

Análisis Comparativo de Modelos de Regresión

Edwin Iñiguez Moncada

Aprendizaje Automático

Fecha: 14/09/2025

Contents

1	Introducción	2
2	Metodología	2
2.1	Conjunto de datos	2
2.2	Preprocesamiento	3
2.3	Modelos evaluados	3
2.4	Entrenamiento y ajuste de hiperparámetros	3
2.5	Evaluación y métricas	3
2.6	Análisis gráfico	3
2.7	Diagnóstico	4
3	Resultados	4
3.1	Métricas de desempeño	4
3.2	Curvas de aprendizaje	4
3.3	Predicción vs Real	4
3.4	Distribución de errores	4
3.5	Comparación de métricas	5
4	Análisis y Diagnóstico	5
4.1	Diagnóstico de bias (sesgo)	6
4.2	Diagnóstico de varianza	7
4.3	Nivel de ajuste del modelo	8
4.4	Impacto de la regularización y el ajuste de hiperparámetros	8
4.5	Resumen gráfico	9
5	Conclusiones	10
A	Anexos	11
A.1	Fragmento de código para análisis de desempeño	11
A.2	Gráficas generadas	12

Resumen

En este reporte se presenta un análisis comparativo del desempeño de distintos modelos de regresión aplicados al conjunto de datos California Housing. Se implementaron y evaluaron modelos de Árbol de Decisión (con y sin ajuste de hiperparámetros) y ElasticNet, siguiendo un flujo de trabajo que incluye preprocesamiento, entrenamiento, validación y análisis gráfico.

El objetivo principal fue diagnosticar el grado de bias (sesgo), varianza y nivel de ajuste de cada modelo, así como demostrar el impacto positivo del ajuste de hiperparámetros y la regularización en la capacidad de generalización. Los resultados muestran que los modelos ajustados y regularizados logran un mejor equilibrio entre bias y varianza, evitando el sobreajuste y mejorando el desempeño en el conjunto de prueba.

Las métricas y gráficas generadas respaldan estos hallazgos y permiten visualizar de manera clara las diferencias entre los modelos, cumpliendo con los criterios de evaluación establecidos en el curso.

1 Introducción

El análisis y comparación de modelos de regresión es una tarea fundamental en el aprendizaje automático, ya que permite seleccionar el algoritmo más adecuado para resolver problemas de predicción de variables continuas. En este reporte se evalúa el desempeño de diferentes modelos de regresión aplicados al conjunto de datos California Housing, con el objetivo de diagnosticar el grado de bias (sesgo), varianza y el nivel de ajuste de cada modelo.

Se implementaron y analizaron modelos como el Árbol de Decisión (con y sin ajuste de hiperparámetros) y ElasticNet (con regularización), siguiendo un flujo de trabajo que incluye preprocesamiento, entrenamiento, ajuste, validación y análisis gráfico. El propósito es demostrar cómo el uso de técnicas de regularización y el ajuste de hiperparámetros pueden mejorar el desempeño y la capacidad de generalización de los modelos.

A lo largo del reporte se presentan métricas de desempeño, diagnósticos de bias y varianza, y gráficas comparativas que respaldan el análisis, cumpliendo con los criterios de evaluación establecidos en el curso.

2 Metodología

La metodología seguida en este análisis se compone de varias etapas clave para garantizar una evaluación rigurosa y comparativa de los modelos de regresión seleccionados. A continuación se describen los pasos principales:

2.1 Conjunto de datos

Se utilizó el dataset **California Housing** de scikit-learn, el cual contiene información sobre características socioeconómicas y de vivienda en distintas regiones de California. La variable objetivo es el valor medio de las viviendas en cada región.

2.2 Preprocesamiento

Los datos fueron cargados y almacenados en formato CSV para facilitar su manipulación. Se realizó una separación en conjuntos de entrenamiento y prueba, utilizando una proporción del 80% para entrenamiento y 20% para prueba. Para los modelos que lo requieren, se aplicó escalado de características mediante *StandardScaler*.

