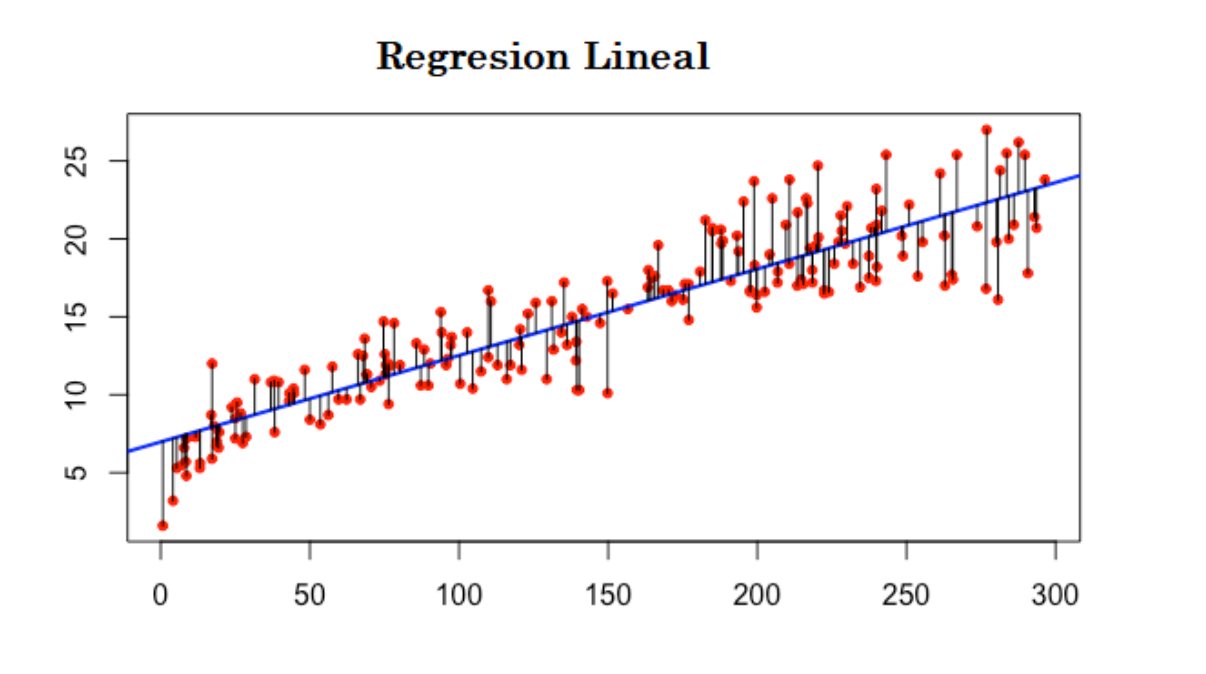
CIENCIA DE DATOS

- Nombre: Lucas Gracés
- Temas de la presentación
  - Temas que vimos en el curso
  - Cosas que vimos
  - Proyecto final

MACHINE LEARNING

- Al principio del curso empezamos viendo un poco lo que ya habíamos visto en el modulo 1 y 2 del curso.



