

Trabajo Práctico Final

Modelo para Predecir Periodos de órbitas de
Asteroides

POR: ALEJANDRO ECHEVERRI, NICOLAS PONTIROLI

07/06/2024





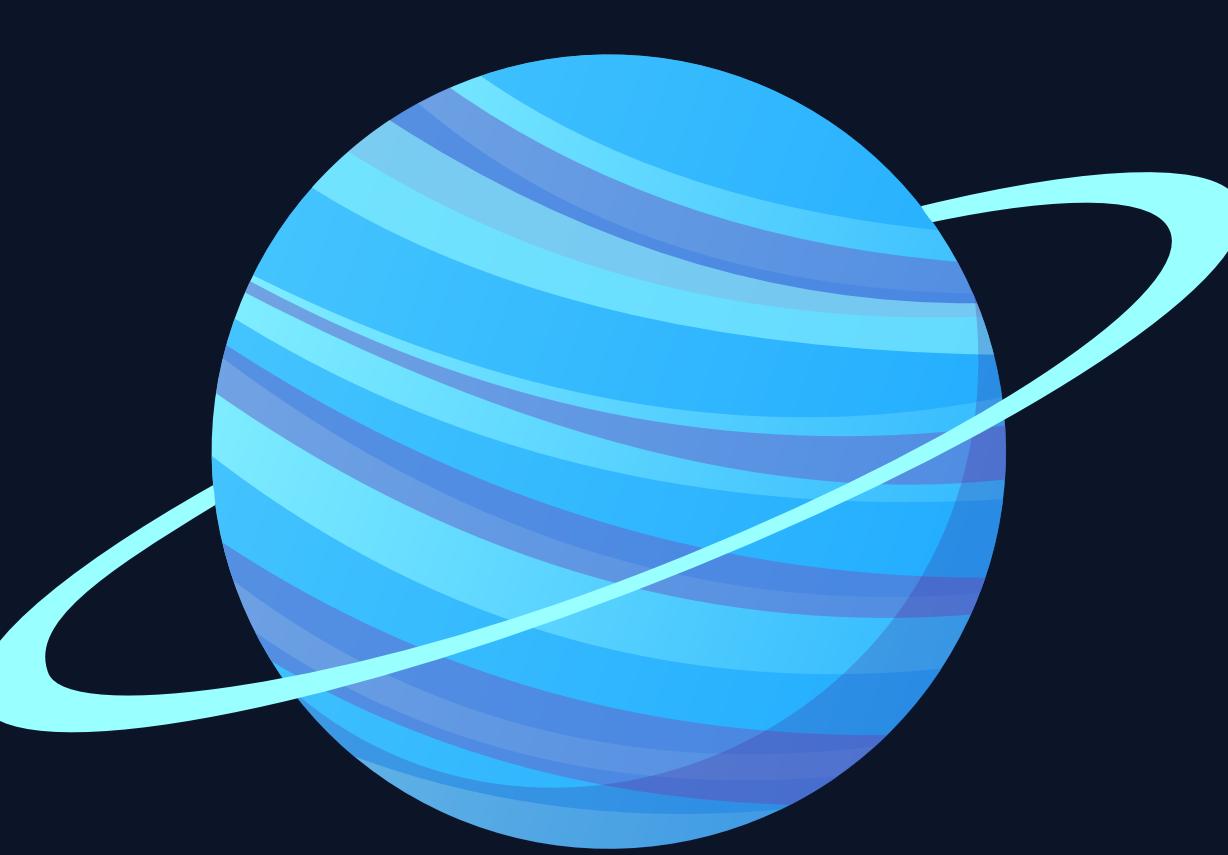
ÍNDICE

03. Introducción al proyecto.
04. Datos Utilizados
05. Exploración de Datos (EDA)
06. Preprocesamiento de Datos
07. Modelado y Evaluación
08. Análisis de Residuos
09. Mejor Modelo
12. Modelo con Polinomios Fraccionarios
13. Mejor Coeficiente
14. Relación Causal y Kepler
16. Residuos Modelo con Kepler
17. Conclusiones
18. Agradecimientos y Preguntas

INTRODUCCIÓN

Estamos trabajando en predecir el período orbital de los asteroides utilizando características físicas y orbitales, como el semieje mayor, la excentricidad y la inclinación. Esto nos ayuda a seguir sus trayectorias y evaluar riesgos de impacto con la Tierra.

Usamos técnicas de Machine Learning para identificar las características más relevantes y construir un modelo preciso. Este proyecto nos permite practicar habilidades en Ciencia de Datos y entender mejor la dinámica de los asteroides. Aquí mostramos nuestros hallazgos preliminares y el proceso seguido hasta ahora.



DATOS UTILIZADOS

Todos los datos provienen de (<http://neo.jpl.nasa.gov/>) . Esta API es mantenida por el equipo de SpaceRocks: David Greenfield, Arezu Sarvestani, Jason English y Peter Baunach.

Tomaremos como target a la variable per por lo tanto eliminaremos el feature per_y ya que por definición es el mismo atributo en distintas variaciones temporales. Originalmente tiene 31 columnas pero por hipótesis y posibles leaks nos quedamos con 13 columnas y 839714 filas.

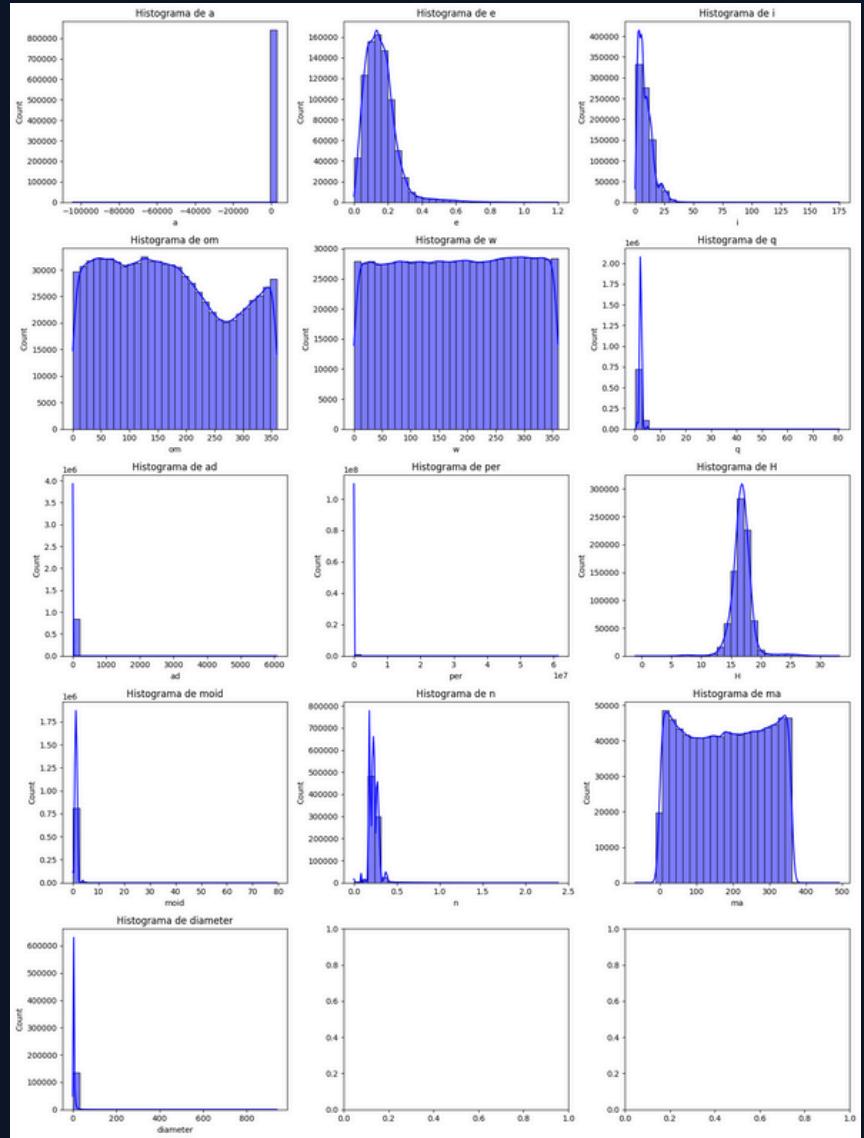
1 UA equivale a 150×10^6 KM

a	semieje_mayor	ad	distancia_afelio
e	excentricidad	H	magnitud_absoluta
i	inclinación	diameter	diámetro
om	longitud_nodo_ascendente	moid	distancia_intersección_orbital_Tierra
w	argumento_perihelio	n	argumento_perihelio
q	distancia_perihelio	ma	anomalía_media