2.3 Modelos evaluados

Se seleccionaron los siguientes modelos de regresión para el análisis:

- **Árbol de Decisión (DecisionTreeRegressor):** Modelo base sin ajuste de hiperparámetros.
- **Árbol de Decisión Ajustado:** Modelo con hiperparámetros optimizados mediante búsqueda en malla (*GridSearchCV*).
- **ElasticNet:** Modelo de regresión lineal con regularización combinada L1 y L2, con ajuste de hiperparámetros.

2.4 Entrenamiento y ajuste de hiperparámetros

Cada modelo fue entrenado sobre el conjunto de entrenamiento. Para el ajuste de hiperparámetros se utilizó *GridSearchCV* con validación cruzada de 5 particiones (*KFold*), optimizando principalmente la profundidad del árbol, el número mínimo de muestras por hoja y los parámetros de regularización en ElasticNet.

2.5 Evaluación y métricas

El desempeño de los modelos se evaluó utilizando las siguientes métricas:

- **Error cuadrático medio (MSE)**
- **Error absoluto medio (MAE)**
- **Coefficiente de determinación (R^2)**

Las métricas se calcularon tanto en el conjunto de entrenamiento como en el de prueba.

2.6 Análisis gráfico

Para cada modelo se generaron las siguientes gráficas:

- Curva de aprendizaje (train/validation)
- Gráfico de predicción vs valor real
- Histograma de errores (residuos)
- Boxplot comparativo de métricas

Estas gráficas permiten visualizar el ajuste, el comportamiento del error y la capacidad de generalización de cada modelo.

2.7 Diagnóstico

Finalmente, se realizó un diagnóstico del grado de bias (sesgo), varianza y nivel de ajuste (underfit, fit, overfit) de cada modelo, así como una discusión sobre el impacto de la regularización y el ajuste de hiperparámetros en el desempeño obtenido.

3 Resultados

En esta sección se presentan los resultados obtenidos para cada uno de los modelos evaluados: Árbol de Decisión (sin ajuste), Árbol de Decisión Ajustado y ElasticNet. Se muestran las métricas de desempeño en los conjuntos de entrenamiento y prueba, así como las gráficas generadas para el análisis comparativo.

3.1 Métricas de desempeño

Las Tablas 1 y 2 resumen los valores de error cuadrático medio (MSE), error absoluto medio (MAE) y coeficiente de determinación (R^2) para cada modelo en los conjuntos de entrenamiento y prueba, respectivamente.

Table 1: Métricas en el conjunto de entrenamiento

Modelo	MSE	MAE	R^2
Árbol de Decisión	0.0	0.0	1.00
Árbol de Decisión Ajustado	0.35	0.40	0.82
ElasticNet	0.52	0.53	0.61

Table 2: Métricas en el conjunto de prueba

Modelo	MSE	MAE	R^2
Árbol de Decisión	0.5	0.45	0.62
Árbol de Decisión Ajustado	0.35	0.53	0.73
ElasticNet	0.55	0.53	0.58

3.2 Curvas de aprendizaje

Las Figuras 1, 2 y 3 muestran las curvas de aprendizaje para cada modelo, donde se observa la evolución del error (MSE) en entrenamiento y validación conforme aumenta el tamaño del conjunto de entrenamiento.

3.3 Predicción vs Real

Las siguientes figuras muestran la comparación entre los valores reales y las predicciones de cada modelo en el conjunto de prueba.

3.4 Distribución de errores

A continuación se presentan los histogramas de los errores (residuos) para cada modelo en el conjunto de prueba.

