Análisis y Reporte de Desempeño del Modelo

Regresión Lineal (scikit-learn) sobre Wine Quality

Fernanda Díaz Gutiérrez A01639572

15 de septiembre de 2025

1. Introducción

En el presente reporte se analiza el desempeño de un modelo de regresión lineal, específicamente aplicado al dataset Wine Quality, utilizando el framework scikit-learn. El objetivo principal del dataset es predecir la variable "quality"(rango 3–8) usando las variables físico/químicas de los vinos. Esta entrega es parte del segundo avance del proyecto, donde usé herramientas de un framework de machine learning para comparar el modelo lineal simple y la versión usando ridge (l2). En el reporte, mencionaré los hallazgos relacionados con los conceptos de sesgo, varianza y ajuste del modelo, interpretados en base a la teoría vista en clase.

2. Metodología

- Se utilizó el dataset winequality-red.csv.
- Se aplicó una división Train/Validation/Test:
  - 60 % train.
  - 20 % validation.
  - 20 % test.
- Se implementaron dos pipelines en scikit-learn:
  1. Regresión lineal con estandarización.
  2. Ridge regression (regularización l2, alpha=1).
- Se evaluaron las métricas: R2, MSE, RMSE y MAE.
- Se generaron curvas de aprendizaje para diagnosticar sesgo, varianza y nivel de ajuste.

3. Resultados

3.1. Métricas Train/Validation/Test

Los resultados principales se encuentran en la Tabla 1.

Modelo	R2 Train	R2 Validation	R2 Test
Linear	0.361	0.284	0.408
Ridge (alpha = 1)	0.361	0.284	0.408

Cuadro 1: Resultados en Train/Validation/Test para Linear y Ridge.

3.2. Comparación de desempeño en Test

Ambos modelos alcanzaron un R2 aproximado de = 0.41 en el conjunto de prueba, lo que nos dice que 'explican' alrededor del 41 % de la variabilidad de la variable quality.

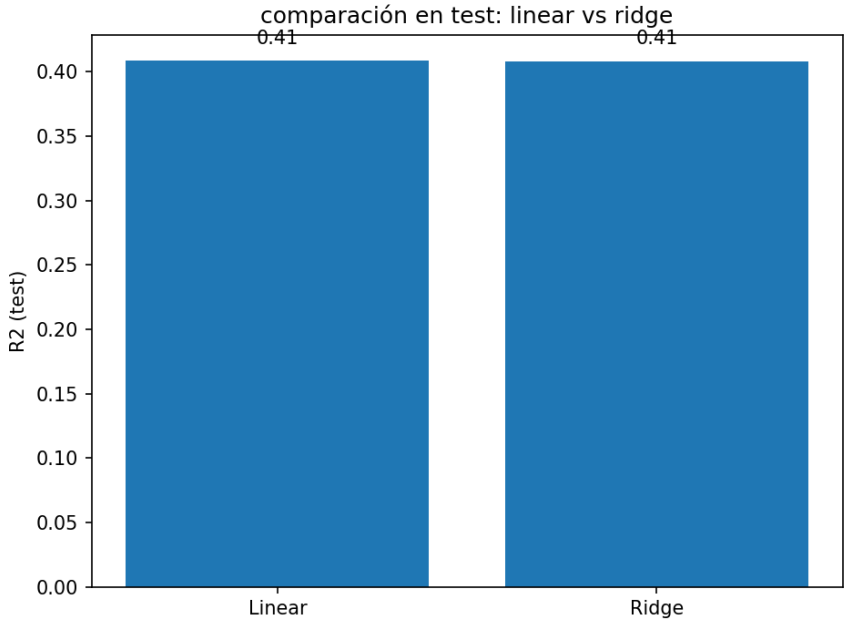

Figura 1: Comparación en test: Linear vs Ridge.

3.3. Curvas de aprendizaje

Las curvas de aprendizaje nos permiten observar/diagnosticar el comportamiento del modelo conforme aumenta el tamaño de los datos de training.

