**Predicción horaria de energía renovable en una micro‑red**

**Proyecto Final Integrador — Diplomatura en Ciencia de Datos y Análisis Avanzado**

**Equipo:** Grupo F — Tomas López Turconi, Hernán Nuñez  
**Fecha:** 6 de octubre de 2025

**Resumen Ejecutivo**

Se desarrolló un modelo para **predecir la energía renovable total por hora** (total\_renewable\_energy) en una micro‑red, con el fin de **mejorar la planificación operativa** (despacho de baterías y compras a la red). Se comparó un **baseline naïve** con **Regresión Lineal** y **Random Forest**.

**Resultado principal:** **Regresión Lineal** obtuvo el **menor error** en test (**RMSE 41.19 kWh**, **MAE 33.90 kWh**, **R² 0.038**, **MASE 0.708**), superando al baseline (**RMSE 59.03 kWh**) con una **reducción del 30.21%**. Un **backtesting temporal (k=5)** confirmó la **robustez** del modelo con **RMSE 40.69 ± 1.08 kWh** (CV≈2.7%).  
**Impacto esperado:** reducción de compras en horas caras, mejor uso del almacenamiento y mayor estabilidad operativa.

**1. Definición del problema y relevancia**

* **Problema:** la generación renovable (solar/eólica) es volátil; pronosticar la potencia generada por hora permite optimizar baterías y compras a la red.
* **Objetivo medible (KPI):** reducir el **RMSE ≥ 15%** respecto del baseline naïve.
* **Criterios de éxito:** RMSE (principal), MAE, R² y MASE ((<1) indica mejora vs. naïve).
* **Valor de negocio:** con menor incertidumbre, se decide **cuándo cargar/descargar** baterías y **cuánta energía comprar** externamente.

**2. Datos y Metodología (CRISP‑DM)**

**2.1 Datos**

* **Fuente:** dataset público de micro‑red renovable (formato horario).
* **Tamaño:** ~3.5k filas horarias (~5 meses).
* **Variables principales:** timestamp, solar\_pv\_output, wind\_power\_output, total\_renewable\_energy (target), clima (solar\_irradiance, wind\_speed, temperature, humidity, atmospheric\_pressure) y battery\_state\_of\_charge.
* **Calidad:** sin nulos críticos en columnas clave; valores extremos esperables (irradiancia=0 de noche, picos de viento).

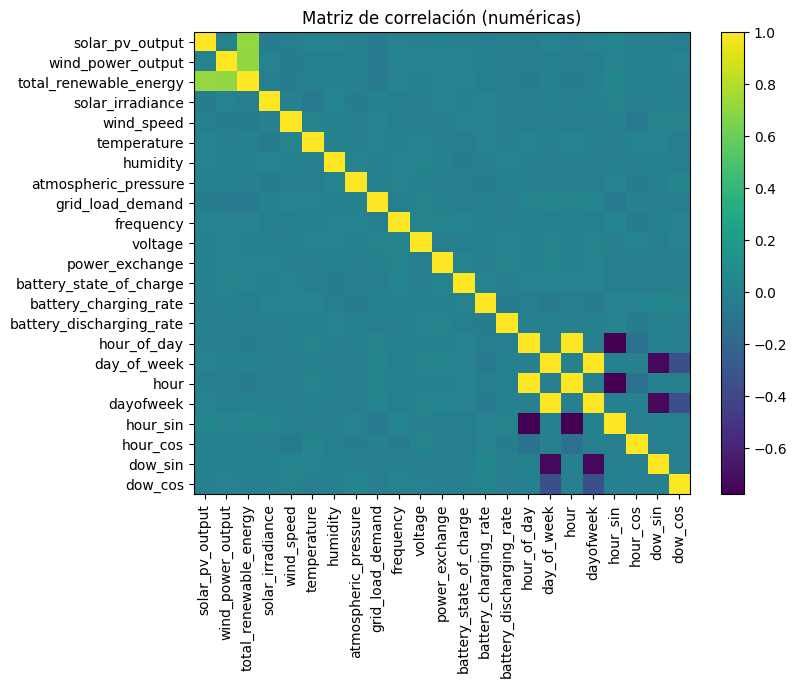
**2.2 Preparación y EDA (resumen)**

* Conversión de fechas y **orden temporal**; remoción de columnas predicted\_\* para evitar fuga.
* **Ingeniería temporal:** armónicos hour\_sin/cos, dow\_sin/cos.
* **Lags sin fuga:** total\_renewable\_energy\_lag1, lag24, roll\_mean\_24; y \*\_lag1 para exógenas (cuando existen).
* **Hallazgos clave:**
  + Correlaciones **contemporáneas** altas de solar\_pv\_output y wind\_power\_output con el target (componentes del total). **No** se usaron en t para evitar fuga; se utilizaron **rezagos**.
  + **Autocorrelación** del target mínima (lag1≈−0.01, lag24≈0); exógenas en t−1 con correlaciones muy bajas (≤0.05).
  + El **patrón horario/semanal** es leve; los armónicos aportan mejora vs. naïve.

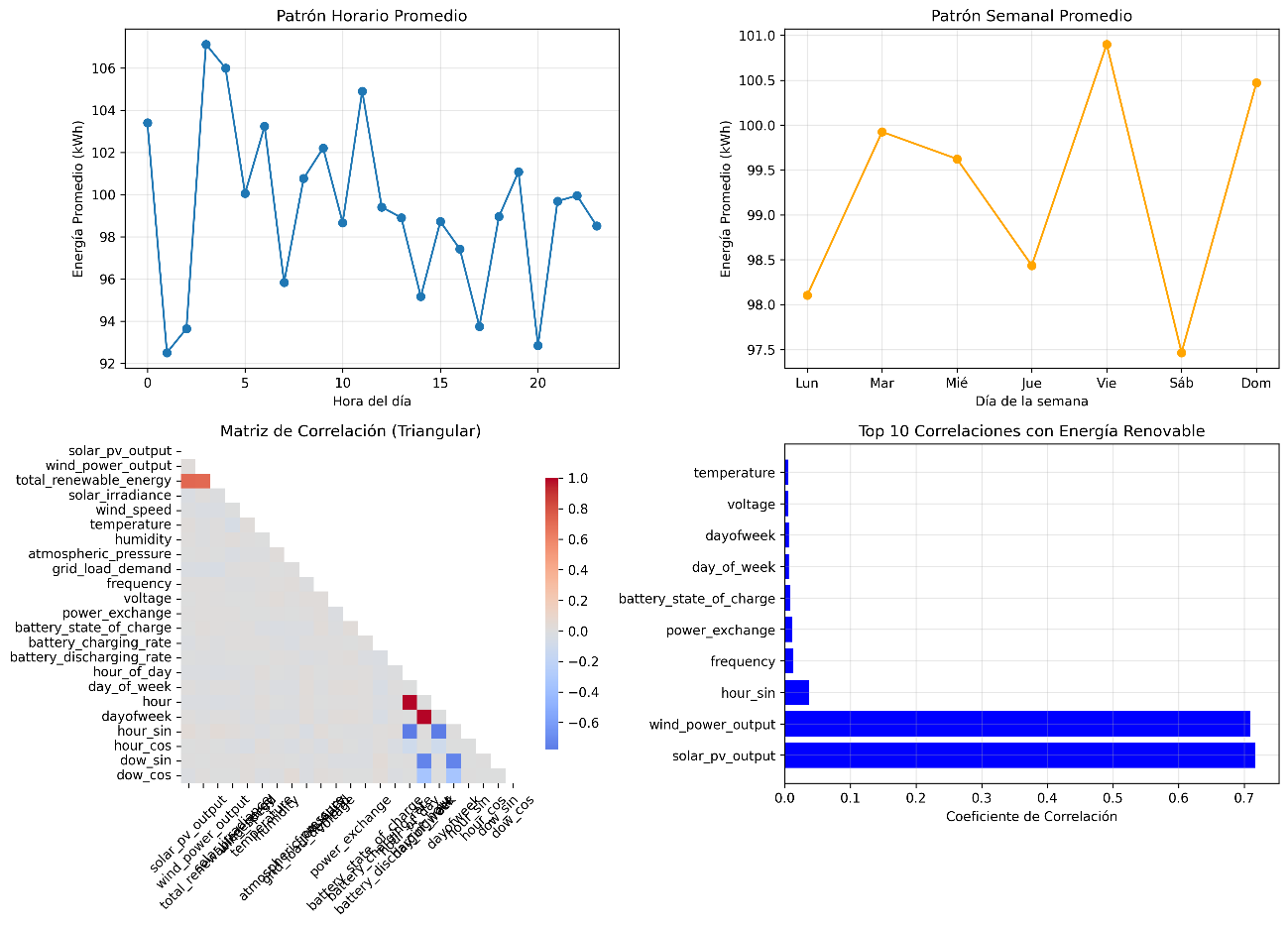
A blue sound wave with white text

AI-generated content may be incorrect.

