Proyecto 1 - Laboratorio aprendizaje estadístico

Predicción de salarios de jugadores de béisbol usando modelos de regresión.

Julia Hernández
Ana Sofia Hinojosa
Sara Hernández

DATASET

Origen de los datos

Grandes Ligas de Béisbol (USA), temporadas 1986–1987.

Contenido

- 322 jugadores.
- 20 variables.
- Estadísticas acumuladas de carrera, información categorica (league, division, newleague) y salario (objetivo)

Muestras

- Cada fila = jugador.
- Cada columna = estadística de desempeño, acumulado de carrera o variable categórica (League, Division, NewLeague).

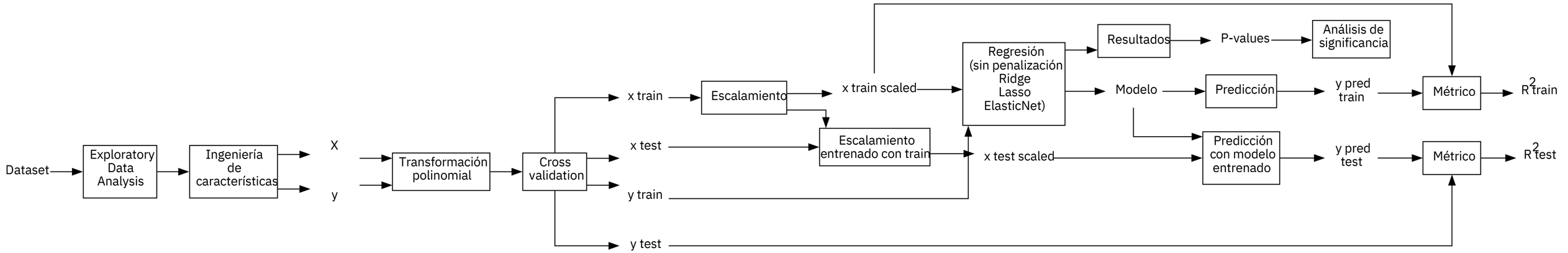
Objetivos
Generales

Predecir el salario de los jugadores de béisbol de las Grandes Ligas usando sus estadísticas de desempeño y determinar qué factores influyen más en la remuneración, aplicando técnicas de regresión lineal, polinomial y penalizada para mejorar la precisión y la interpretación del modelo.

Objetivos Específicos

- Analizar el dataset de los jugadores: cantidad de datos, qué variables hay, y qué tipo de variables son (numéricas o categóricas).
- Preparar los datos para el modelado: Limpieza del dataset, quitar duplicados y NaNs, creación de dummies para las variables de League, Division y NewLeague.
- Construir modelos de regresión: Regresión lineal simple, polinomial de grado 2 y de grado 3. Aplicar penalizaciones de Ridge, Lasso y ElasticNet.
- Evaluar el desempeño de los modelos: Comparar los R^2 de cada modelo y las variables con p-values significativos.
- Conclusiones o recomendaciones: Determinar si este modelo se ajusta correctamente a los datos, si se puede generalizar y si hace predicciones correctas.

Pipeline


Dataset
Exploratory Data Analysis
Ingeniería de características
X
y
Transformación polinomial
Cross validation
x train
x test
y train
y test
Escalamiento
x train scaled
Escalamiento entrenado con train
x test scaled
Regresión (sin penalización Ridge Lasso ElasticNet)
Resultados
P-values
Análisis de significancia
Modelo
Predicción
y pred train
Métrico
R2 train
Predicción con modelo entrenado
y pred test
Métrico
R2 test

Análisis de R2

R^2 de los modelos utilizados

Modelo	R2_train	R2_test
Lineal	0.60326	0.380623
Lineal con penalización Ridge	0.594157	0.403631
Lineal con penalización Lasso	0.5971	0.404124
Lineal con penalización ElasticNet	0.529303	0.353081
Grado 2	0.861026	0.407619
Grado 2 con penalización Ridge	0.799778	0.475286
Grado 2 con penalización Lasso	0.825276	0.470247
Grado 2 con penalización ElasticNet	0.631365	0.428992
Grado 3	0.968363	-1.95472
Grado 3 con penalización Ridge	0.875765	0.412473
Grado 3 con penalización Lasso	0.860321	0.417089
Grado 3 con penalización ElasticNet	0.72348	0.411073

Análisis de significancia

Factores de modelo sin penalización lineal

	Std.Err.	t	P>|t|
const	21.184727	25.293612	1.734893e-58
Hits	109.671043	2.709519	7.454154e-03
Walks	41.686921	1.964468	5.116571e-02
CAtBat	386.421413	-2.359500	1.947659e-02
PutOuts	24.105469	3.643922	3.599557e-04
Division_E	11.094461	1.991211	4.811834e-02
Division_W	11.094461	-1.991211	4.811834e-02

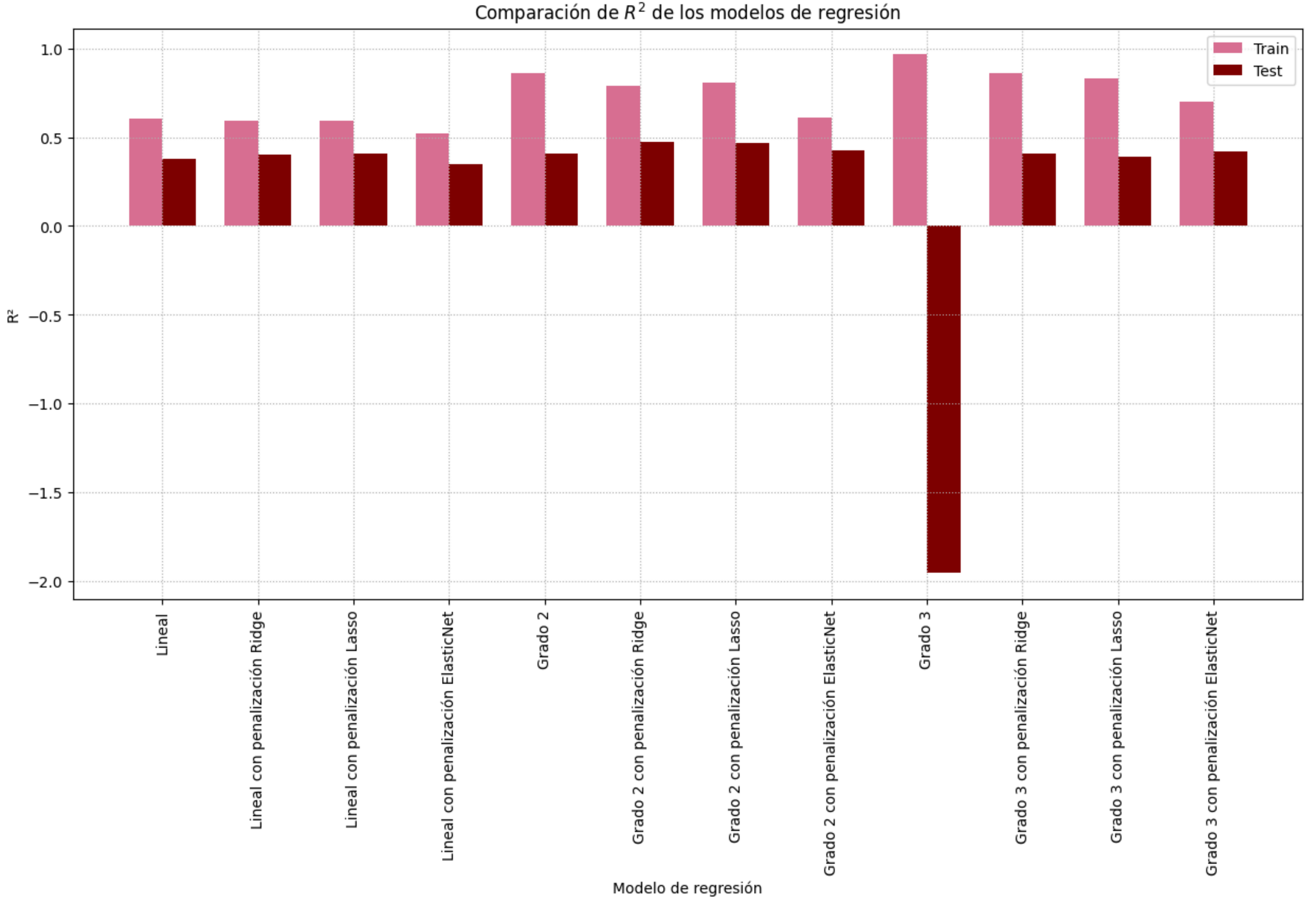
Factores de modelo sin penalización grado 2

	Std.Err.	t	P>|t|
const	13.718228	39.060311	3.896241e-76
HmRun	58.372915	-2.647887	9.049732e-03
CAtBat	315.093959	-5.400962	2.847152e-07
PutOuts	17.786483	4.454480	1.731491e-05
League_A	14.718682	-2.121073	3.571676e-02
League_N	14.718682	2.121073	3.571676e-02
CHits	348.454835	4.257769	3.809349e-05
Runs RBI	293.501866	2.692509	7.976805e-03
Runs CHits	588.493696	2.486511	1.410193e-02
Runs CHmRun	510.513215	2.640559	9.237886e-03
Runs CRBI	1038.504925	-2.342454	2.059760e-02
Runs Errors	88.771920	2.389759	1.822065e-02
RBI CHits	815.467008	-2.582522	1.085724e-02
RBI CHmRun	595.447035	-2.697813	7.857243e-03
RBI CRBI	1339.426009	2.631404	9.477923e-03
RBI Errors	94.037066	-3.792701	2.227164e-04
CHits NewLeague_A	224.543654	3.158808	1.949427e-03
CHits NewLeague_N	227.148252	4.185462	5.059210e-05

Análisis de significancia

Factores de modelo sin penalización grado 3

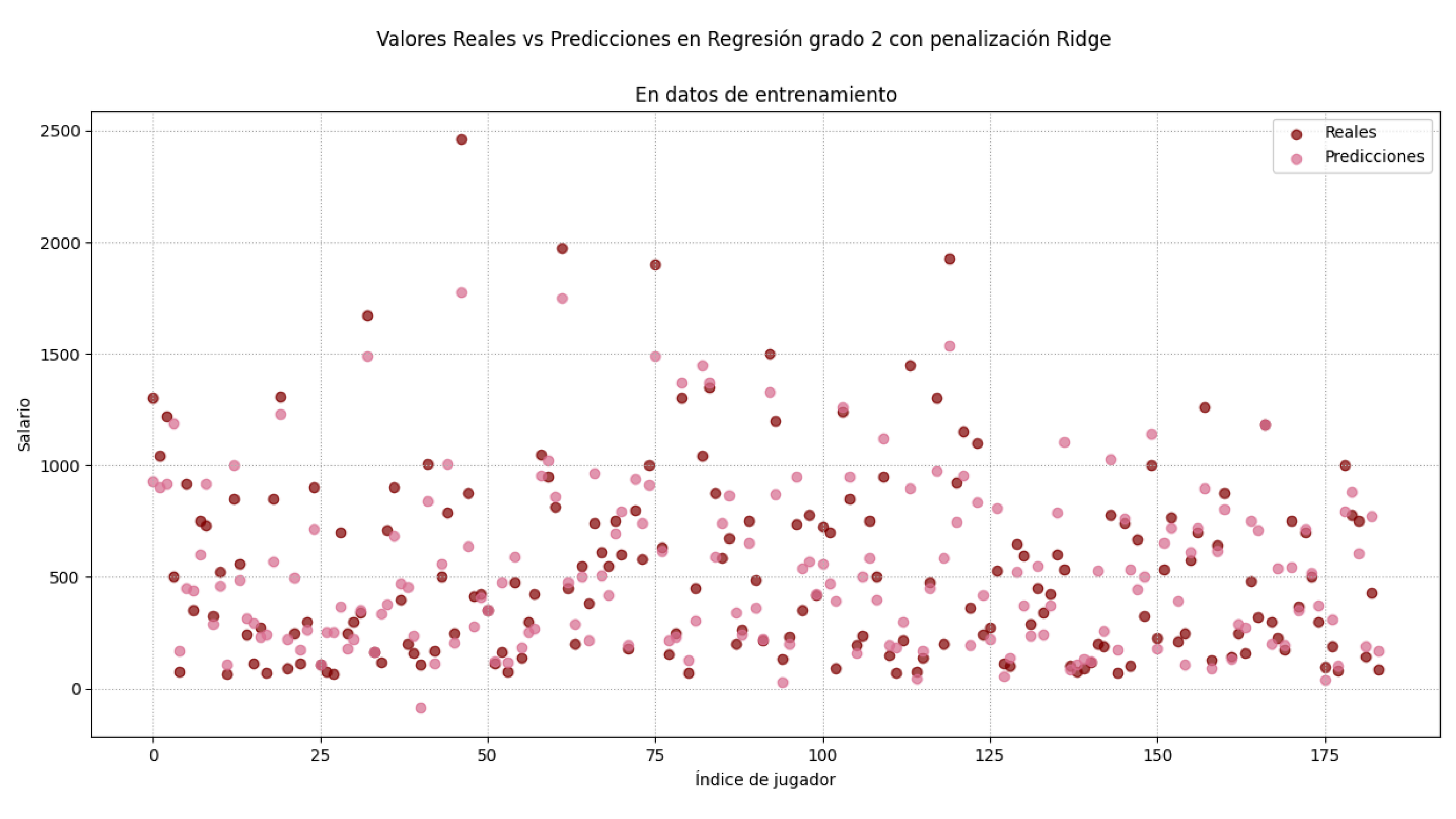
	Std.Err.	t	P>|t|
const	9.890472	54.177215	1.167337e-52
CAtBat	597.147665	-3.603041	6.399260e-04
PutOuts	21.467394	2.069339	4.282918e-02
League_A	16.992281	-2.422815	1.843607e-02
League_N	16.992283	2.422814	1.843612e-02
CHits	879.220517	1.947867	5.611377e-02
Errors	88.736854	2.012127	4.870264e-02
Runs NewLeague_A	104.510590	-2.026575	4.715741e-02
Errors NewLeague_N	54.764276	2.276575	2.639290e-02
Runs NewLeague_A^2	104.510591	-2.026575	4.715742e-02
Errors^2 NewLeague_N	137.757127	-2.175923	3.350984e-02
Errors NewLeague_N^2	54.764276	2.276575	2.639289e-02

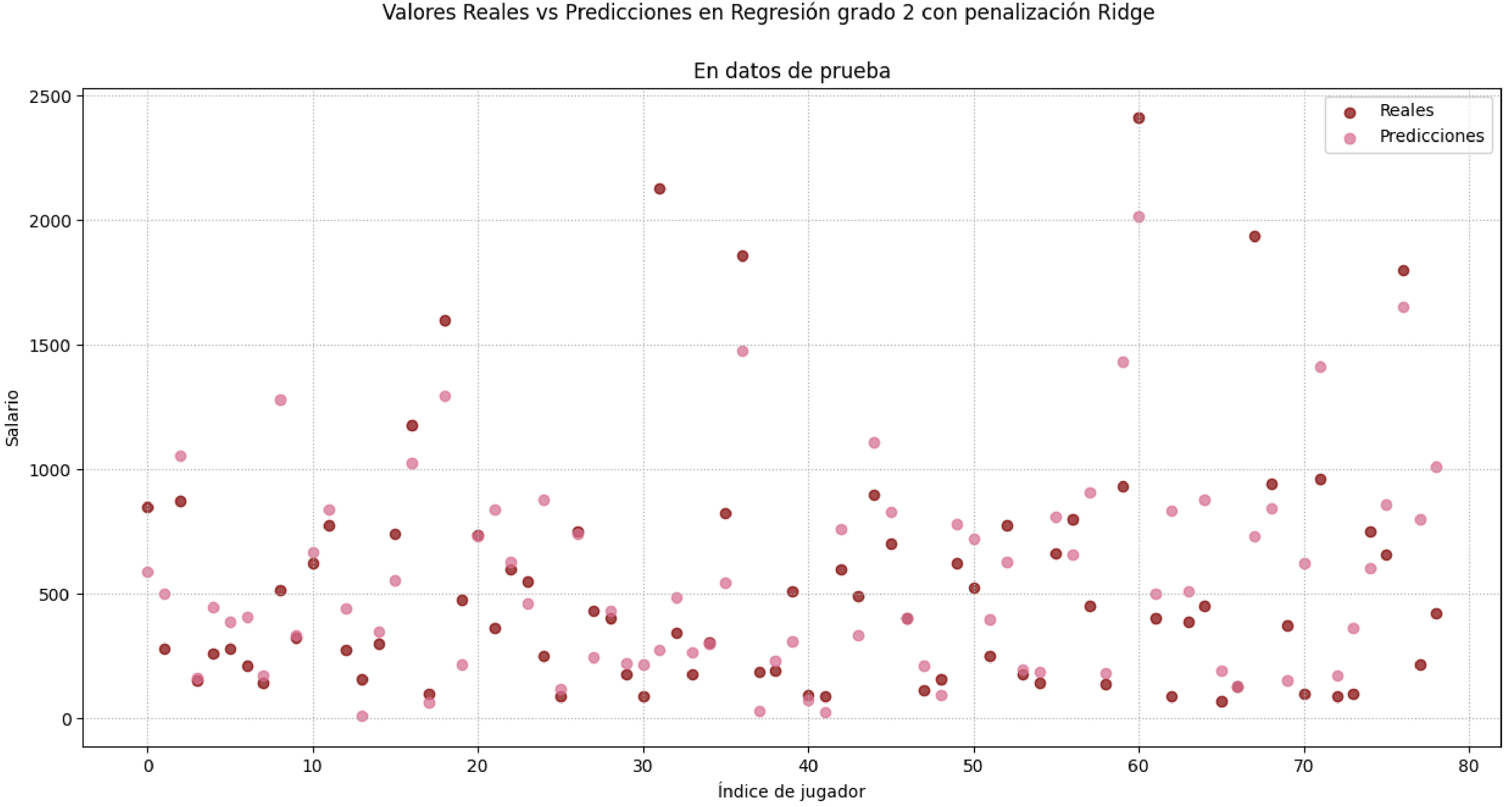

Comparación de R^2 de los modelos de regresión
Train
Test
R²
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Lineal
Lineal con penalización Ridge
Lineal con penalización Lasso
Lineal con penalización ElasticNet
Grado 2
Grado 2 con penalización Ridge
Grado 2 con penalización Lasso
Grado 2 con penalización ElasticNet
Grado 3
Grado 3 con penalización Ridge
Grado 3 con penalización Lasso
Grado 3 con penalización ElasticNet
Modelo de regresión

Mejor modelo

Regresión polinomial cuadrada con penalización Ridge

R2 Train	R2 Test
0.799778	0.475286

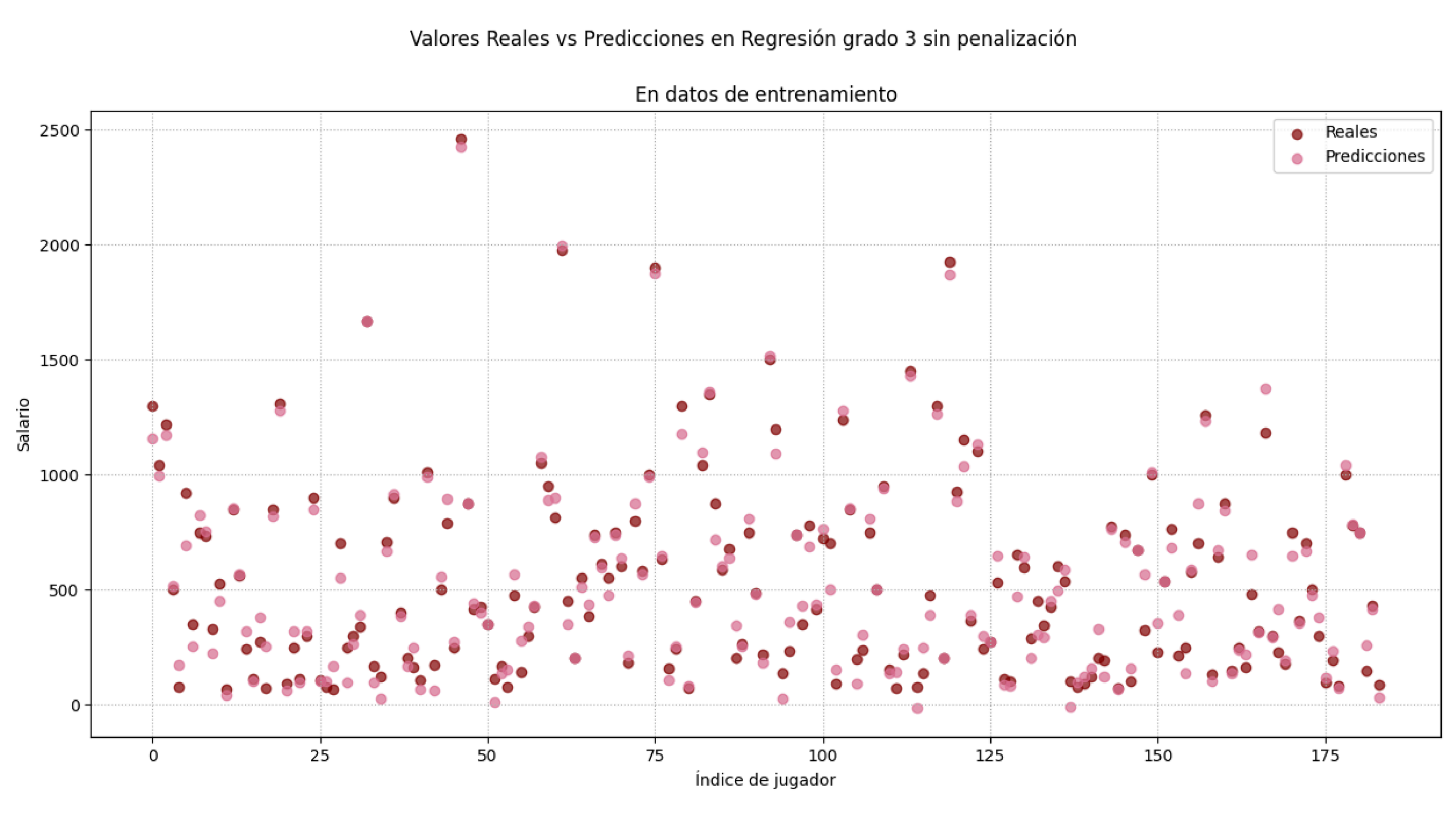

Valores Reales vs Predicciones en Regresión grado 2 con penalización Ridge
En datos de entrenamiento
Reales
Predicciones
Salario
Índice de jugador
2500
2000
1500
1000
500
0
0
25
50
75
100
125
150
175

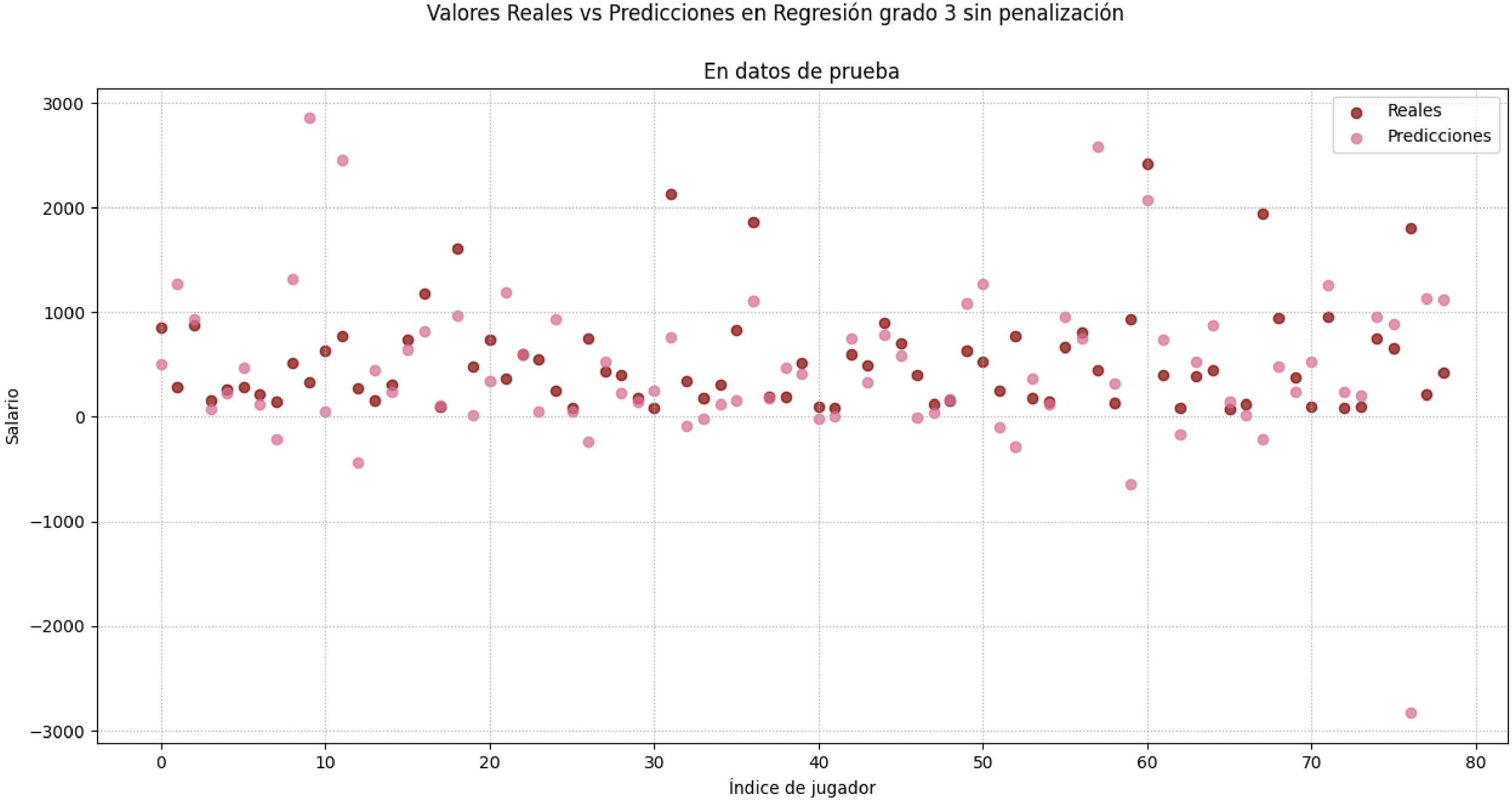

Valores Reales vs Predicciones en Regresión grado 2 con penalización Ridge
En datos de prueba
Reales
Predicciones
Salario
Índice de jugador
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
70
80

Ejemplo overfitting

Regresión polinomial cúbica sin penalización

R2 Train	R2 Test
0.968363	-1.95472


Valores Reales vs Predicciones en Regresión grado 3 sin penalización
En datos de entrenamiento
Reales
Predicciones
Salario
Índice de jugador
2500
2000
1500
1000
500
0
0
25
50
75
100
125
150
175


Valores Reales vs Predicciones en Regresión grado 3 sin penalización
En datos de prueba
Reales
Predicciones
Salario
Índice de jugador
3000
2000
1000
0
−1000
−2000
−3000
0
10
20
30
40
50
60
70
80

Conclusiones

Variables significativas que más influyen en el salario:

Regresión lineal sin penalización:

- Hits
- CAtBat
- Walks

Regresión polinomial grado 2 sin penalización:

- RBI CRBI
- RBI CHits
- Runs CHits

Regresión polinomial grado 3 sin penalización:

- CHits
- CAtBat

Conclusiones finales:

- El modelo creado no es confiable para predecir los salarios. Todos los modelos mostraron tener un overfitting significante y los R^2 de los datos de test no sobrepasan el 50% en ninguno de los casos.

- A su vez, creemos que muy pocas variables analizadas realmente son significativas para predecir el salario de los jugadores de béisbol. Podrían existir otros factores externos que influyan en esto (la inversión en los diferentes equipos, apoyo gubernamental, país originario del equipo).