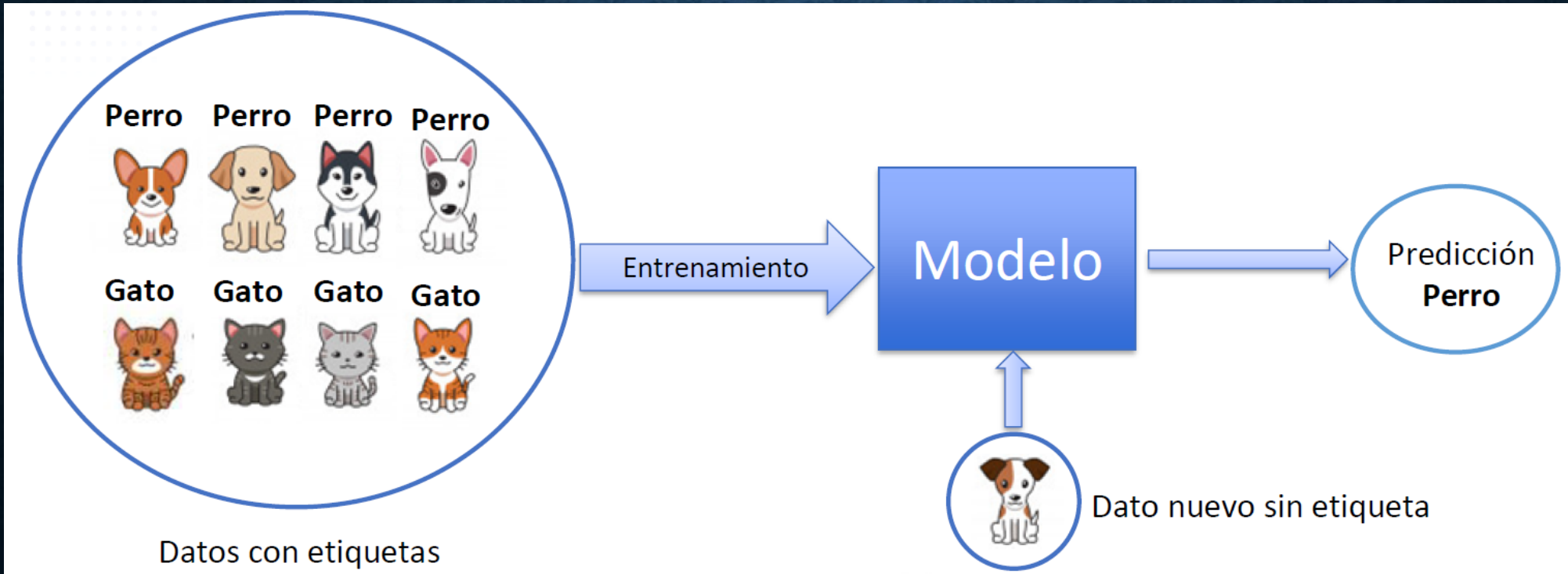
Que es Regresión Lineal?

La Regresión lineal es una técnica que se utiliza para predecir variables.

En otras palabras en un método que predice una variable dependiente (y) en función a los valores de las variables independientes (x).

COMO LO UTILIZAMOS?

- Lo podemos utilizar de varias formas uno de los ejemplos mas claros que vimos es la de los perros y los gatos.



CONCLUSIÓN

La regresión lineal la podemos utilizar para predecir algo. Para este mundo tan amplio hay un montón de ideas que pueden servir para hacer machine learnig.

