



PREDICTION WITH MULTIPLE REGRESSION

RED WINE QUALITY

Facundo Sanabria

Presentación y consideraciones

Objetivo: Generar un modelo de regresión múltiple que asigne a cada vino, según su composición, el puntaje que recibiría por parte de catadores expertos en esta variedad.

Elección del dataset

Selección de preditore

Generación del modelo

Obtención y análisis de métricas

Selección de predictores

Selección 1 - Relevancia en el sector vitivinícola

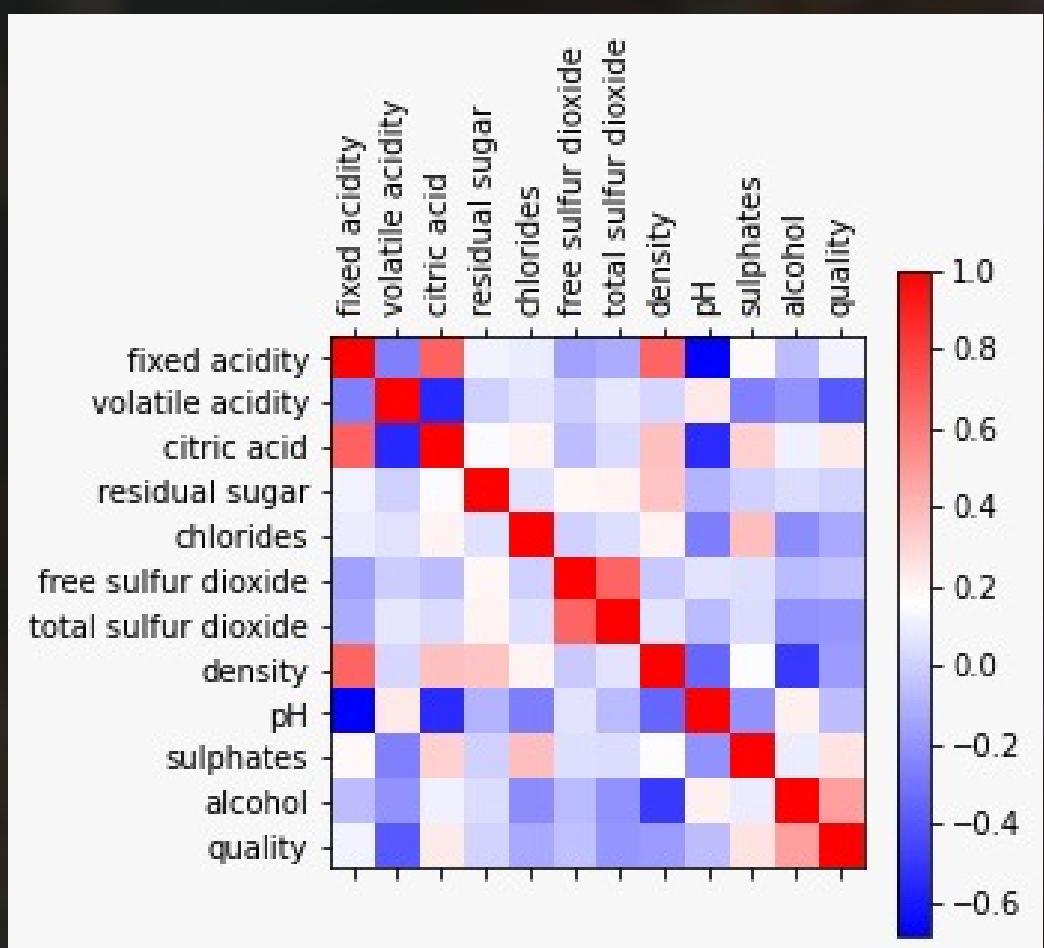
Según se investigó las características fundamentales de un buen vino son: dulzor, acidez, tanino, ácido acético y alcohol.