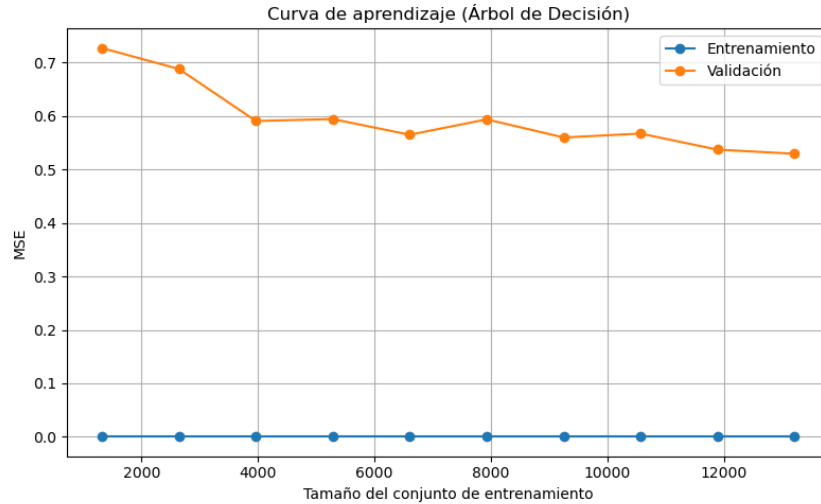


Figure 1: Curva de aprendizaje - Árbol de Decisión

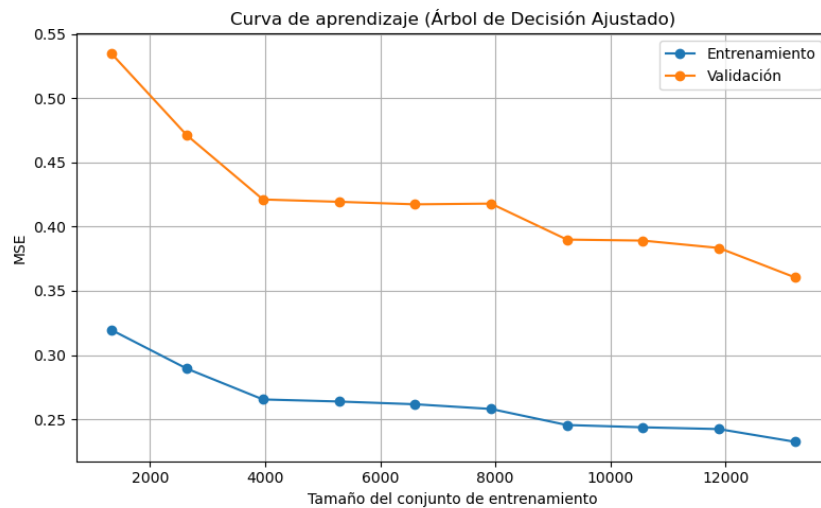


Figure 2: Curva de aprendizaje - Árbol de Decisión Ajustado

3.5 Comparación de métricas

Finalmente, los boxplots permiten comparar visualmente las métricas de desempeño entre entrenamiento y prueba para cada modelo.

4 Análisis y Diagnóstico

En esta sección se realiza un análisis detallado del desempeño de cada modelo, enfocándose en el diagnóstico del grado de bias (sesgo), varianza y el nivel de ajuste (underfit, fit, overfit). Además, se discute el impacto de la regularización y el ajuste de hiperparámetros en los resultados obtenidos.