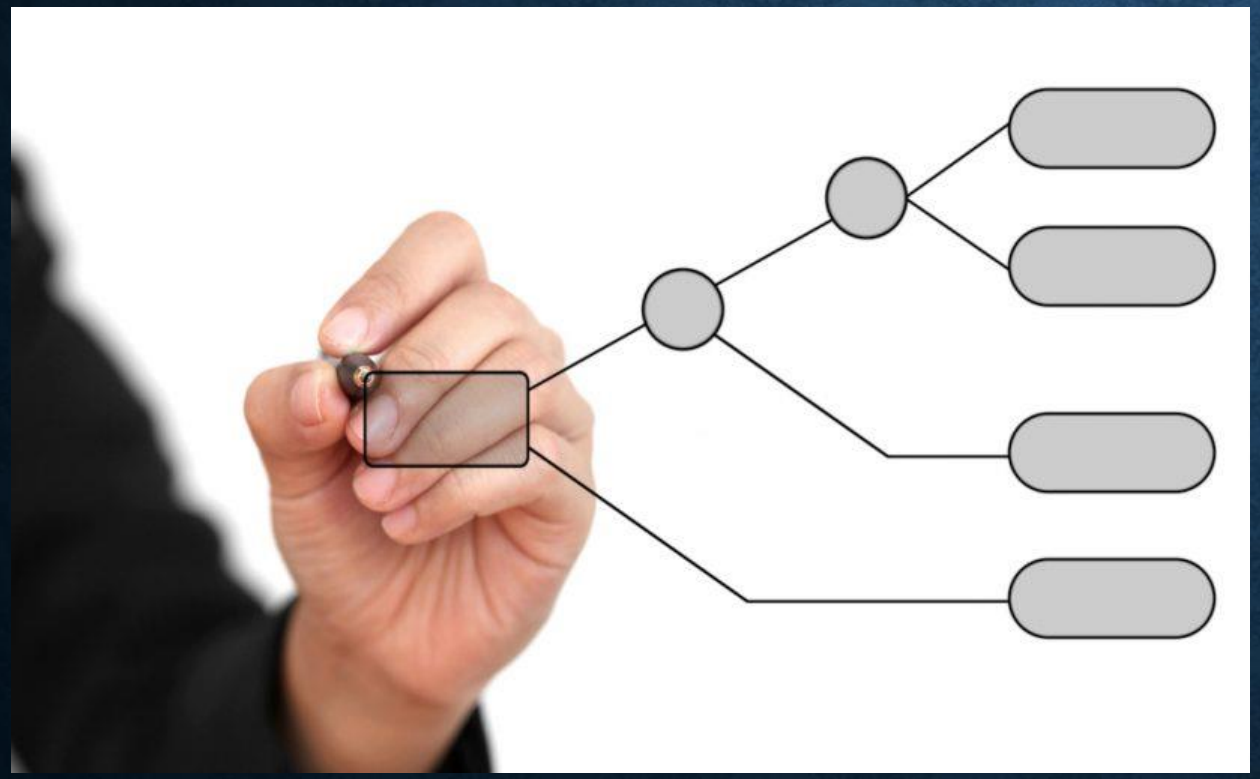
COSAS QUE VIMOS

Algo que vimos en el curso fue como utilizar git

Git es un software de control de versiones que se puede enlazar con github que es un lugar donde se pueden almacenar proyectos y se pueden compartir a la vez con otras personas.

Una vez nos

ÁRBOL DE DECISIONES

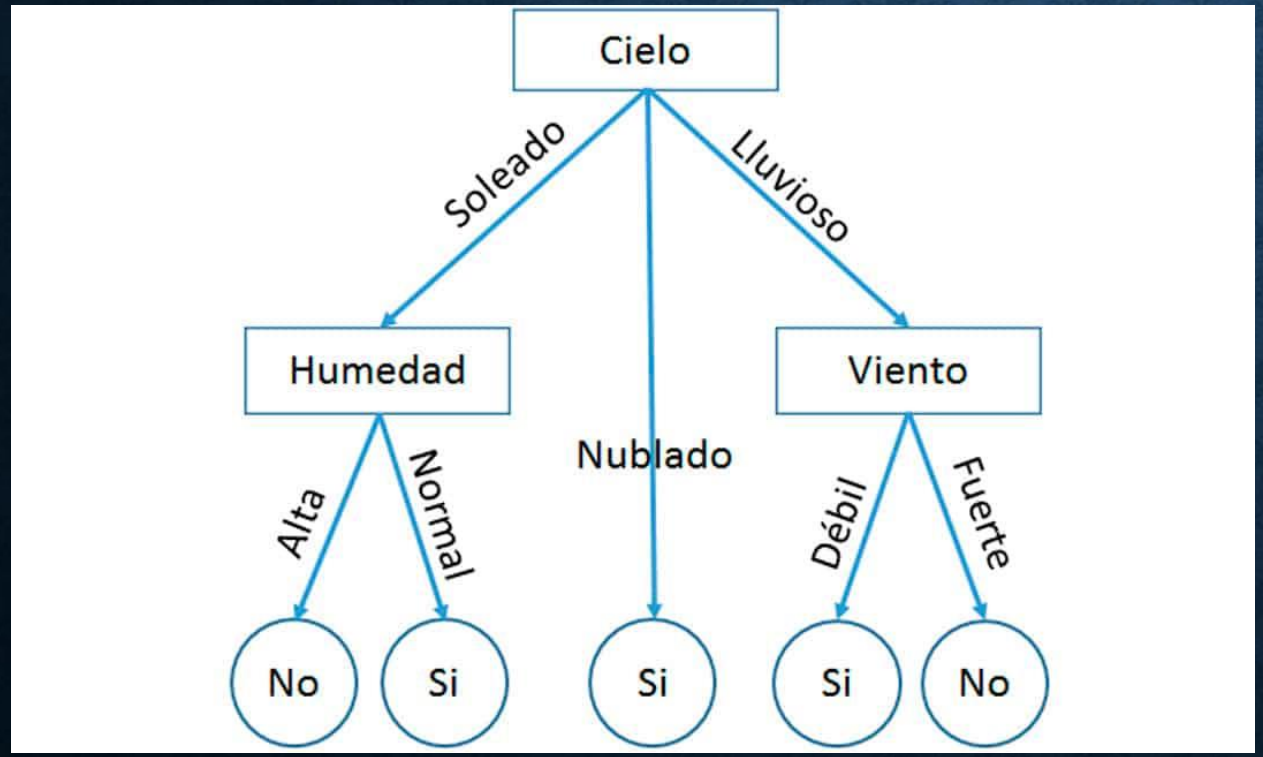


Que es el árbol de decisiones?