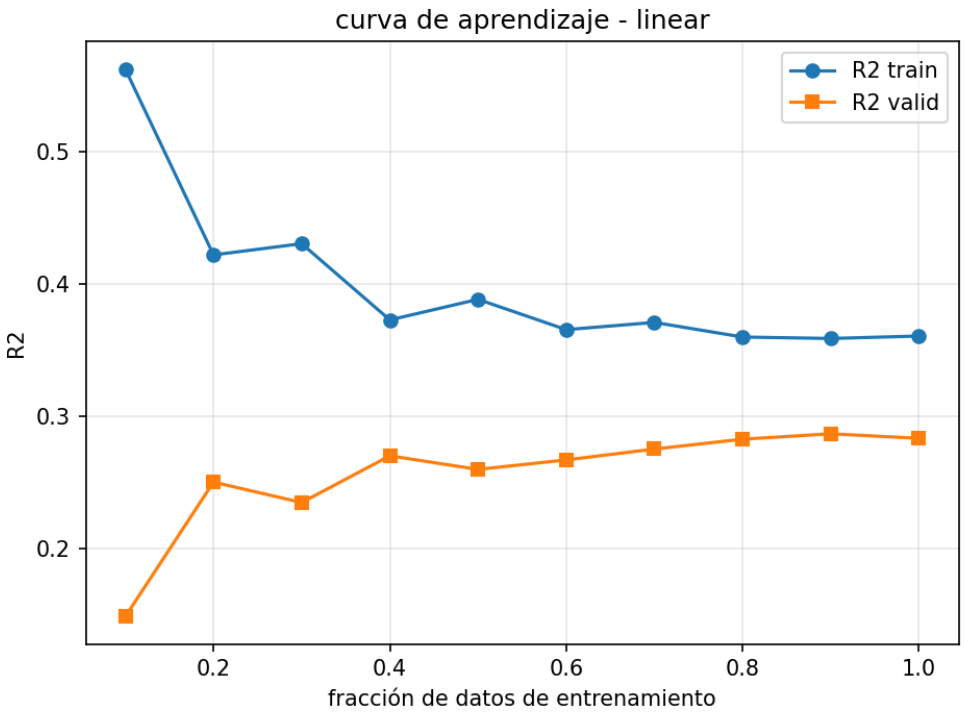

Figura 2: Curva de aprendizaje – Linear.

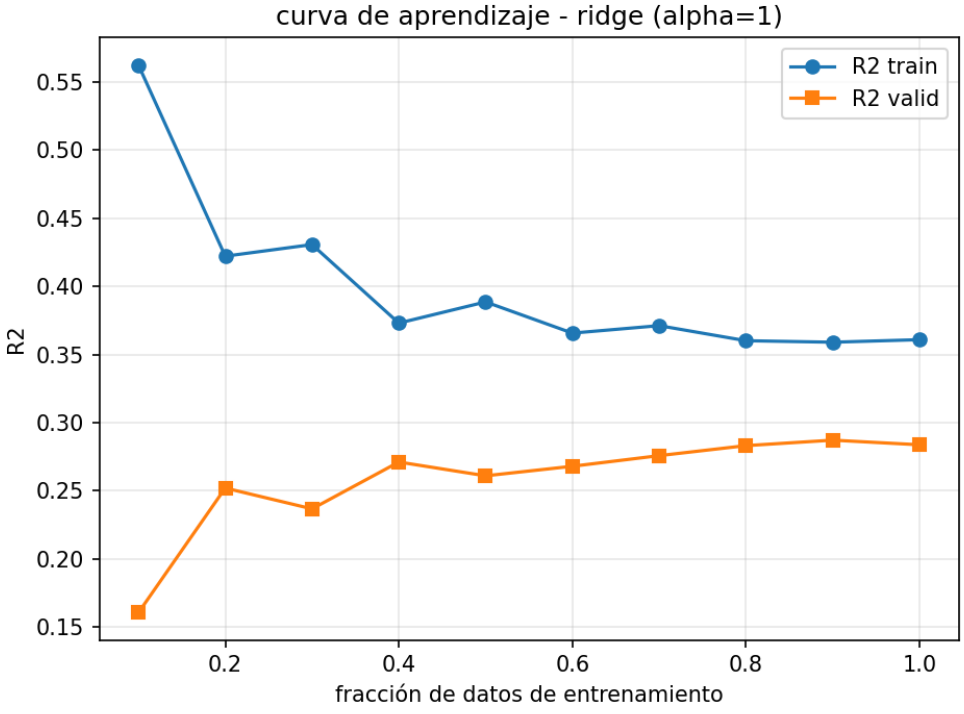

Figura 3: Curva de aprendizaje – Ridge (alpha = 1).

4. Diagnóstico Teórico-Práctico

De acuerdo con la teoría de sesgo y varianza aprendida en clase, los hallazgos de mi modelo son los siguientes:

4.1. Sesgo (bias)

El R2 en entrenamiento es bajo (aprox = 0.36). Esto significa que el modelo no logra capturar completamente la relación entre las variables predictoras/la variable target. **Diagnóstico: sesgo medio-alto.**

4.2. Varianza

La diferencia entre R2 train y R2 validation es moderada (brecha de aprox = 0.08). Esto nos dice que el modelo no se está sobreajustando de manera grave. **Diagnóstico: varianza media.**

4.3. Nivel de ajuste

- No se observa underfitting fuerte: R2 en validation/test es positivo.
- Tampoco hay overfitting fuerte: la brecha entre train y valid es controlada.

Diagnóstico general: ligero underfitting debido al sesgo alto.

5. Regularización y mejoras

Se aplicó regularización ridge l2 con alpha = 1. Los resultados en validación y prueba fueron esencialmente iguales a los de la regresión lineal simple (R2 aprox = 0.41).

Esto nos confirma que, al menos para este dataset:

- La regularización no genera una mejora significativa o relevante, ya que el modelo no presenta una varianza alta que necesite ser corregida.
- El principal problema entonces, es el sesgo alto, lo que sugiere que el modelo lineal es demasiado limitado para la complejidad del fenómeno.

De acuerdo con la teoría, el siguiente paso sería explorar modelos más flexibles (como una regresión polinómica o métodos no lineales como árboles), que nos permitan reducir el sesgo manteniendo bajo control la varianza.

6. Conclusiones

- Ambos modelos (linear y ridge) alcanzaron un desempeño muy similar, con R2 aprox = 0.41.
- El modelo presenta sesgo medio-alto y varianza media, lo que podemos interpretar como un poco de underfitting.
- La regularización de ridge l2 no mejoró significativamente el desempeño, lo que nos confirma que el principal problema es el sesgo.
- Según la teoría vista en clase, idealmente buscaríamos aumentar la complejidad del modelo para que sus predicciones sean mejores.