Selección 2 - Correlación entre variables

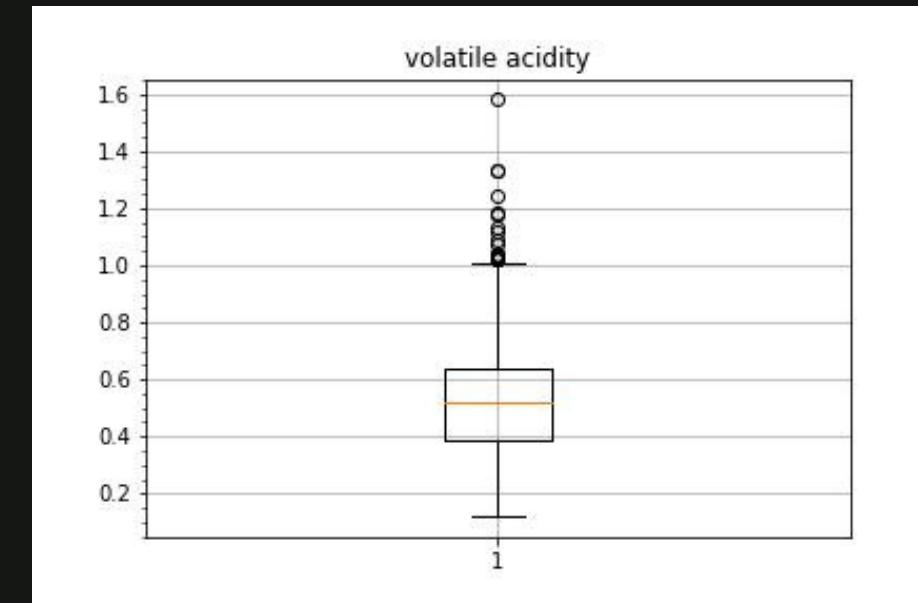
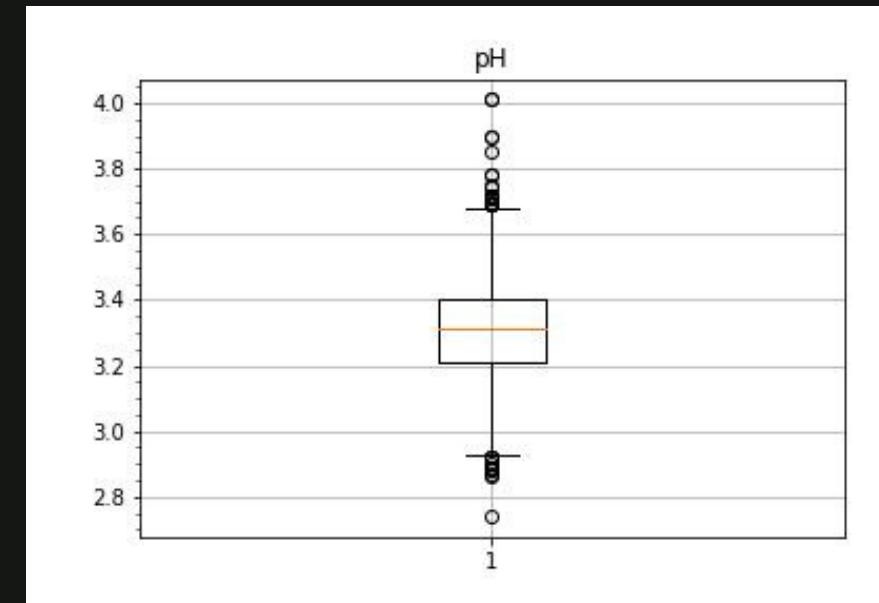
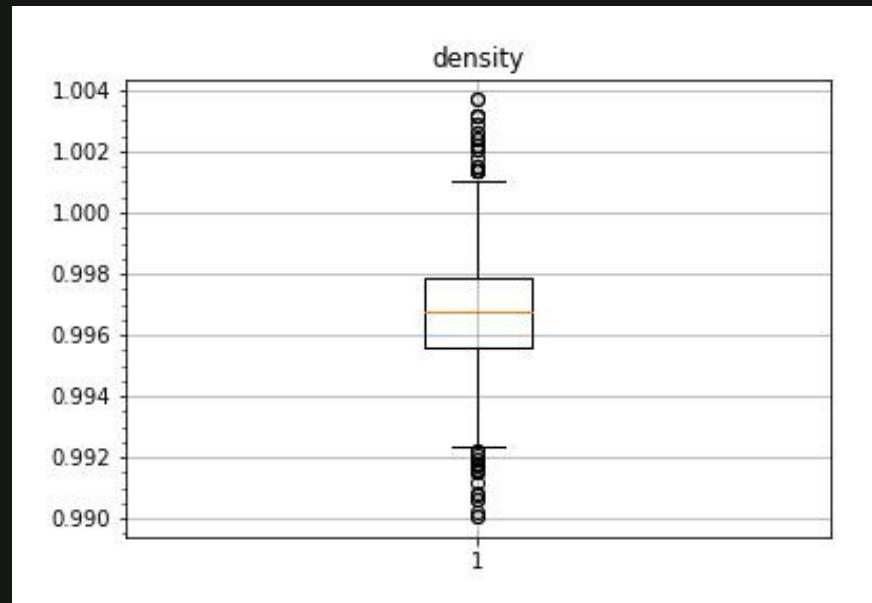
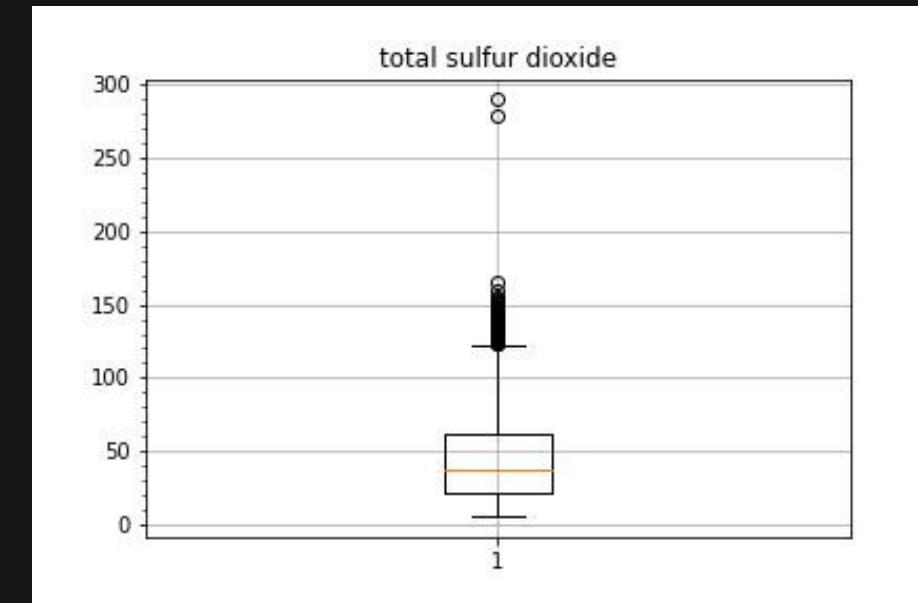
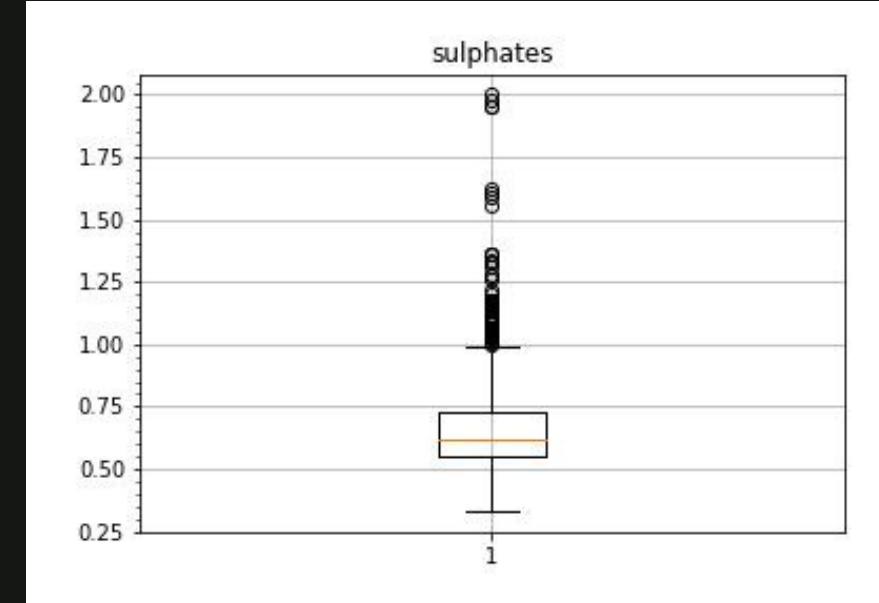
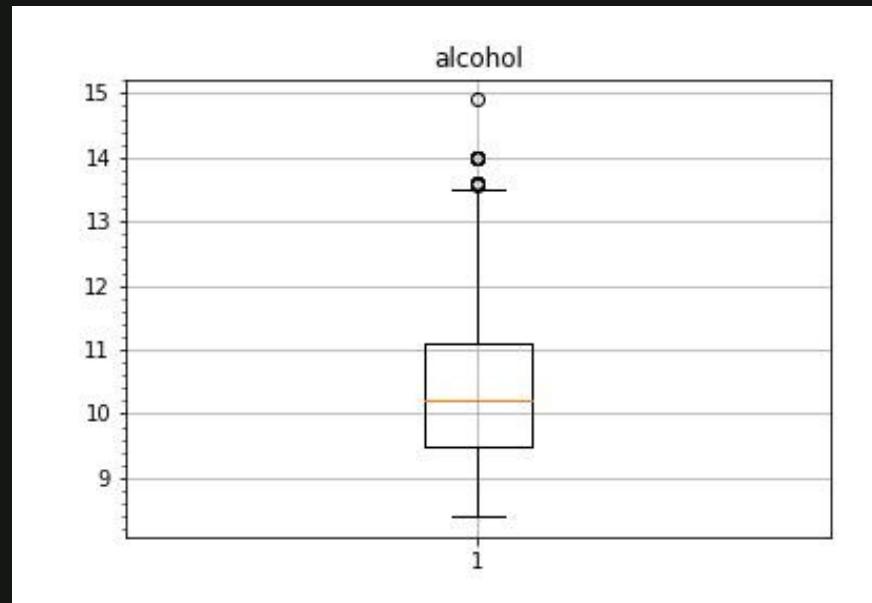
Analizamos la correlación a través del método `.corr()` y un mapa de color para una interpretación mas fluida.

Selección 3 - Stepwise Selection

Con el método mencionado, realizamos una selección en base a los datos estadísticos de cada atributo para con el modelo.



Predictores elegidos: Análisis



Luego de definir quienes van a ser nuestros predictores para el modelo, se reviso su distribución en busca de outliers para ser limpiados, y así proceder con la generación del modelo.

Modelo principal

OLS Regression Results									
Dep. Variable:	quality	R-squared:	0.366						
Model:	OLS	Adj. R-squared:	0.363						
Method:	Least Squares	F-statistic:	136.1						
Date:	Wed, 25 May 2022	Prob (F-statistic):	3.33e-136						
Time:	12:06:03	Log-Likelihood:	-1369.7						
No. Observations:	1422	AIC:	2753.						
Df Residuals:	1415	BIC:	2790.						
Df Model:	6								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			
volatile acidity	-0.9064	0.107	-8.461	0.000	-1.117	-0.696			
total sulfur dioxide	-0.0020	0.001	-3.027	0.003	-0.003	-0.001			
alcohol	0.2702	0.021	12.587	0.000	0.228	0.312			
density	-18.2525	12.715	-1.435	0.151	-43.195	6.690			
pH	-0.5650	0.134	-4.221	0.000	-0.828	-0.302			
sulphates	1.6248	0.161	10.121	0.000	1.310	1.940			
const	22.4191	12.873	1.742	0.082	-2.834	47.672			
Omnibus:	27.137	Durbin-Watson:	1.769						
Prob(Omnibus):	0.000	Jarque-Bera (JB):	38.766						
Skew:	-0.203		Prob(JB):	3.82e-09					
Kurtosis:	3.700		Cond. No.	5.47e+04					

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.47e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Modelo - Selección 1