El árbol de decisiones es un modelo de predicciones que parte de una raíz y va haciendo preguntas de apoco al principio son preguntas genéricas y al final son preguntas mas especificas.

COMO LO UTILIZAMOS?

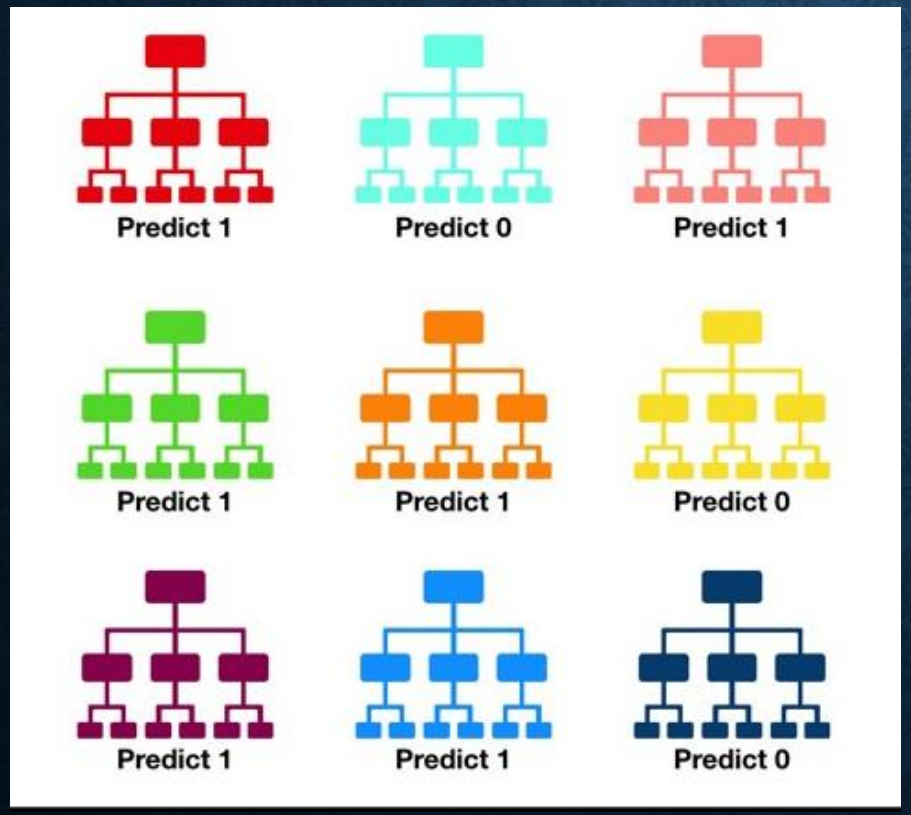
- Hay varias formas de utilizarlo esta es una de ellas



Este es un ejemplo bastante claro de como se puede implementar el árbol de decisiones

RAMDOM FOREST

- También vimos ramdom forest.



El ramdom forest es un conjunto de arboles de decisiones. Los arboles de decisiones son modelos rápidos en su entrenamiento y fáciles de interpretar pero muchas veces tienen overfitting (sobreajuste).

PROYECTO FINAL

Ideas:

+ La probabilidad de que crezca un árbol.
+ Precio de un diamante.
+ Probabilidad de ventas de una tienda.
+ Enfermedades crónicas.

IDEA QUE ELEGI

- La idea que elegí al final fue la de predecir el valor de los diamantes. Busque en diferentes paginas y al final encotre este dataset en kaggle

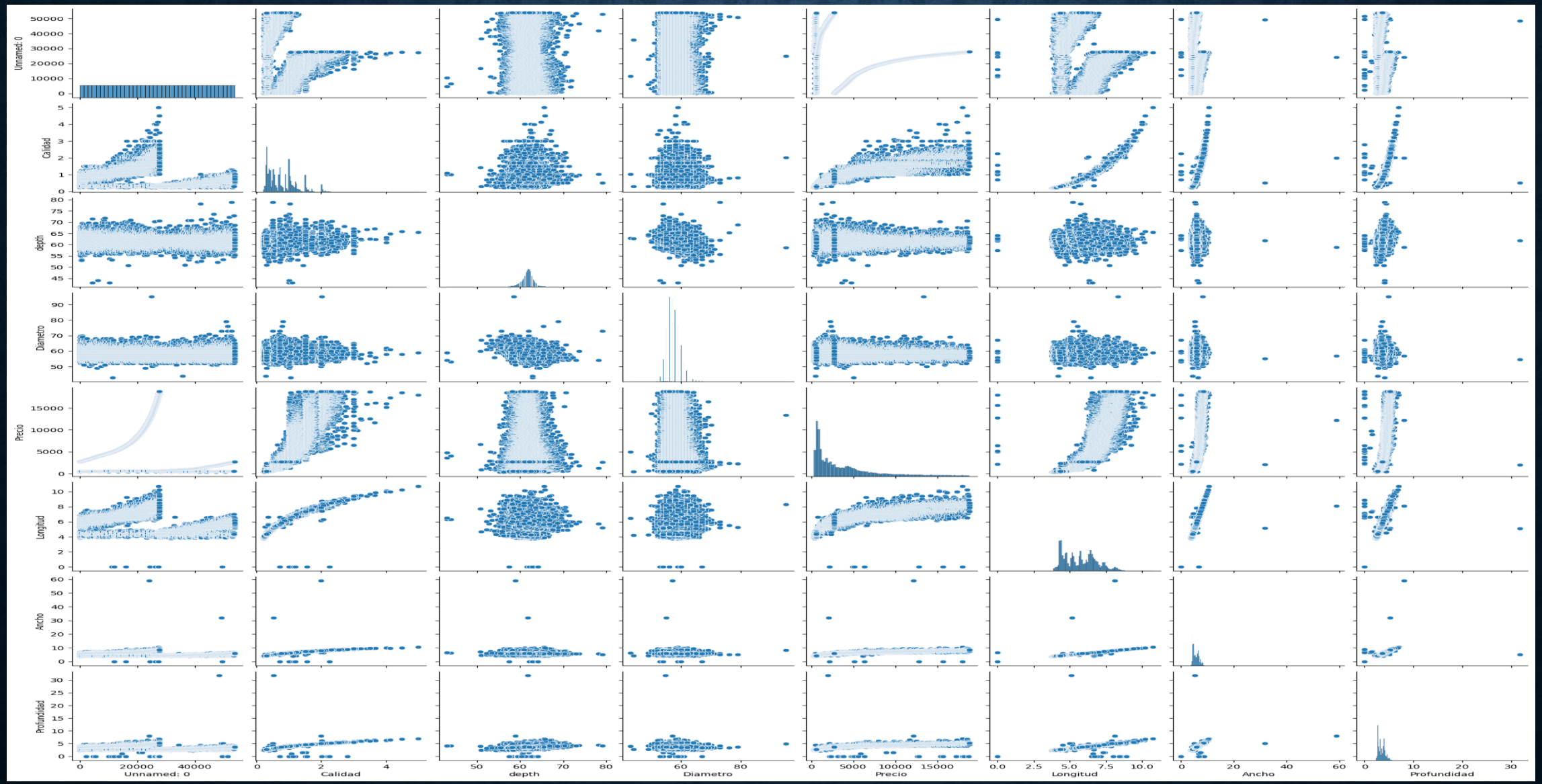
Diamonds | Kaggle

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
5	6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
6	7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
7	8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
8	9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
9	10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39

- Una vez que revise los datos por arriba empecé cambiando los nombres de las columnas traduciéndolas al español

	Unnamed: 0	Calidad	Corte	color	claridad	depth	Diametro	Precio	Longitud	Ancho	Profundidad
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...	...
53935	53936	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	53937	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	53938	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

Acá podemos ver la comparación de las columnas



LIMPIEZA DE DATOS

- Me fije si las columnas tenían datos null

```
data.isnull().sum()
```

```
Unnamed: 0     0
Calidad        0
Corte          0
color          0
claridad       0
depth          0
Diametro       0
Precio         0
Longitud       0
Ancho          0
Profundidad    0
dtype: int64
```

LIMPIEZA

- Ya que no hay datos null empiezo eliminar las columnas que no me sirven.