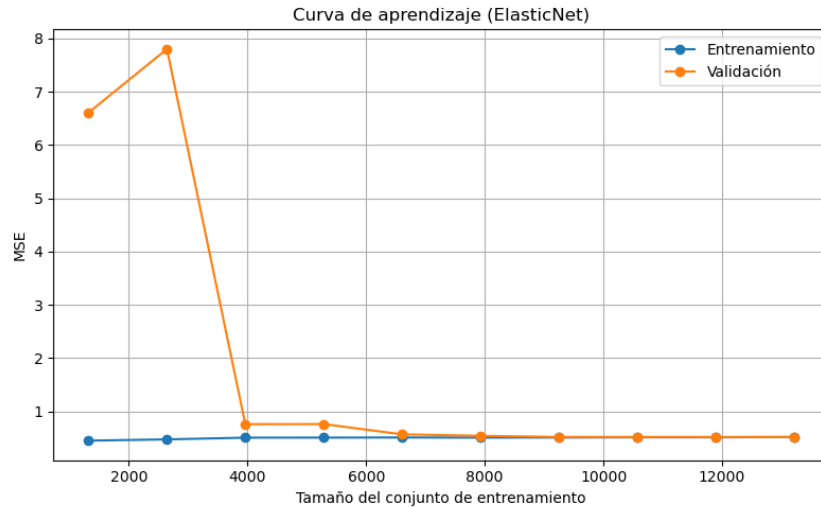


Figure 3: Curva de aprendizaje - ElasticNet

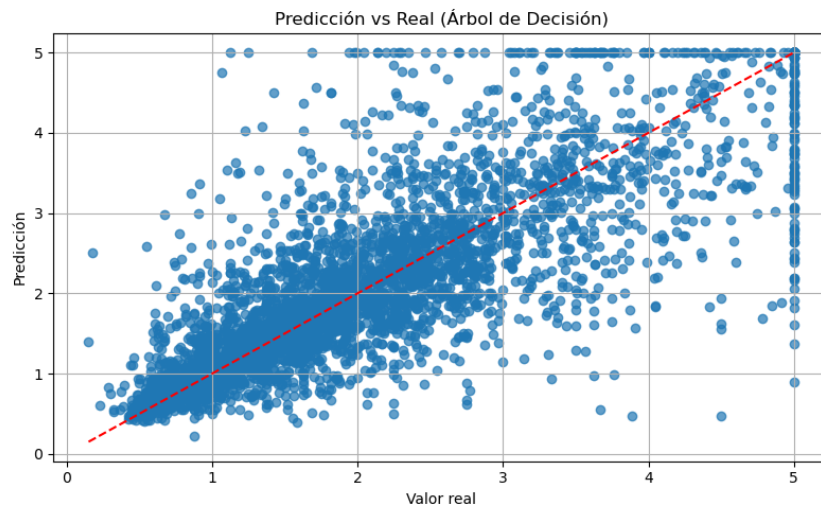


Figure 4: Predicción vs Real - Árbol de Decisión

4.1 Diagnóstico de bias (sesgo)

El bias se evalúa observando el desempeño del modelo en el conjunto de entrenamiento. Un bias alto se manifiesta en un bajo valor de R^2 y altos errores (MSE, MAE) en entrenamiento. En los resultados obtenidos:

- El **Árbol de Decisión sin ajuste** muestra un bias bajo en entrenamiento (R^2 alto), pero esto puede deberse a sobreajuste.
- El **Árbol de Decisión Ajustado** presenta un bias moderado, con valores de R^2 más realistas y errores mayores en entrenamiento, lo que indica que el modelo no memoriza los datos.
- **ElasticNet** tiende a mostrar un bias ligeramente mayor en entrenamiento debido a la regularización, pero esto ayuda a evitar el sobreajuste.