Score: 0.3209

Modelo - Selección 2

Score: 0.3581

Modelo - Selección 3

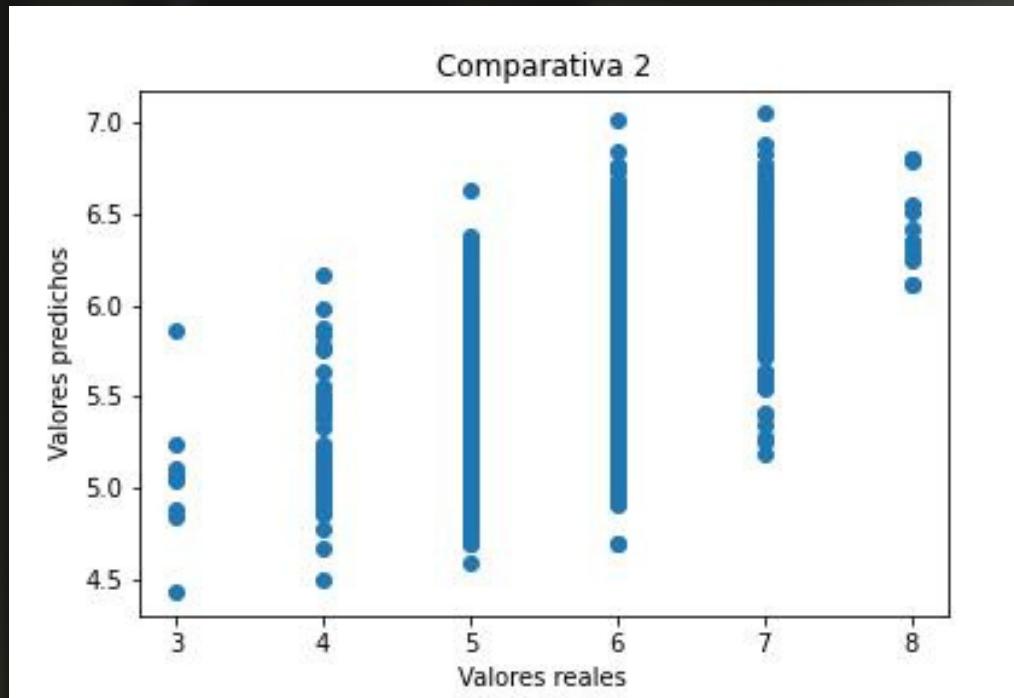
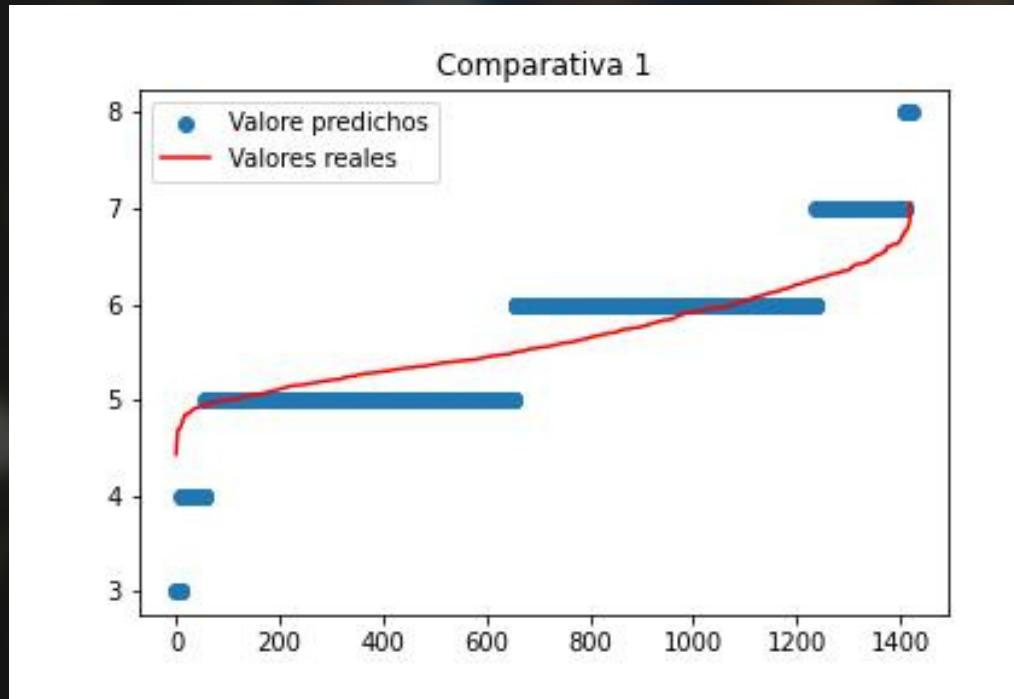
Score: 0.3708

Modelo - All features

Score: 0.3732

Como nuestra puntuación está entre las demás, concluimos que es la correcta. No elegiremos trabajar con todas las variables ya que su mejor puntuación se debe a un overfitting del modelo.

Valores predichos vs valores reales



Análisis

El modelo se ajusta bastante a las pendientes presentadas en los datos, pero debido a la gran cantidad de registros con puntuaciones medias, la recta del modelo no logra ajustarse correctamente.

Cross Validation & Confidence Intervals

¿Que se podrá esperar?

Cross Validation Result:

```
0.24167288013297927  
0.35514341368427704  
0.3952779498392931  
0.3736510891246756  
0.2544798969225658
```

Valor de la media: 0.3240450459407581

Confidence intervals:

volatile acidity	-1.031453	volatile acidity	-0.779890
total sulfur dioxide	-0.002649	total sulfur dioxide	-0.001296
alcohol	0.245330	alcohol	0.293248
density	-32.416357	density	-4.746803
pH	-0.706802	pH	-0.408709
sulphates	1.453195	sulphates	1.788076

Name: 0.05, dtype: float64 Name: 0.95, dtype: float64

Análisis

Al igual que trabajar solo con los extremos, al dividir nuestros datos en grupo tan pequeños, homo/heterogenos, las metricas se dispersan demasiado.

Conclusión

36.6%

Score del modelo de regresión múltiple

Apesar del conocimiento del sector, limpieza de outliers, y de ser ajustado según las estadísticas de sus variables, debido a su bajo rendimiento, no se recomienda este tipo de modelo para la predicción de la calidad de vinos.

Muchas gracias por su atención

Facundo Sanabria