float64

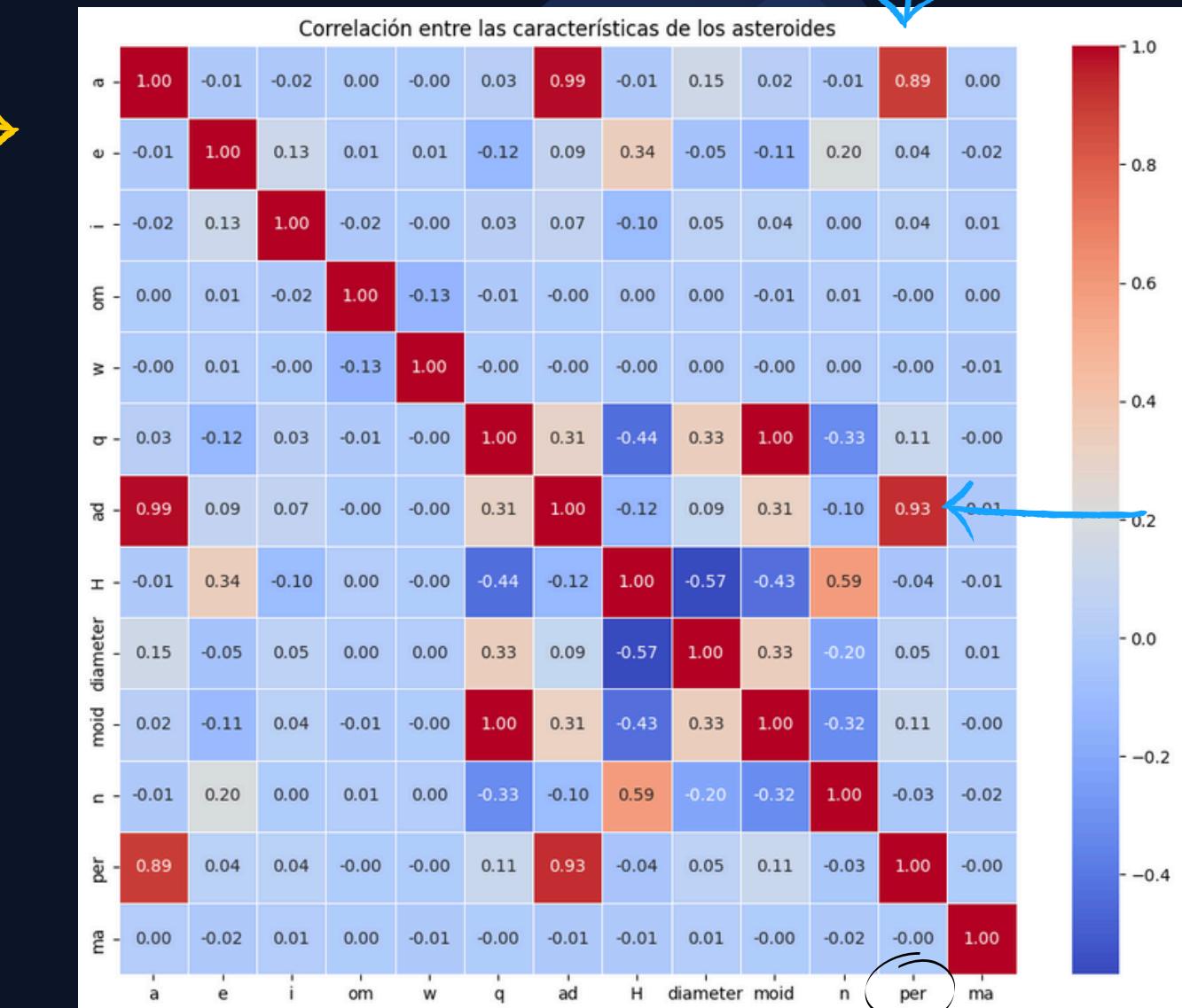
EXPLORACIÓN DE DATOS

	a	e	i	om	w	q	ad	H	diameter	moid	n	per	ma
count	839712.000000	839714.000000	839714.000000	839714.000000	839714.000000	839708.000000	837025.000000	137635.000000	8.232720e+05	8.397120e+05	8.397080e+05	839706.000000	
mean	2.757514	0.155636	8.949826	168.499466	181.075796	2.404728	3.385710	16.786249	5.481626	1.423371e+00	2.374145e-01	2.505533e+03	180.659892
std	114.384959	0.093897	6.666087	103.096307	104.023854	2.233172	12.748733	1.821574	9.366928	2.250450e+00	8.095014e-02	9.213979e+04	106.562235
min	-104279.220927	0.000000	0.007546	0.000388	0.001666	0.070511	0.773684	-1.100000	0.002500	3.437640e-07	2.926897e-08	1.511339e+02	-67.136826
25%	2.385258	0.091454	4.069077	80.211400	91.041603	1.971941	2.775350	15.900000	2.770000	9.784998e-01	1.900553e-01	1.345555e+03	86.642618
50%	2.644219	0.143655	7.257101	160.294860	181.669478	2.225510	3.037761	16.800000	3.956000	1.237810e+00	2.292228e-01	1.570524e+03	181.517775
75%	2.996048	0.199400	12.255653	252.201519	271.521717	2.578162	3.357967	17.600000	5.741500	1.590560e+00	2.675475e-01	1.894184e+03	274.301731
max	3043.149073	1.201134	175.188725	359.999800	359.999833	80.424175	6081.841956	33.200000	939.400000	7.950130e+01	2.381994e+00	6.131733e+07	491.618014