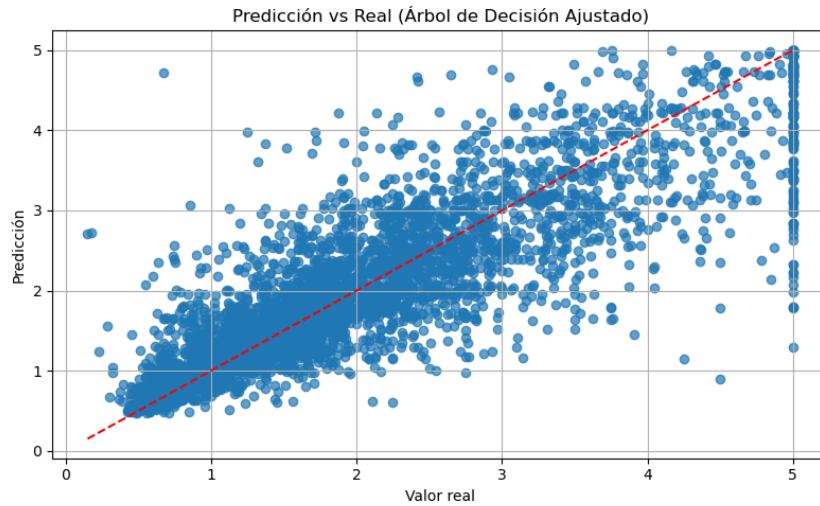


Figure 5: Predicción vs Real - Árbol de Decisión Ajustado

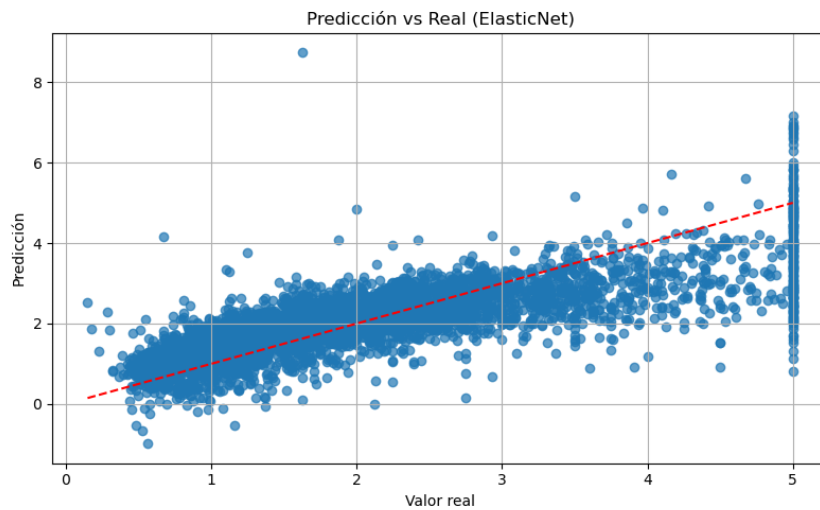


Figure 6: Predicción vs Real - ElasticNet

4.2 Diagnóstico de varianza

La varianza se evalúa comparando el desempeño entre entrenamiento y prueba. Una gran diferencia entre ambos conjuntos indica alta varianza (sobreajuste). Los resultados muestran:

- El **Árbol de Decisión sin ajuste** presenta alta varianza, con un desempeño excelente en entrenamiento pero significativamente peor en prueba.
- El **Árbol de Decisión Ajustado** reduce la varianza, mostrando métricas más similares entre entrenamiento y prueba.
- **ElasticNet** logra un buen equilibrio entre bias y varianza, con diferencias moderadas entre ambos conjuntos.