*“Serie temporal de total\_renewable\_energy (hora a hora). Se observa alta variabilidad intra-día y leve tendencia estacional; motiva el uso de armónicos temporales y validación temporal.”*



*“Matriz de correlación entre variables numéricas. Las correlaciones altas de solar\_pv\_output/wind\_power\_output con el total son contemporáneas (componentes del target). Para evitar fuga de información, estas variables en t* ***no*** *se usan como features; se utilizan sus rezagos.”*



*“Patrón horario y semanal promedio (arriba); matriz triangular y top de correlaciones (abajo). La estacionalidad diaria es leve pero presente; se codifica con armónicos hour\_sin/cos y dow\_sin/cos. Las correlaciones fuertes con solar\_pv\_output/wind\_power\_output son contemporáneas y no se emplean en el modelado para evitar fuga.”*

**2.3 Esquema de validación**

* **Split 80/20 cronológico** (train: primeras observaciones; test: últimas).
* Métricas calculadas **exclusivamente** en test.
* **Backtesting (k=5)** adicional para robustez: **40.69 ± 1.08 kWh**.

**3. Modelado y evaluación**

**3.1 Modelos**

* **Baseline naïve:** (\hat{y}*t = y*{t-1}).
* **Regresión Lineal (LR):** con StandardScaler.
* **Random Forest (RF):** 500 árboles, random\_state=42.

**Features utilizadas:** lags y rolling del target (lag1, lag24, roll\_mean\_24), lags de exógenas (\*\_lag1 cuando existen) y armónicos hour\_sin/cos, dow\_sin/cos. No se incluyeron variables **en t** para evitar fuga.

**3.2 Resultados en test (holdout 20%)**

| **Modelo** | **RMSE (kWh)** | **MAE (kWh)** | **R²** | **MASE** |
| --- | --- | --- | --- | --- |
| Naive (t‑1) | 59.03 | 48.08 | −0.975 | 1.004 |
| LinearRegression | **41.19** | **33.90** | 0.038 | 0.708 |
| RandomForest | 41.95 | 34.59 | 0.002 | 0.722 |

**Selección:** **LinearRegression** por **RMSE mínimo** y **MASE<1** (mejora clara vs. naïve).  
**ΔRMSE vs. baseline:** **30.21%**.

**3.3 Backtesting (k=5)**

* **LinearRegression:** **40.69 ± 1.08 kWh** (CV≈2.7%).  
  Confirma **estabilidad** del desempeño y generalización temporal.

**3.4 Interpretabilidad**

* Importancias de RF (y XGB si se ejecuta) muestran predominio de **lags del target** y **armónicos**.
* Exógenas en t−1 aportan marginalmente, consistente con su baja correlación.
* Con autocorrelación casi nula, un modelo **parco** (lineal) resulta competitivo.

A graph with blue and white bars

AI-generated content may be incorrect.

*“Importancias de variables (RF). Destacan total\_renewable\_energy\_roll\_mean\_24, wind\_speed\_lag1, atmospheric\_pressure\_lag1, temperature\_lag1 y los armónicos temporales; consistente con el aporte moderado de exógenas con rezago y la utilidad de patrones temporales.”*

**4. Impacto en el negocio**

* **Optimización de baterías:** decidir **cuándo** cargar/descargar reduce ciclos innecesarios y prolonga la vida útil.
* **Menor compra de energía externa:** evitar picos en horas caras gracias a pronósticos más precisos.
* **Estabilidad operativa:** planificación proactiva ante horas de baja generación esperada.

**5. Riesgos, limitaciones y consideraciones**

* **Dataset sintético/público:** puede no reflejar una micro‑red local; riesgo de **overfitting contextual**.
* **Horizonte acotado (~5 meses):** no captura estacionalidad anual.
* **R² bajo:** esperable por alta variabilidad intra‑día y señal débil de exógenas en t−1; el KPI principal (RMSE) mejora sustancialmente.
* **Ética/transparencia:** el modelo es apoyo a la decisión; requiere monitoreo y recalibración.

**6. Conclusiones y recomendaciones**

* **Conclusión:** **LinearRegression** cumple el KPI (−30.21% RMSE vs. naïve) y muestra **robustez** en backtesting (40.69 ± 1.08 kWh). Es **interpretable, parsimonioso y estable** para el horizonte horario.
* **Recomendaciones técnicas (opcionales para futuras iteraciones):**
  1. **Backtesting** sistemático y monitoreo post‑despliegue.
  2. **Regularización** (Ridge/Lasso) y **tuning** ligero de RF/HGB.
  3. Integrar **pronósticos meteorológicos de t+1** (irradiancia/viento previstos) para un forecasting operativo real.
  4. **Análisis de residuales** por hora/nivel de generación para detectar sesgos.

**7. Anexos**

* **Artefactos generados por el código:**
  + outputs/metrics.csv — tabla de métricas por modelo.
  + outputs/timeseries\_target.png — serie temporal del target.
  + outputs/corr\_heatmap.png — matriz de correlación.
  + outputs/rf\_top15.png — importancias de RF (si entrenado).
  + outputs/summary.json — metadatos de ejecución.
* **Repositorio / ejecución:** notebook o script TFI\_Renewables\_EDA\_Modelado + dataset.