Instituto Tecnológico y de Estudios Superiores de Monterrey
Inteligencia artificial avanzada para la ciencia de datos I

Análisis del desempeño del modelo Random Forest en la predicción de demanda de bicicletas compartidas

Profesor:
Benjamín Valdés Aguirre

Paul Park - A01709885

11 de septiembre de 2025

Introducción

Los sistemas de bicicletas compartidas facilitan la renta y devolución de bicicletas alrededor del mundo. Esta tecnología ha generado un gran interés por su importancia en áreas de tráfico, ambientalismo y salud. Además, dado que los datos de uso (como el tiempo de renta y las estaciones de origen y destino) están explícitamente registrados, estos sistemas ofrecen mayores oportunidades para investigaciones comparado con otros transportes, como autobuses o metro.

En este reporte se analiza el desempeño de un modelo de Random Forest Regressor implementado con el framework scikit-learn, comparando sus resultados en los conjuntos de entrenamiento, validación y prueba. El objetivo es evaluar sesgo, varianza y nivel de ajuste, así como aplicar técnicas de regularización para mejorar la generalización del modelo.

Descripción del dataset

Se utilizó el dataset Bike Sharing del UC Irvine Machine Learning Repository, que contiene registros diarios y horarios de 2011 a 2012 del sistema Capital Bikeshare en Washington D.C. Por cuestiones de simplicidad, se eligió el conjunto de registros diarios, dado que contiene 731 filas (a diferencia del de registros horarios con 17379 filas), las cuales representan el total de bicicletas rentadas por día. Las variables seleccionadas y su tratamiento son:

Variable	Tipo	Descripción	Uso
season	categórica	Estación del año (1: primavera, 2: verano, 3: otoño, 4: invierno)	Usada (one-hot encoding)
yr	categórica	Año del registro (0: 2011, 1: 2012)	Usada (one-hot)
mnth	categórica	Mes del registro (1-12)	Usada (one-hot)
weekday	categórica	Día de la semana (0-6)	Usada (one-hot)
weathersit	categórica	Tipo de clima (1: soleado, 2: nublado, 3: lluvia ligera, 4: tormenta/nieve)	Usada (one-hot)
temp	continua	Temperatura normalizada (°C)	Usada (normalizada)
atemp	continua	Sensación térmica normalizada (°C)	Usada (normalizada)
hum	continua	Humedad normalizada (%)	Usada (normalizada)
windspeed	continua	Velocidad del viento normalizada	Usada (normalizada)

casual	integer	Número de usuarios casuales	Descartada
registered	integer	Número de usuarios registrados	Descartada
instant	integer	Índice del registro	Descartada
dteday	date	Fecha del registro	Descartada
cnt	integer	Objetivo: Total de bicicletas rentadas	Usada

Se descartaron las variables irrelevantes para el análisis (casual, registered, instant, dteday).

Metodología

El flujo de trabajo consistió en:

Preprocesamiento:

- One-hot encoding con pandas.
- Normalización min–max en variables continuas.

Separación de datos:

- Train: 70%
- Validation: 15%
- Test: 15%

Modelo:

- Random Forest Regressor (n_estimators=100, max_depth=None, random_state=42).

Métricas de evaluación:

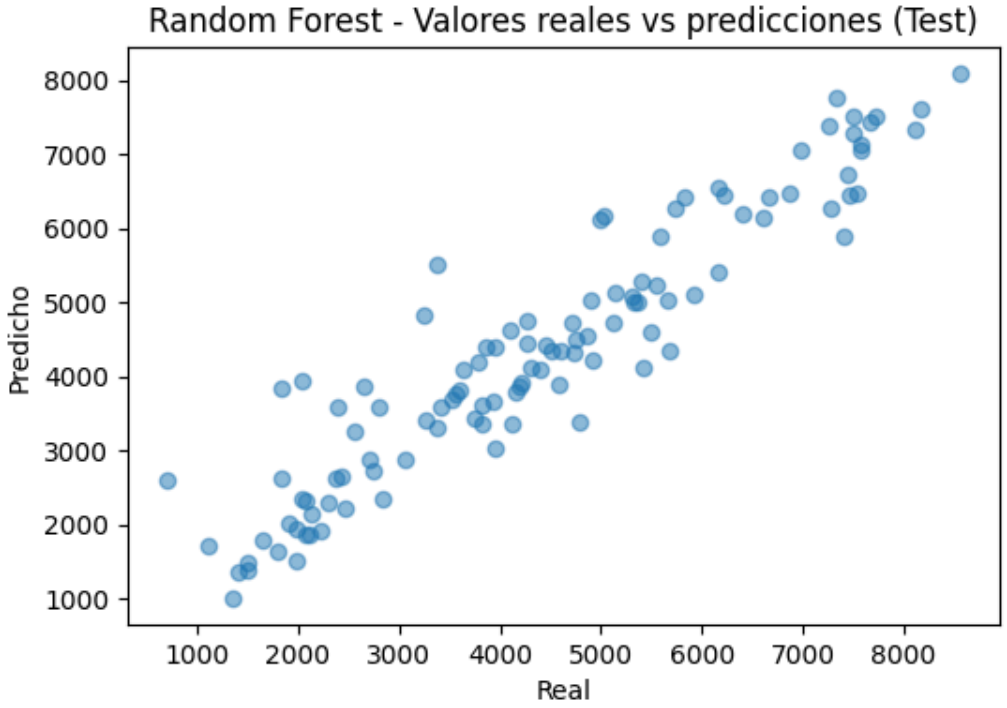
- Coeficiente de determinación (R^2).
- Error absoluto medio (MAE).
- Raíz del error cuadrático medio (RMSE).