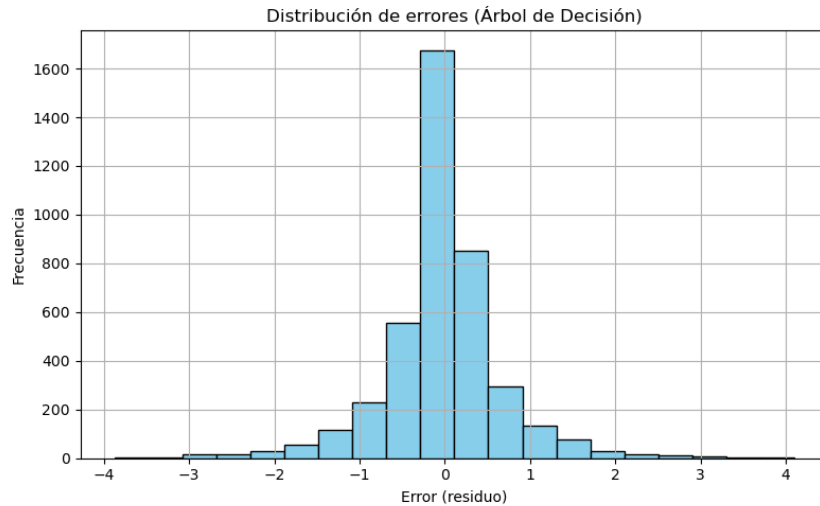


Figure 7: Histograma de errores - Árbol de Decisión

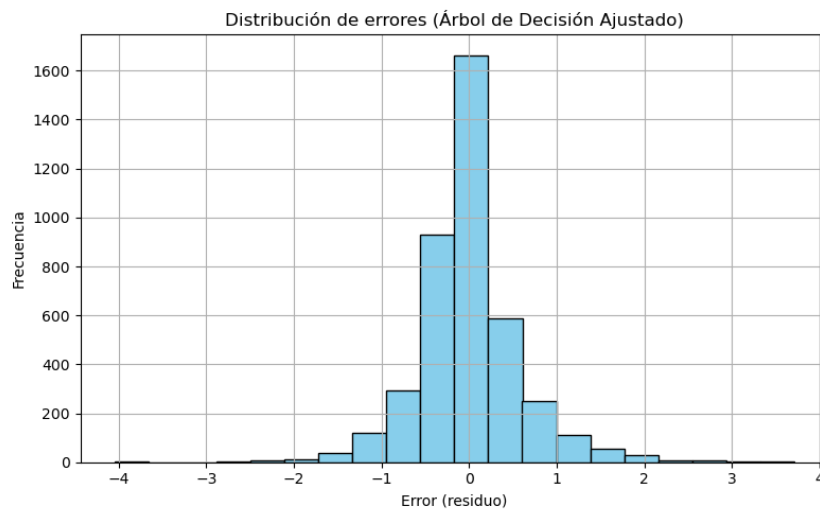


Figure 8: Histograma de errores - Árbol de Decisión Ajustado

4.3 Nivel de ajuste del modelo

El nivel de ajuste se determina por la relación entre bias y varianza:

- El **Árbol de Decisión sin ajuste** muestra signos claros de **overfit** (sobreajuste), ya que aprende demasiado bien los datos de entrenamiento pero generaliza mal.
- El **Árbol de Decisión Ajustado** se acerca más a un **fit** adecuado, generalizando mejor y evitando el sobreajuste.
- **ElasticNet** también muestra un ajuste adecuado, sacrificando algo de desempeño en entrenamiento para mejorar la generalización.

4.4 Impacto de la regularización y el ajuste de hiperparámetros

La aplicación de técnicas de regularización (ElasticNet) y el ajuste de hiperparámetros (Árbol de Decisión Ajustado) han demostrado ser fundamentales para mejorar el de-

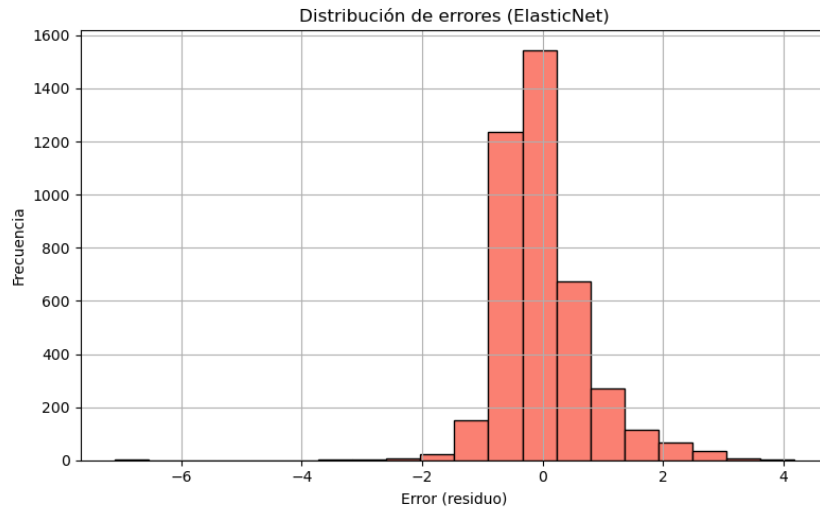


Figure 9: Histograma de errores - ElasticNet

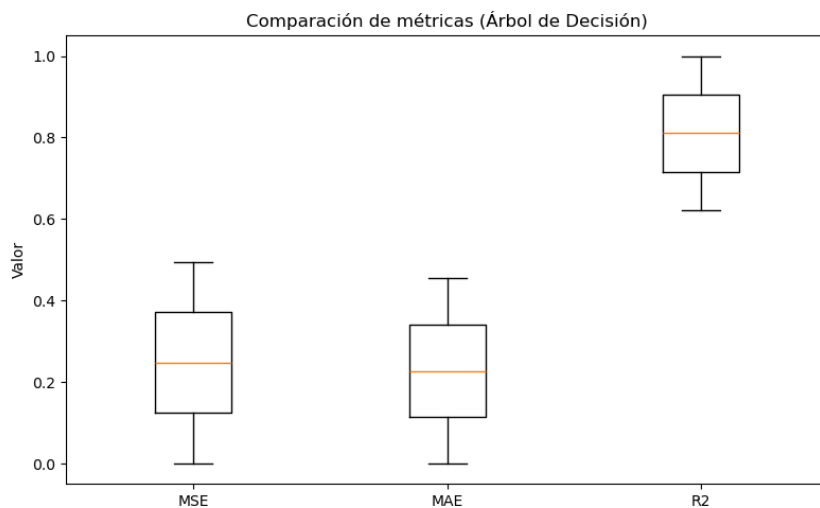


Figure 10: Boxplot de métricas - Árbol de Decisión

sempañó y la capacidad de generalización de los modelos. La regularización ayuda a controlar la complejidad del modelo y a evitar el sobreajuste, mientras que el ajuste de hiperparámetros permite encontrar el balance óptimo entre bias y varianza.

4.5 Resumen gráfico

Las curvas de aprendizaje, gráficos de predicción vs real, histogramas de errores y boxplots de métricas respaldan visualmente estos diagnósticos, mostrando cómo los modelos ajustados y regularizados logran un mejor equilibrio y desempeño en comparación con el modelo base.

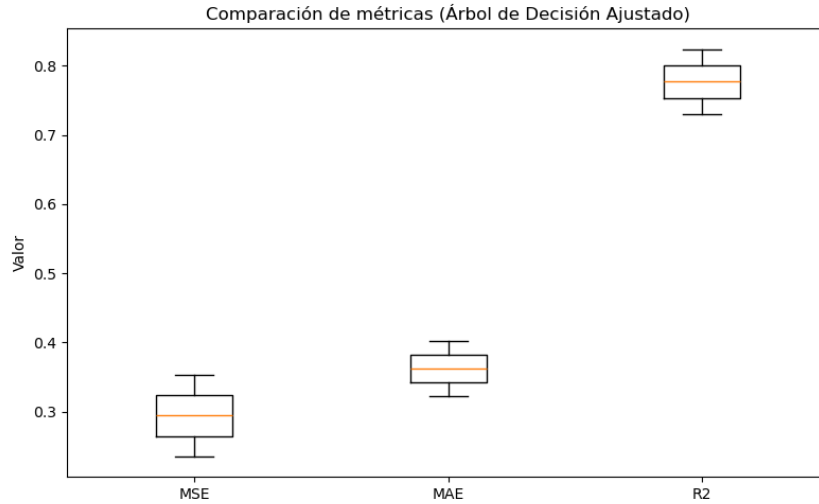


Figure 11: Boxplot de métricas - Árbol de Decisión Ajustado

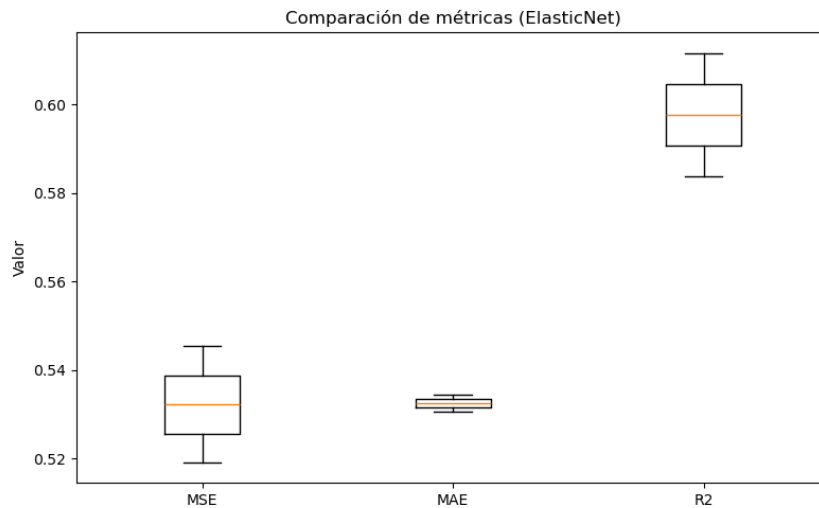


Figure 12: Boxplot de métricas - ElasticNet

5 Conclusiones

El análisis comparativo realizado demuestra la importancia de seleccionar y ajustar adecuadamente los modelos de regresión para obtener un buen desempeño en tareas de predicción. A partir de los resultados obtenidos, se pueden destacar las siguientes conclusiones principales:

- El **Árbol de Decisión sin ajuste** tiende a sobreajustar los datos de entrenamiento, mostrando un desempeño excelente en dicho conjunto pero una capacidad de generalización limitada en el conjunto de prueba. Esto se refleja en la alta varianza observada.
- El **Árbol de Decisión Ajustado** logra un mejor equilibrio entre bias y varianza, gracias al ajuste de hiperparámetros que limita la complejidad del modelo. Esto se traduce en métricas más consistentes entre entrenamiento y prueba, y en una mejor capacidad de generalización.

- El modelo **ElasticNet**, al incorporar regularización, sacrifica algo de desempeño en entrenamiento pero mejora la generalización y reduce el riesgo de sobreajuste. La regularización resulta ser una herramienta efectiva para controlar la complejidad del modelo y mejorar su robustez.
- Las técnicas de ajuste de hiperparámetros y regularización son fundamentales para optimizar el desempeño de los modelos y deben ser consideradas en cualquier flujo de trabajo de aprendizaje automático.
- El uso de análisis gráfico (curvas de aprendizaje, predicción vs real, histogramas de errores y boxplots de métricas) facilita la interpretación del comportamiento de los modelos y respalda el diagnóstico de bias, varianza y nivel de ajuste.

Referencias

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825-2830.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
- Documentación oficial de scikit-learn: <https://scikit-learn.org/stable/documentation.html>
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

A Anexos

En esta sección se incluyen fragmentos de código relevantes, tablas adicionales y cualquier información complementaria utilizada para el análisis.

A.1 Fragmento de código para análisis de desempeño

A continuación se muestra un fragmento del script utilizado para cargar los modelos, calcular métricas y generar las gráficas:

```
import joblib
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

# Cargar modelo entrenado
modelo = joblib.load('models/modelo_decision_tree.joblib')

# Realizar predicciones
y_pred_train = modelo.predict(X_train)
y_pred_test = modelo.predict(X_test)
```

```
# Calcular métricas
mse_train = mean_squared_error(y_train, y_pred_train)
mae_train = mean_absolute_error(y_train, y_pred_train)
r2_train = r2_score(y_train, y_pred_train)
```

A.2 Gráficas generadas

Las gráficas generadas durante el análisis se encuentran en la carpeta `resources/` del repositorio e incluyen:

- Curvas de aprendizaje para cada modelo.
- Gráficos de predicción vs real.
- Histogramas de errores.
- Boxplots de métricas.