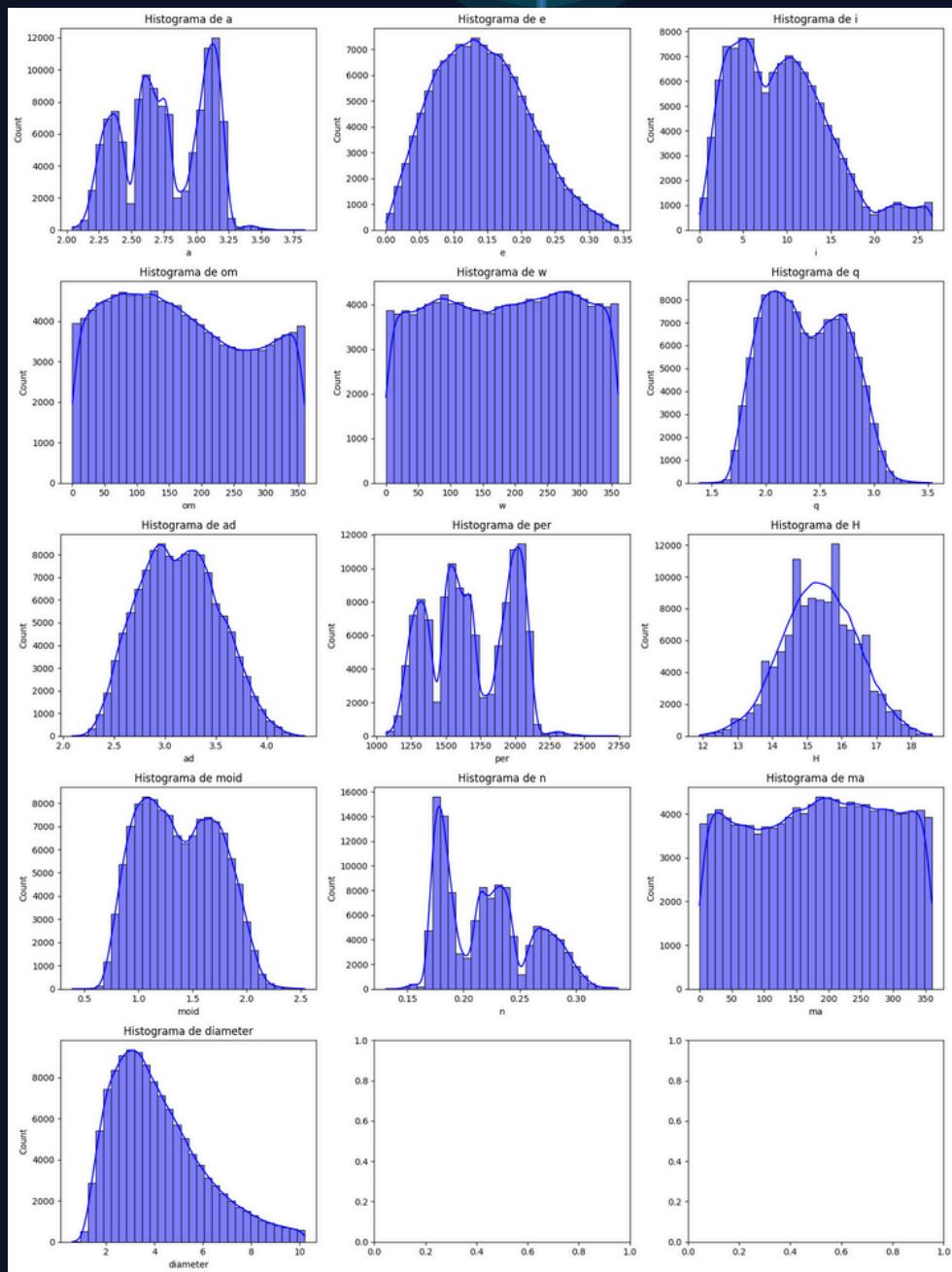


Observamos una alta correlación de a y ad con nuestro target

Mucha varianza



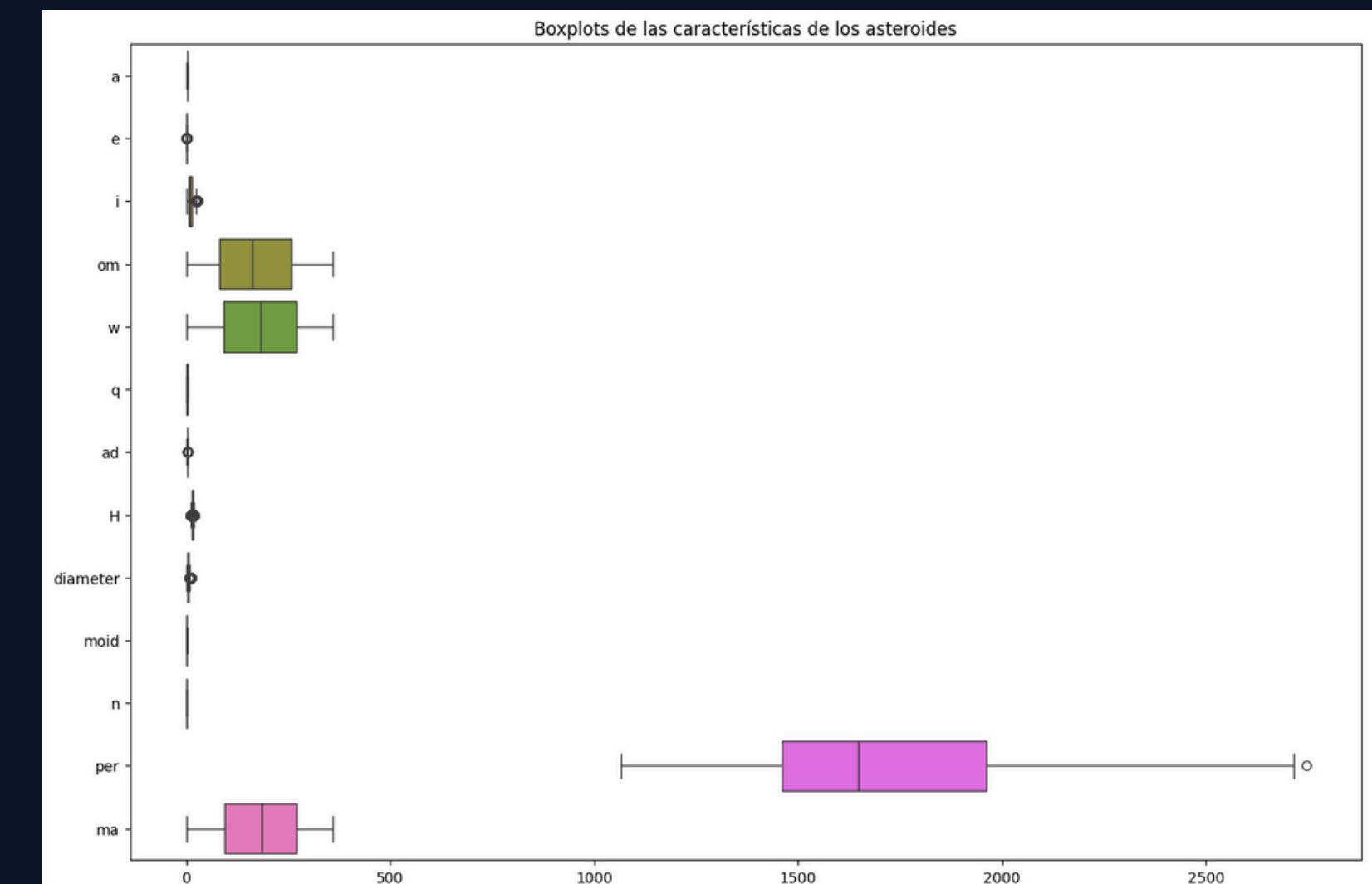
PREPROCESAMIENTO DE DATOS



0	a	120435	non-null	float64
1	e	120435	non-null	float64
2	i	120435	non-null	float64
3	om	120435	non-null	float64
4	w	120435	non-null	float64
5	q	120435	non-null	float64
6	ad	120435	non-null	float64
7	H	120435	non-null	float64
8	diameter	120435	non-null	float64
9	moid	120435	non-null	float64
10	n	120435	non-null	float64
11	per	120435	non-null	float64
12	ma	120435	non-null	float64

Quitamos NA y outliers nos quedamos con 120.435 instancias

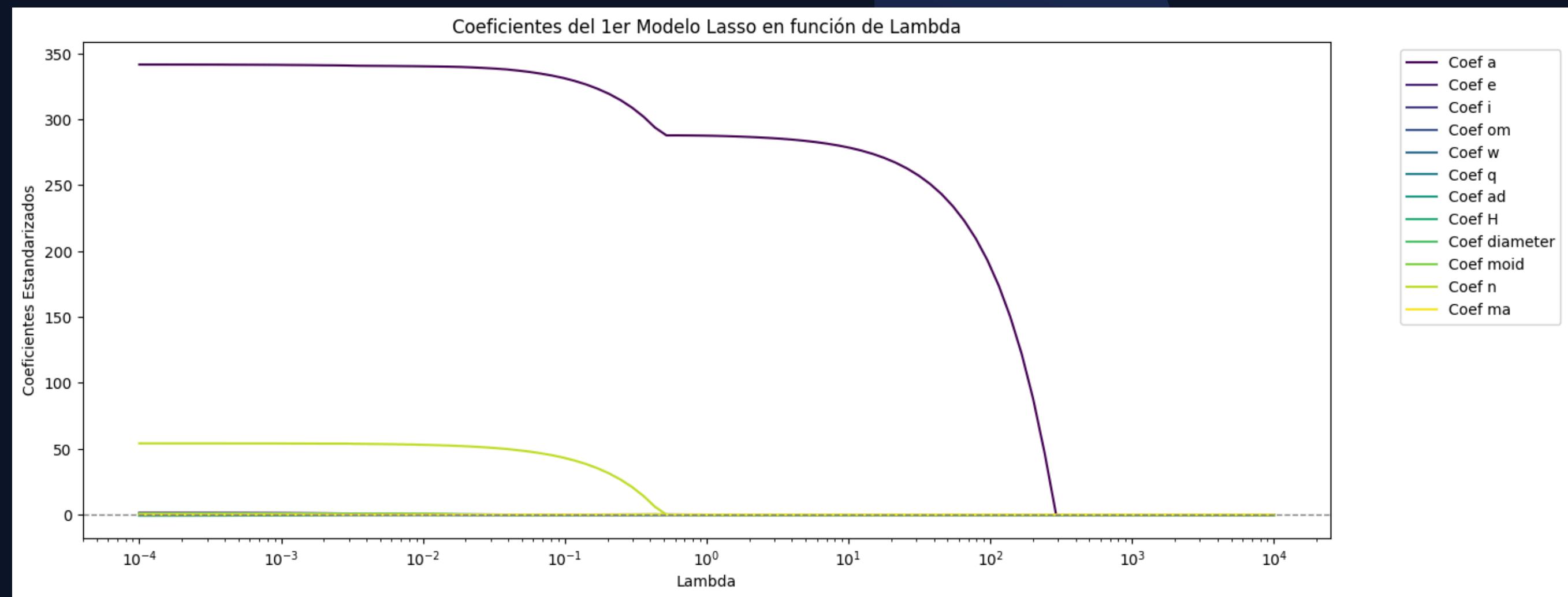
Quitamos los outliers →



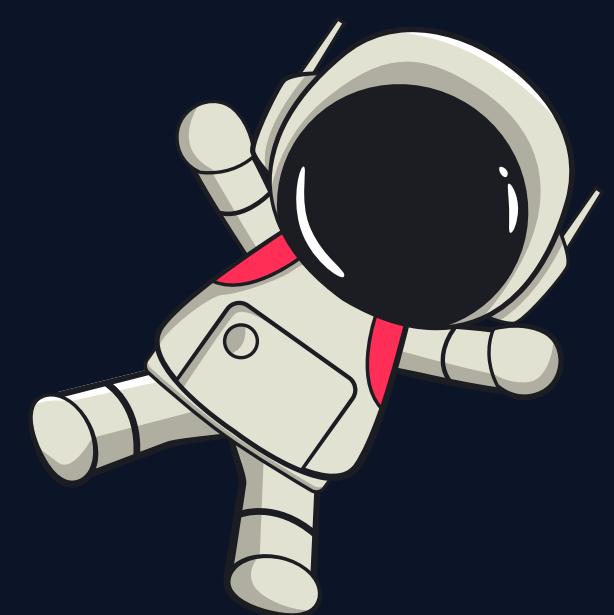
← Mejoró la varianza

MODELADO Y EVALUACIÓN

Primer Modelo: Probemos lo simple...

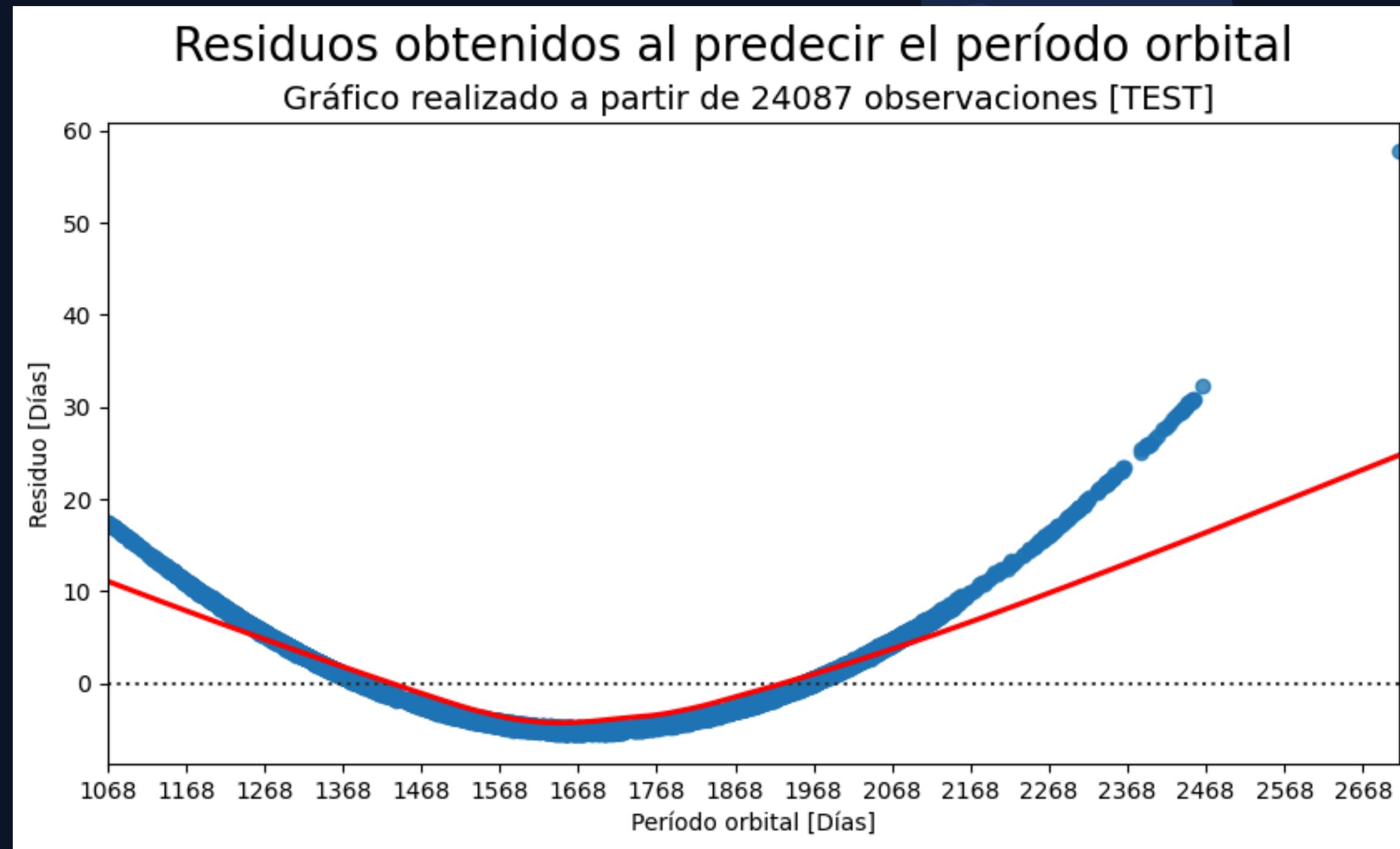


Utilizando LassoCV elegimos el mejor Lambda

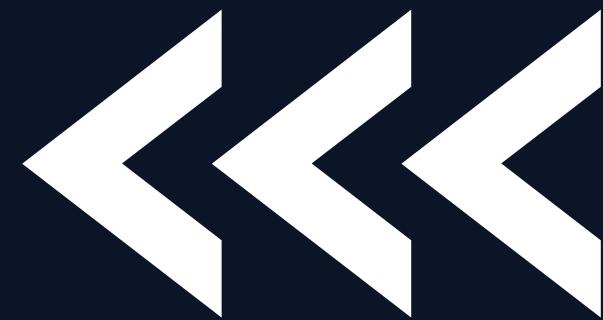


MODELADO Y EVALUACIÓN

Residuos del primer modelo

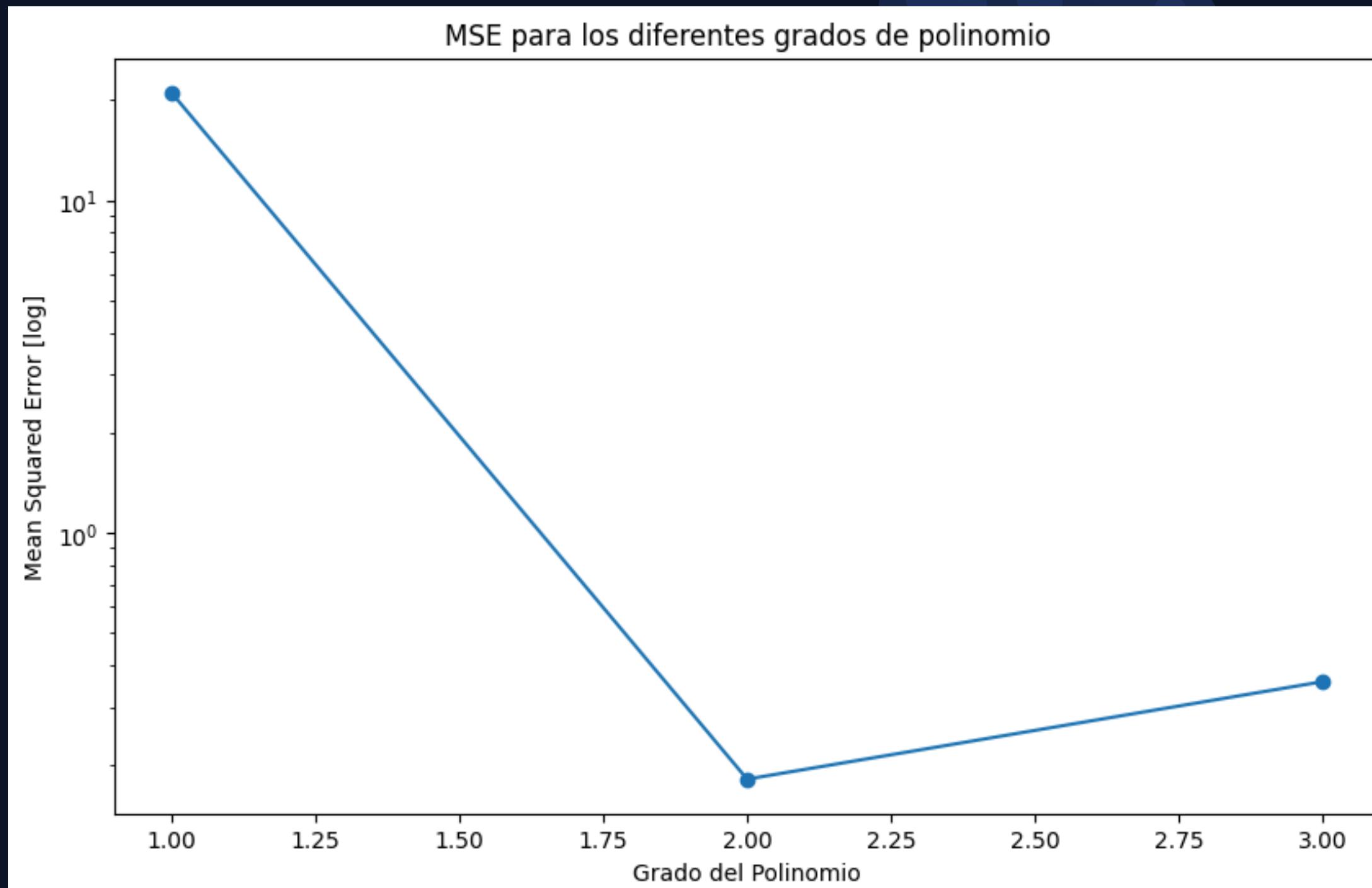


Los residuos tienen tendencias



MODELADO Y EVALUACIÓN

Intentamos con modelos LassoCV polinómicos:

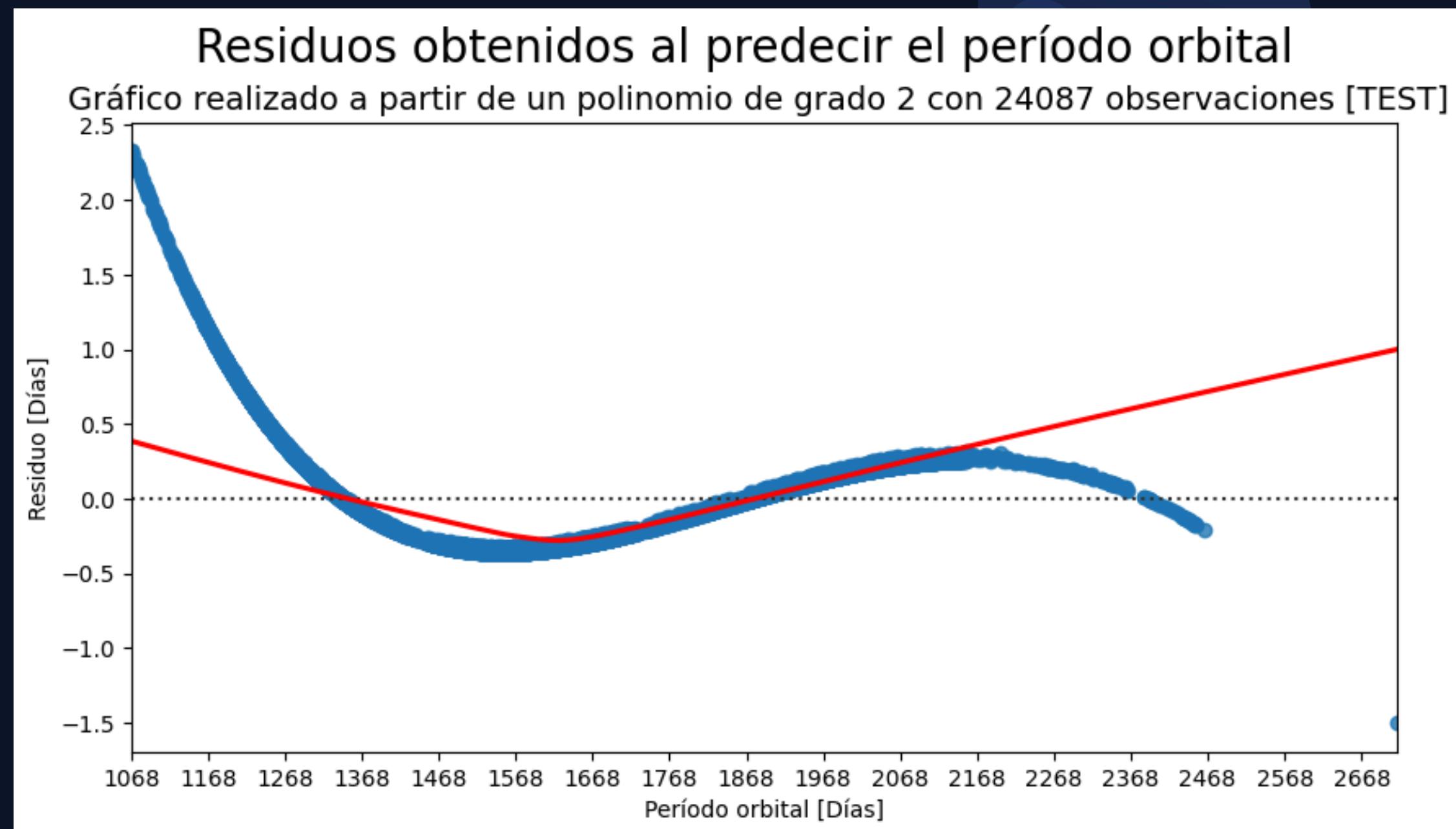


Grado 1: MSE	21.0590560589443
Grado 2: MSE	0.18197008618362143
Grado 3: MSE	0.3574131432804117

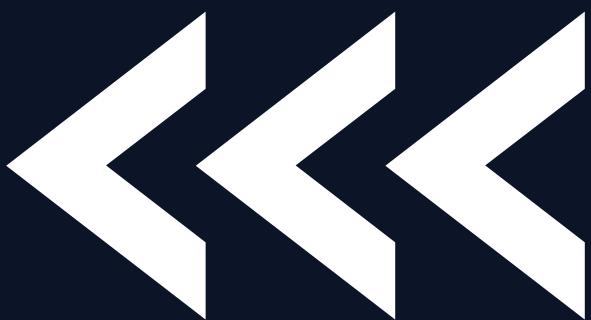


MODELADO Y EVALUACIÓN

Residuos del modelo con coeficiente polinómico de grado 2

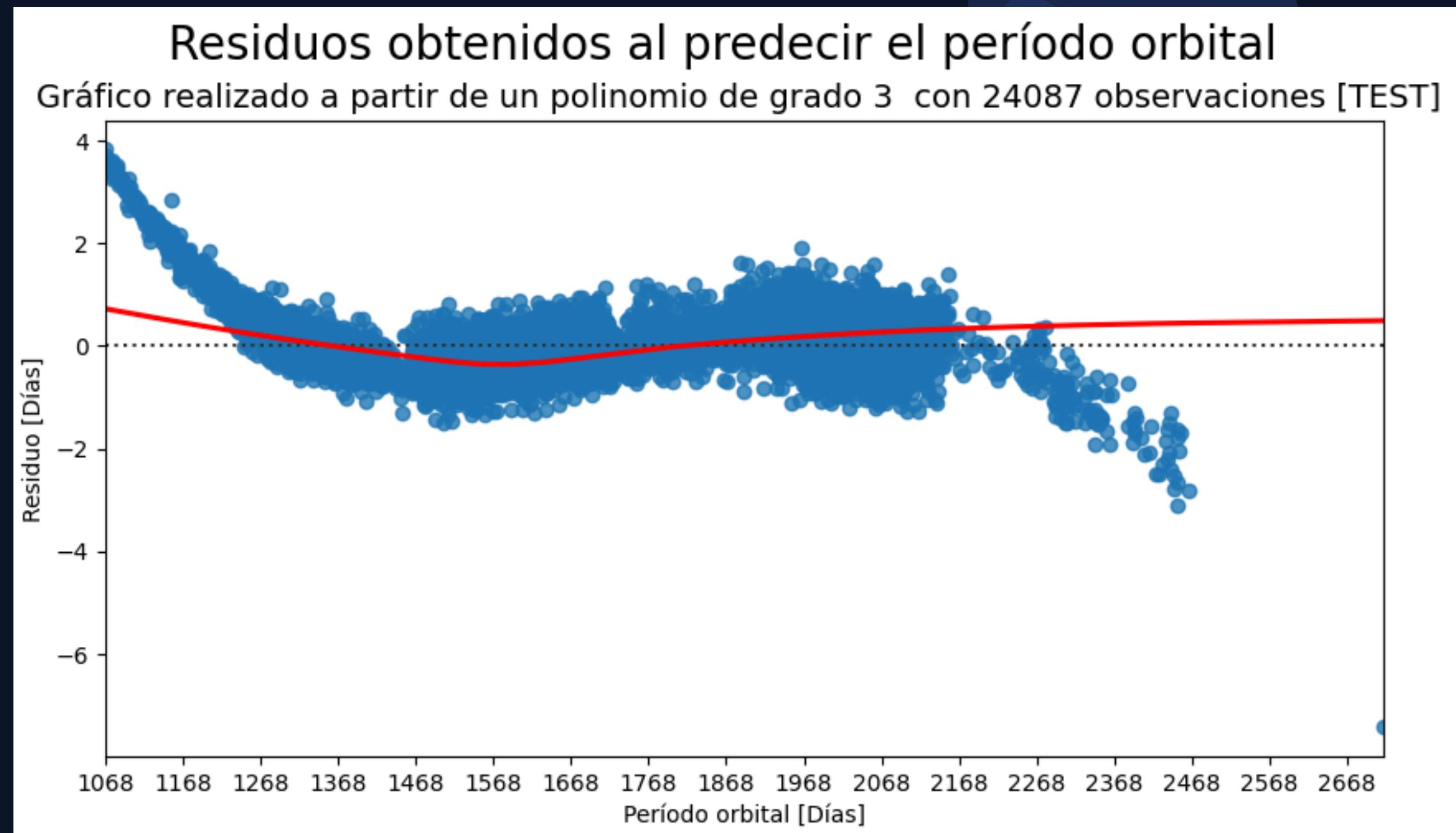


Los residuos tienen tendencias



MODELADO Y EVALUACIÓN

Residuos del modelo con coeficiente polinómico de grado 3

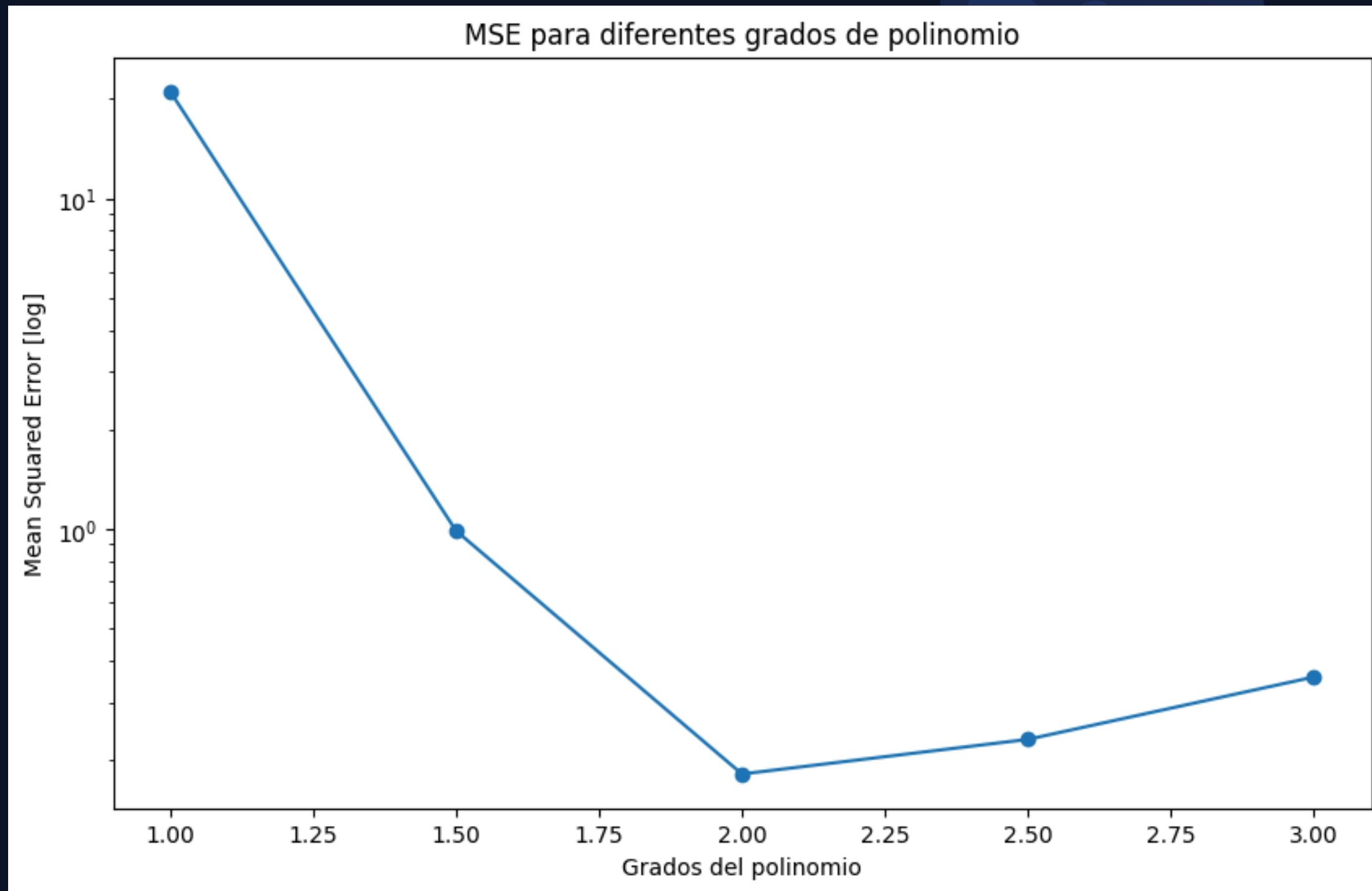


Los residuos tienen tendencias



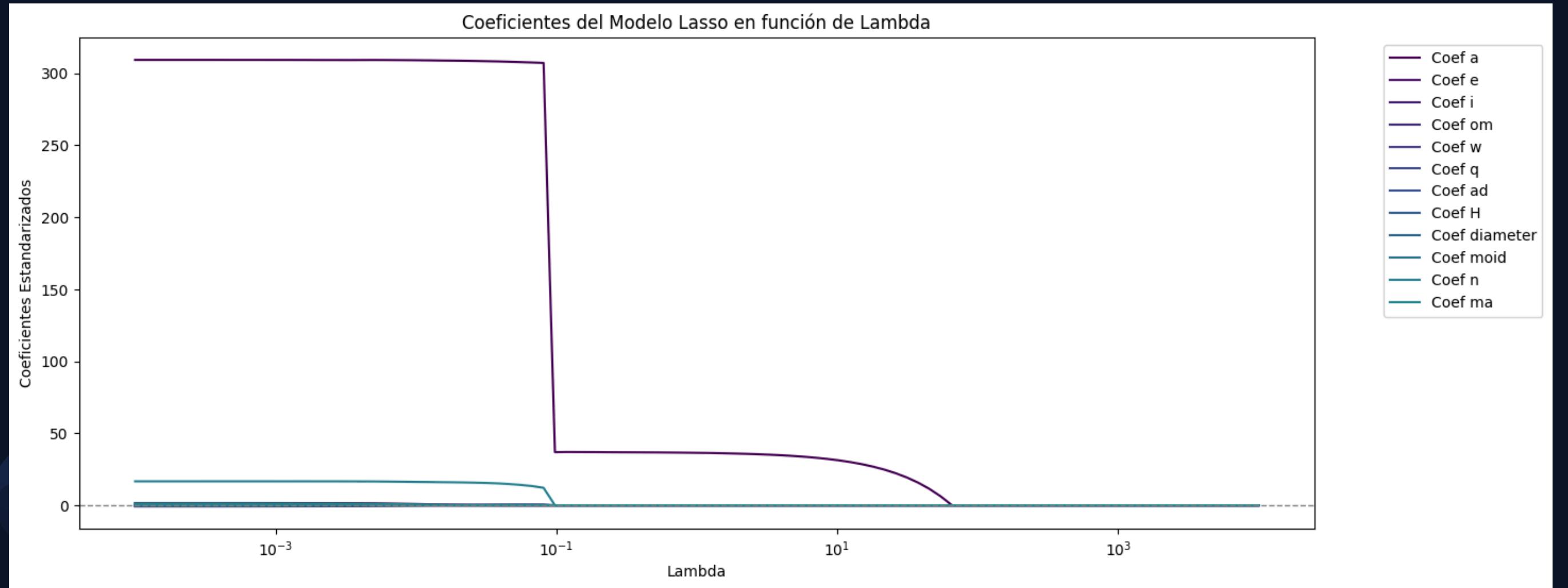
MODELADO Y EVALUACIÓN

Probamos con modelos con polinomios fraccionarios



Grado 1: MSE	21.0590560589443
Grado 1.5: MSE	0.9890734919153321
Grado 2: MSE	0.18197008618362143
Grado 2.5: MSE	0.2315864722851204
Grado 3: MSE	0.3574131432804117

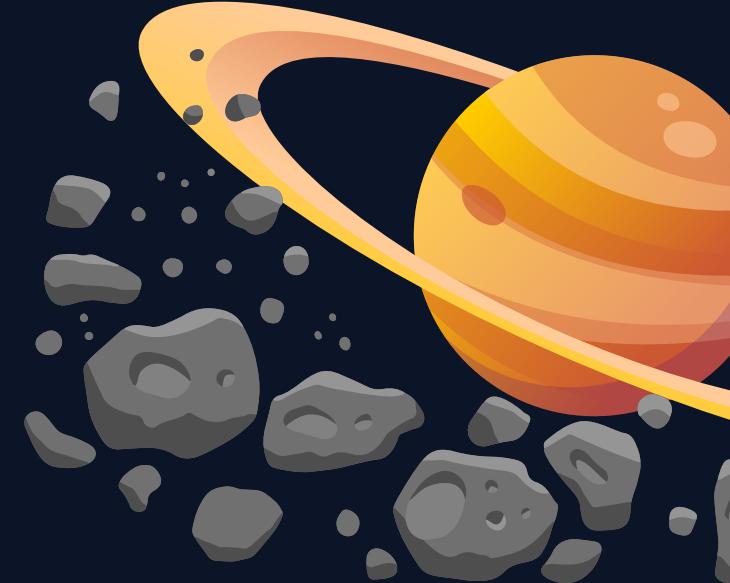
MEJOR COEFICIENTE: Semieje Mayor



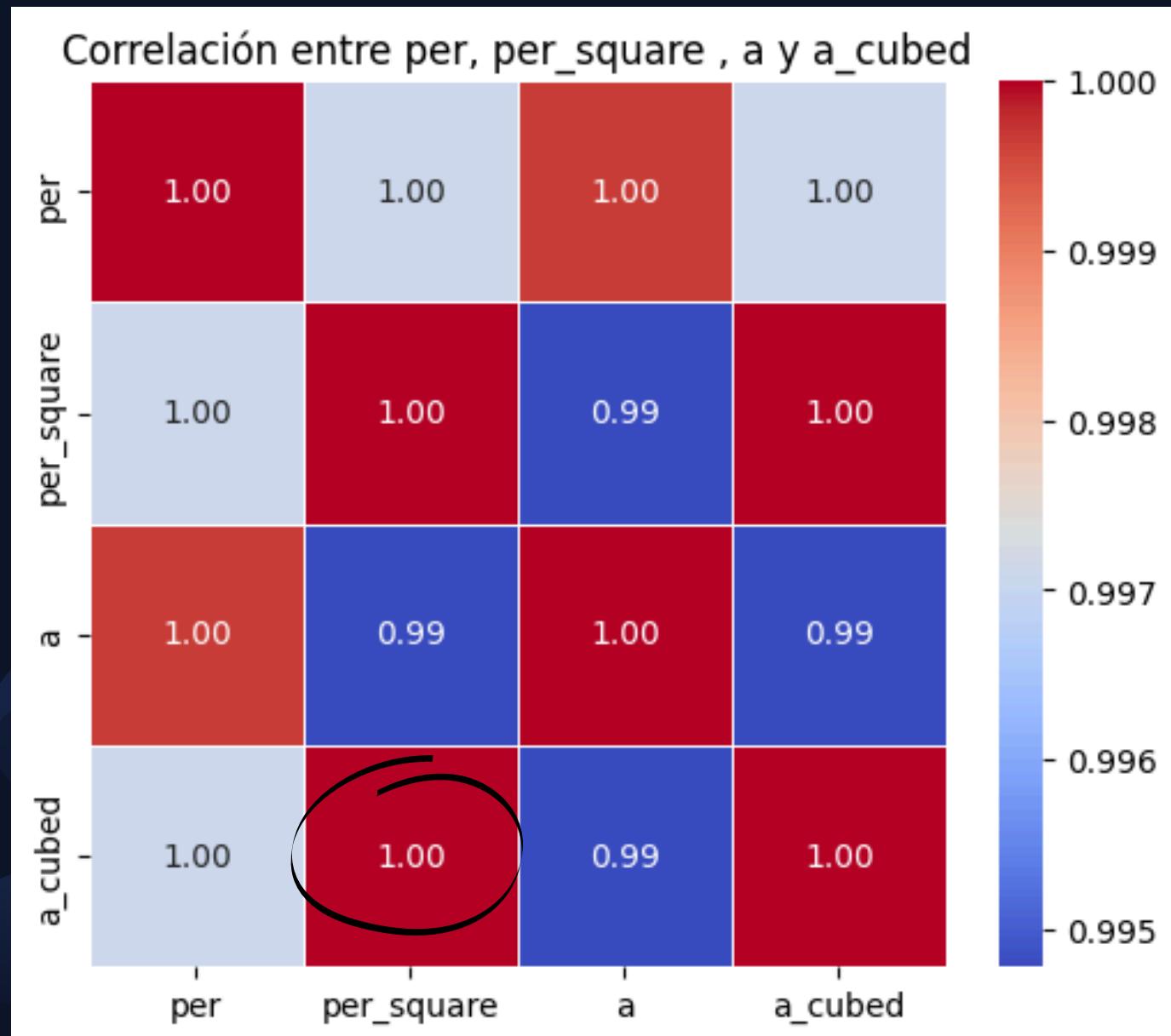
¿ Existirá una relación causal que explica la alta correlación ?

TERCERA LEY DE KEPLER

$$T^2 = A^3$$



Comprobamos esta relación en nuestros datos:

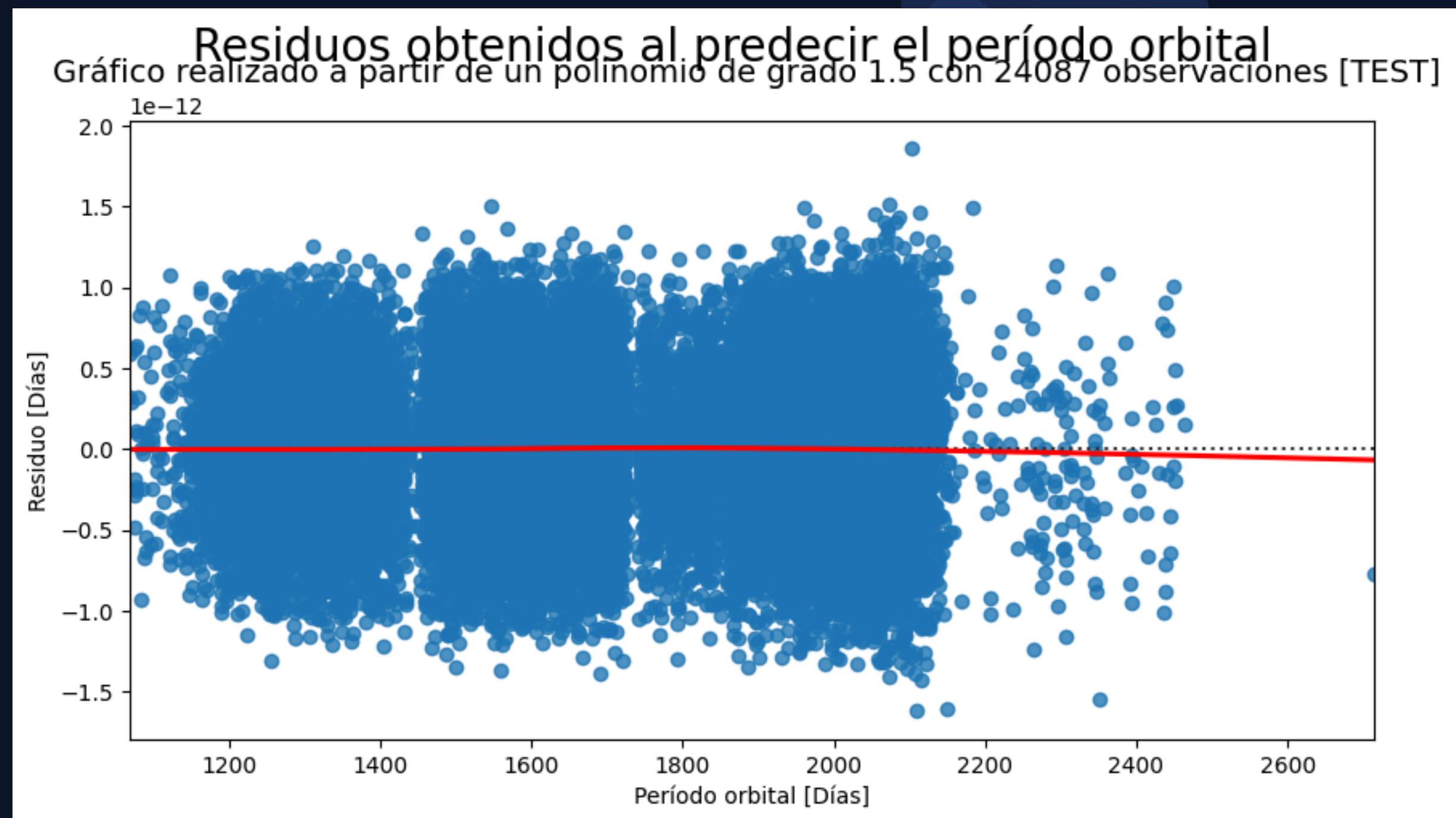


Vemos como per_square y a_cubed tienen un 100% de correlación



MODELADO Y EVALUACIÓN

Residuos del modelo con coeficiente polinómico de grado 1.5



Los residuos carecen de sesgo



CONCLUSIONES



En resumen.

✓ Hemos logrado construir un modelo predictivo robusto para estimar el período orbital de los asteroides utilizando una regresión polinómica de grado 2.

Lasso y Kepler

✓ La creación de características polinomiales y el uso de Lasso nos permitió identificar la característica más importante y a raíz de eso poder conocer la ley que rige la naturaleza de los periodos orbitales.

Gracias Por su atención!

¿PREGUNTAS?