Resultados iniciales

```
=== Evaluación del modelo ===
Train: R²=0.9781, MAE=198.33, RMSE=282.05
Validation: R²=0.8761, MAE=441.71, RMSE=717.54
Test: R²=0.8799, MAE=494.42, RMSE=675.61
```

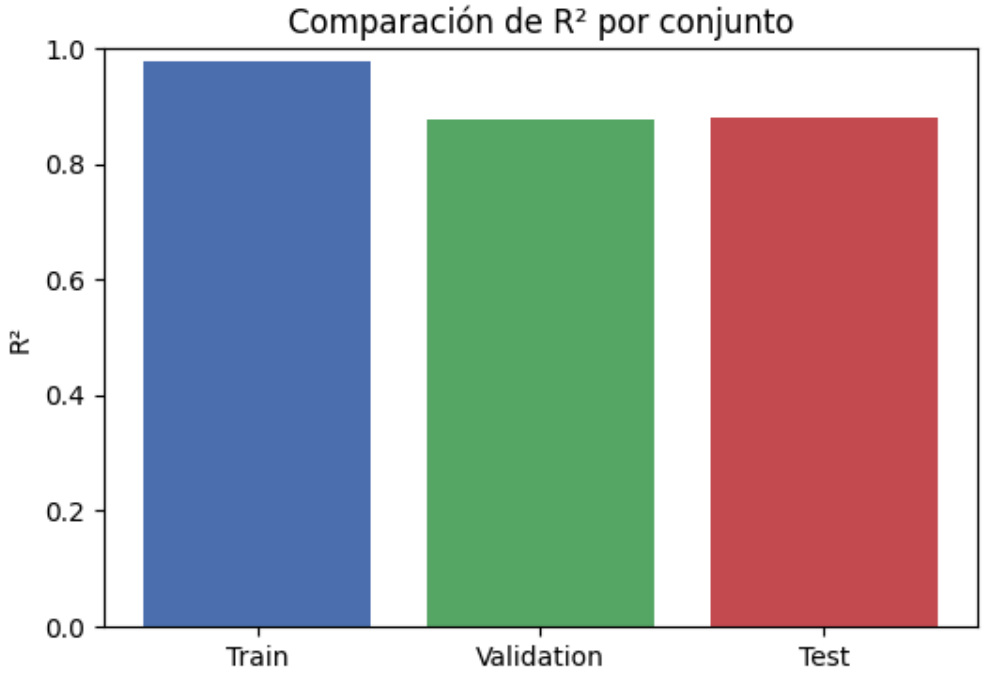
Gráficas

- **Dispersión real vs predicho (Test)**



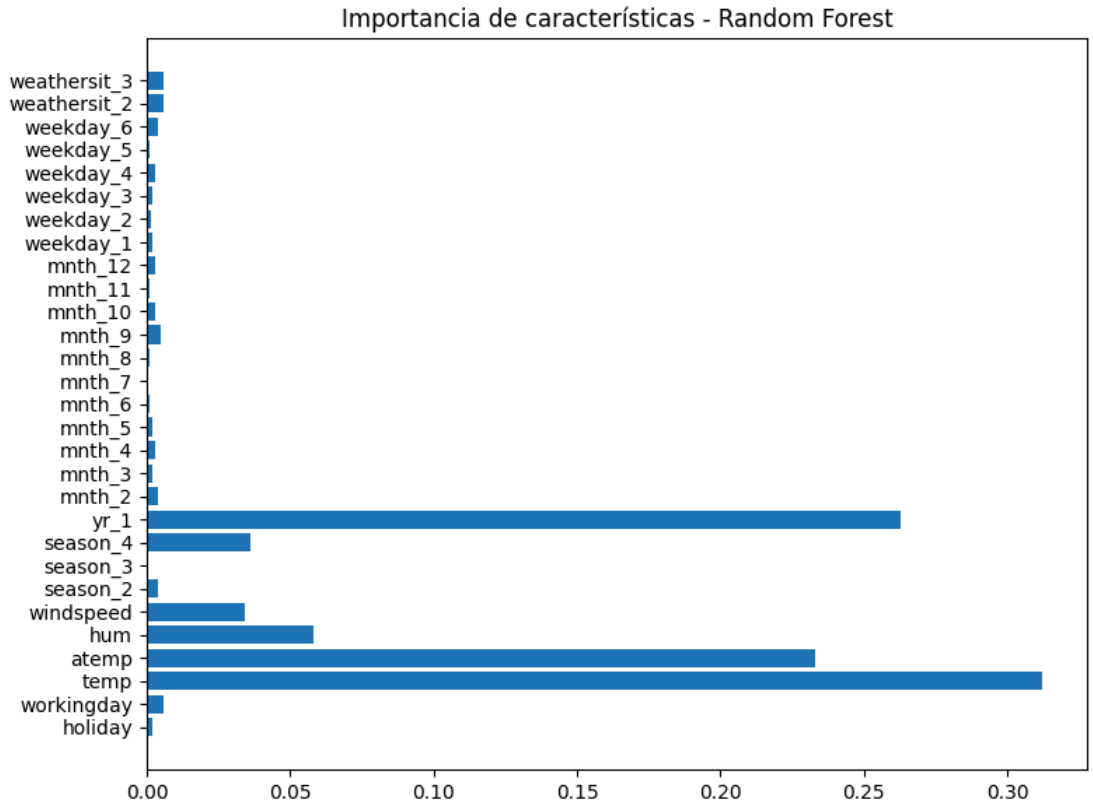
Se observa una buena correlación, aunque con ligera dispersión en valores altos.

- **Comparación de R^2 (barras)**



Se evidencia una diferencia notable entre Train y Validation/Test.

- **Importancia de características:**



Aquí se puede observar que las variables climáticas y estacionales tienen mayor peso.

Diagnóstico del modelo

- **Sesgo (bias): Bajo.** El modelo explica alrededor del 88% de la varianza en validación y prueba, lo que indica que aprende patrones relevantes.
- **Varianza: Media.** Existe diferencia entre Train (R^2=0.97) y Validation/Test (~0.88), lo que sugiere un ligero sobreajuste.
- **Nivel de ajuste:** El modelo se encuentra **cercano al fit ideal**, con una leve tendencia al *overfitting*.

Regularización y mejora

Para reducir la varianza y mejorar la generalización, se ajustaron los hiperparámetros del modelo:

- **max_depth=10** (limitando la profundidad de los árboles).
- **min_samples_split=5** (requiere más datos para crear divisiones).

Resultados con regularización:

```
=== Evaluación del modelo regularizado ===
Train (Reg): R²=0.9621, MAE=271.74, RMSE=371.42
Validation (Reg): R²=0.8717, MAE=458.46, RMSE=730.33
