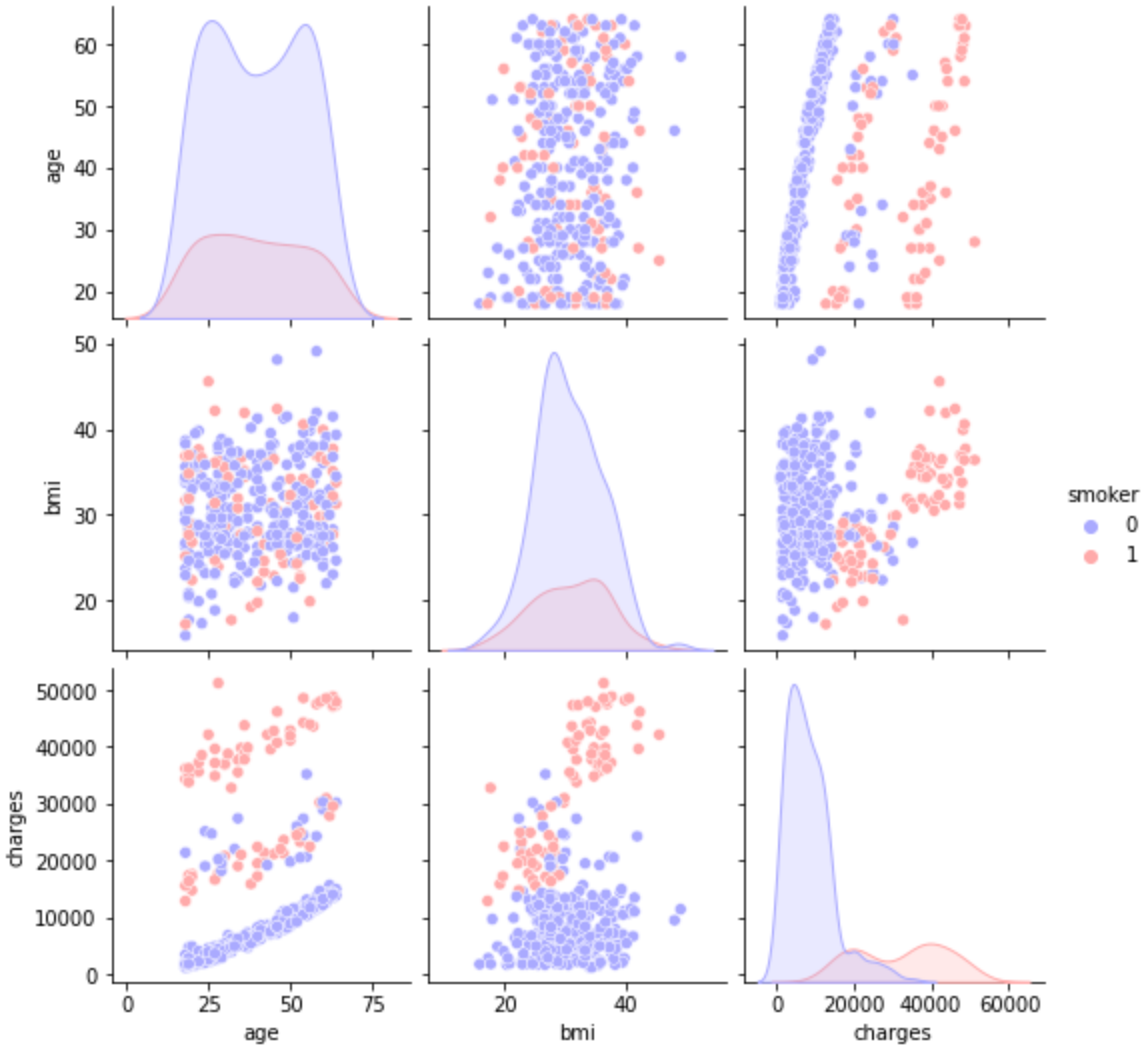
Pairwase agrupado por el sexo:



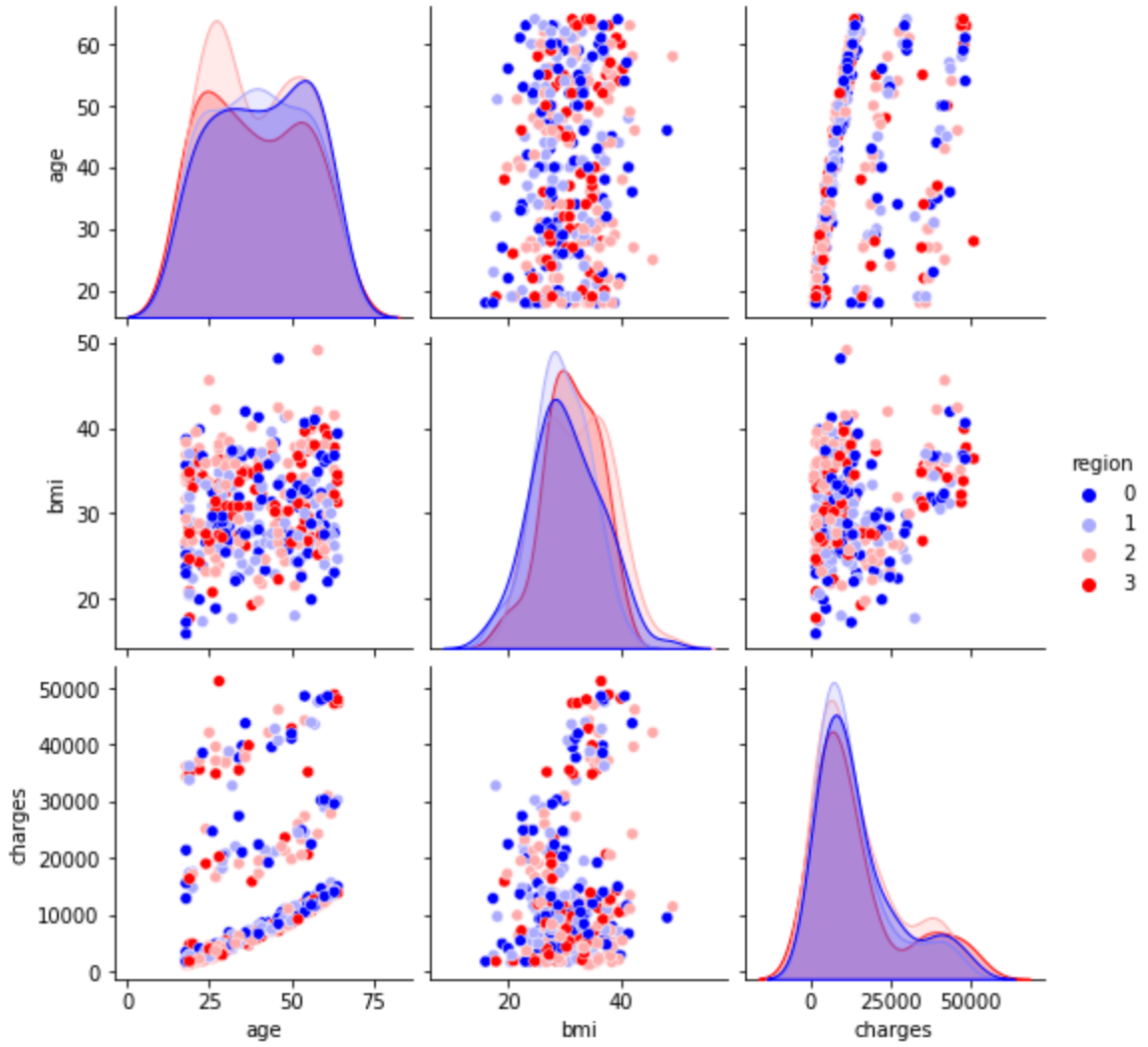
Pairwise agrupado por cantidad de hijos:



Pairwise agrupado si fuma o no:



Pairwise agrupado por región:



Luego se busco la cantidad de nulls en las columnas del dataset para poderlos analizar sin ningún inconveniente. Luego de buscar, no se encontró ningún caso en ninguna columna por lo que no se tuvo que hacer ninguna modificación.

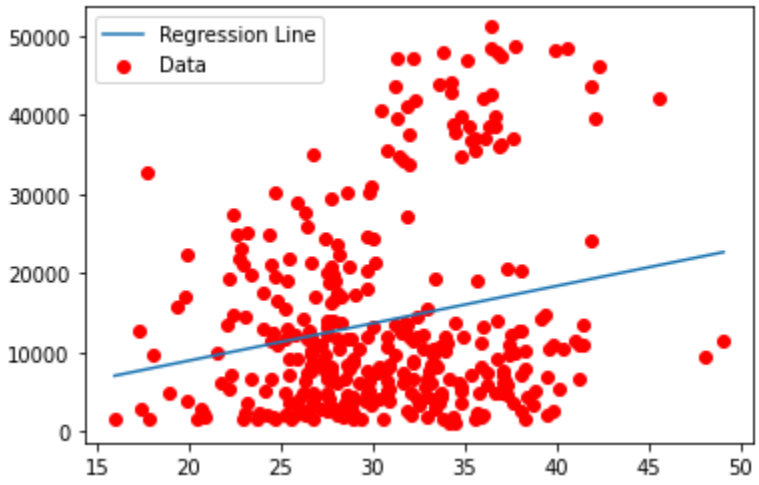
Luego de esto se dividieron los datos en el dataset en x y y dividiendo las variables independietes de las variables dependientes. Para el modelo 1 y 2 solo se tomó en cuenta el feature bmi para realizar el análisis, mientras que para el modelo 3 se tomaron en cuenta todas las features del dataset. Luego de esto se generaron los arreglos de entrenamiento, y los arreglos para probar nuestros modelos.

Antes de crear los modelos se realizo la preparación de datos en donde se aplicó normalización y estandarización al dataset para tener todas las variables independientes en una misma escala y de esta manera conseguir mejores resultados y que el resultado de la regresión no esté sesgado por una diferencia de escalas.

Por último, se crearon los modelos 1, 2 y 3. El primer modelo se creó desarrollando el código de least squares desde 0 sin usar inguna librería, mientras que para el modelo 2 y 3 se utilizó la librería de sklearn.

Resultados: puntuales y respaldados por gráficos, tablas y visualizaciones en general

Análisis del modelo 1:



RMSE	R^2
101.11133200191263	0.04412123949701052

Análisis del modelo 2:

Model:	OLS	Adj. R-squared (uncentered):	0.008
Dependent Variable:	charges	AIC:	5258.2014
Date:	2021-04-19 17:24	BIC:	5261.6524
No. Observations:	233	Log-Likelihood:	-2628.1
Df Model:	1	F-statistic:	2.848
Df Residuals:	232	Prob (F-statistic):	0.0928
R-squared (uncentered):	0.012	Scale:	3.6860e+08

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
x1	2122.7439	1257.7692	1.6877	0.0928	-355.3656	4600.8535

Omnibus:	36.155	Durbin-Watson:	0.909
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49.531
Skew:	1.128	Prob(JB):	0.000
Kurtosis:	3.125	Condition No.:	1

RMSE	MSE	MAE	R^2
124939624.84899	11177.639502551	8881.578890658	0.08370071861348

Análisis del modelo 3:

Model:	OLS	Adj. R-squared (uncentered):	0.008
Dependent Variable:	charges	AIC:	5258.2014
Date:	2021-04-19 17:24	BIC:	5261.6524
No. Observations:	233	Log-Likelihood:	-2628.1
Df Model:	1	F-statistic:	2.848
Df Residuals:	232	Prob (F-statistic):	0.0928
R-squared (uncentered):	0.012	Scale:	3.6860e+08

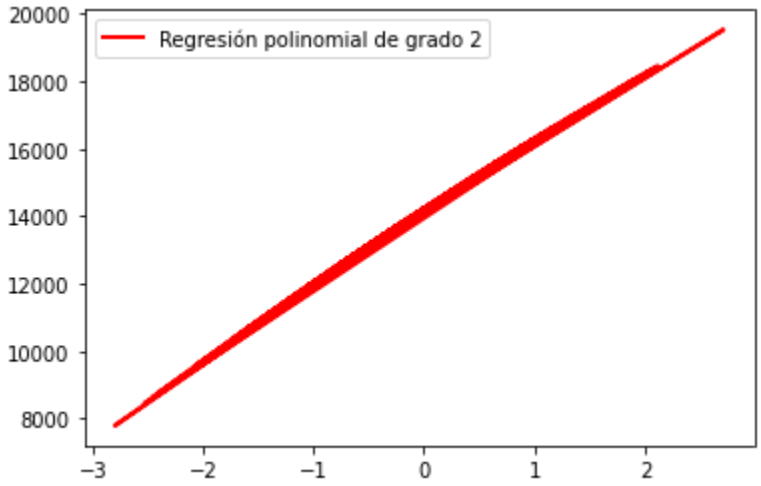
	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
x1	2122.7439	1257.7692	1.6877	0.0928	-355.3656	4600.8535

Omnibus:	36.155	Durbin-Watson:	0.909
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49.531
Skew:	1.128	Prob(JB):	0.000
Kurtosis:	3.125	Condition No.:	1

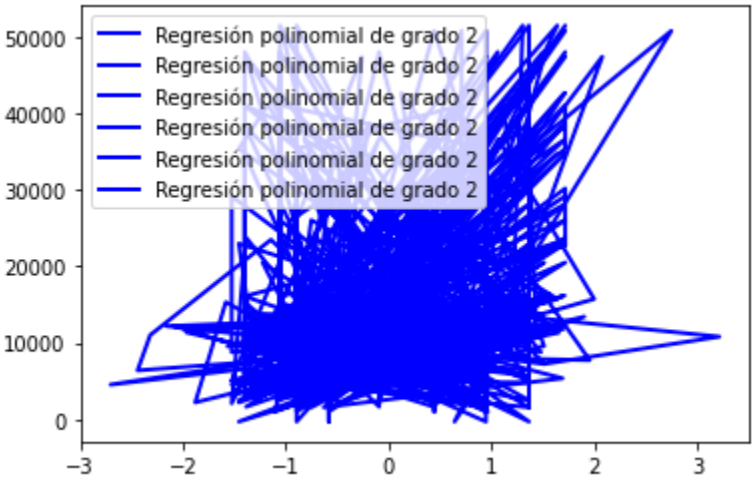
RMSE	MSE	MAE	R^2
39848575.38912	6312.5728026795	4315.2772908440	0.7163723630799

Gráficas aplicando regresión lineal:

Modelo 1:



Modelo 2:



Podemos ver que los resultados del modelo 1 y 2 son muy parecidos y el R^2 es casi el mismo, esto es muy lógico ya que para los dos modelos se utilizó el mismo feature que fue el bmi, por lo que los resultados tendrían que ser mucho más parecidos que con el modelo 3.

El modelo 3 mostró mejores resultados que el modelo 1 y 3. Estos eran los resultados esperados ya que al utilizar más features el algoritmo de least squares pudo identificar de mejor manera el dataset y el cargo del seguro tiene mucho más relación con todas las features juntas y por lo tanto el modelo entrenado y generado es mucho más útil para predecir cuanto se cobrará a una nueva persona dependiendo de sus características personales.