	Calidad	Corte	color	claridad	Diametro	Precio	Longitud	Ancho	Profundidad
0	0.23	Ideal	E	SI2	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...
53935	0.72	Ideal	D	SI1	57.0	2757	5.75	5.76	3.50
53936	0.72	Good	D	SI1	55.0	2757	5.69	5.75	3.61
53937	0.70	Very Good	D	SI1	60.0	2757	5.66	5.68	3.56
53938	0.86	Premium	H	SI2	58.0	2757	6.15	6.12	3.74
53939	0.75	Ideal	D	SI2	55.0	2757	5.83	5.87	3.64

PASAJE DE DATOS

- Los arboles y las regresiones lineales no pueden ser manipuladas con datos char ni datos de tipo string entonces lo pasamos a int con el método `get_dummies.`

	Corte_Fair	Corte_Good	Corte_Ideal	Corte_Premium	Corte_Very Good
0	0	0	1	0	0
1	0	0	0	1	0
2	0	1	0	0	0
3	0	0	0	1	0
4	0	1	0	0	0
...	...	...	...	...	...
53935	0	0	1	0	0
53936	0	1	0	0	0
53937	0	0	0	0	1
53938	0	0	0	1	0
53939	0	0	1	0	0

El método get_dummies separa los datos únicos de la columna y hace una columna para cada 1 para identificarlos con 0 y 1 si es 1 es que es ese dato.

LUEGO DE UTILIZAR EL GET_DUMMIES

	Calidad	Diametro	Precio	Longitud	Ancho	Profundidad	Corte_Fair	Corte_Good	Corte_Ideal	Corte_Premium	...	calridad_VS2	calridad_VVS1	calridad_VVS2	color_D	color_E	color_F	color_G	color_H	color_I	color_J
0	0.23	55.0	326	3.95	3.98	2.43	0	0	1	0	...	0	0	0	0	1	0	0	0	0	0
1	0.21	61.0	326	3.89	3.84	2.31	0	0	0	1	...	0	0	0	0	1	0	0	0	0	0
2	0.23	65.0	327	4.05	4.07	2.31	0	1	0	0	...	0	0	0	0	1	0	0	0	0	0
3	0.29	58.0	334	4.20	4.23	2.63	0	0	0	1	...	1	0	0	0	0	0	0	0	1	0
4	0.31	58.0	335	4.34	4.35	2.75	0	1	0	0	...	0	0	0	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
53935	0.72	57.0	2757	5.75	5.76	3.50	0	0	1	0	...	0	0	0	1	0	0	0	0	0	0
53936	0.72	55.0	2757	5.69	5.75	3.61	0	1	0	0	...	0	0	0	1	0	0	0	0	0	0
53937	0.70	60.0	2757	5.66	5.68	3.56	0	0	0	0	...	0	0	0	1	0	0	0	0	0	0
53938	0.86	58.0	2757	6.15	6.12	3.74	0	0	0	1	...	0	0	0	0	0	0	0	1	0	0
53939	0.75	55.0	2757	5.83	5.87	3.64	0	0	1	0	...	0	0	0	1	0	0	0	0	0	0

53940 rows × 26 columns

PRUEBA DE ARBOLES

- Ya que pase todos las columnas a int empiezo a hacer las pruebas

Error absoluto medio

```
125284.18531102511
```

Error cuadrático medio

```
1298802.6734890619
```

REGRESION LINEAL

```
La Media del Error Absoluto  del modelo es 748.71
La Media del Error cuadrático del modelo es 1285918.02
La raíz del error cuadrático medio del modelo es 1133.98
El R2 del modelo es 0.92
```

	hiperparametros_default	sin_intercepto
mae	7.487138e+02	7.487138e+02
mse	1.285918e+06	1.285918e+06
rmse	1.133983e+03	1.133983e+03
r2	9.200687e-01	9.200687e-01