Test (Reg): R²=0.8775, MAE=508.36, RMSE=682.34
```

Comparación:

El R^2 en entrenamiento disminuyó de 0.98 a 0.96, lo cual indica menor sobreajuste. En validación y prueba, el desempeño se mantuvo prácticamente igual (~0.88). Aunque no hubo mejora directa en métricas de generalización, la reducción de la diferencia entre Train y Validation/Test muestra un mejor balance entre bias y varianza.

Conclusiones

El modelo **Random Forest** aplicado al dataset de bicicletas compartidas alcanzó un **desempeño sólido** con bajo sesgo y varianza moderada. La inclusión de un conjunto de validación permitió identificar una ligera tendencia al *overfitting*, corregida con técnicas de regularización.

El análisis muestra que:

- El modelo logra generalizar adecuadamente a datos no vistos.
- La importancia de las variables climáticas y temporales sugiere que la demanda de bicicletas está fuertemente relacionada con condiciones ambientales y estacionales.
- Con ajustes adicionales de hiperparámetros, el desempeño podría mejorar aún más.

Este ejercicio permitió no sólo aplicar un framework de Machine Learning, sino también comprender la importancia de separar los datos en conjuntos Train/Validation/Test y evaluar el balance entre sesgo y varianza.

Referencias

Scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/
UC Irvine Machine Learning Repository – Bike Sharing Dataset.
https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset