Documentación ML

Big Data II TerraWatt

Alma Gutierrez
Rafael Borge
Deniz Alcobendas
Íñigo Pérez

Índice

Introducción

Planteamiento del problema

El aumento constante de los precios de la electricidad, junto con los esfuerzos de la Unión Europea por promover un consumo energético más eficiente, pone de manifiesto la necesidad de contar con herramientas que permitan gestionar mejor el uso de energía en los hogares. En este contexto, TerraWatt surge como una solución innovadora diseñada para proporcionar a los usuarios una previsión precisa de sus facturas de electricidad y promover un uso más consciente y optimizado de la energía.

Los precios de la electricidad son sumamente variables y están influenciados por múltiples factores, como las condiciones económicas, los cambios sociales y el clima. Además, el consumo energético en los hogares depende de aspectos diversos, como el tipo de vivienda, la potencia contratada, el número de habitantes y las condiciones meteorológicas locales.

Por estas razones, el proyecto se enfoca en desarrollar dos modelos predictivos con el objetivo de:

1. **Anticipar los cambios en los precios de la electricidad.**
2. **Prever el consumo energético en los hogares para optimizar el uso de los recursos.**

La meta final es dotar a los usuarios de herramientas prácticas que les permitan:

- Planificar con antelación su consumo energético.
- Identificar momentos de alto costo y consumo.
- Fomentar un uso eficiente y sostenible de la energía.

Estos modelos predictivos se apoyan en la integración de datos históricos de consumo, precios y variables meteorológicas, enriquecidos con información adicional, como los días festivos y los patrones de actividad semanal. Al consolidar esta información, se generan datasets completos y estructurados que facilitan el análisis y la construcción de predicciones precisas de la factura eléctrica.

Preparación y limpieza de datos

Descripción del dataset

Para la construcción de los modelos predictivos en el proyecto TerraWatt, se consolidaron varios conjuntos de datos relevantes que permiten capturar los factores que influyen tanto en el precio de la electricidad como en el consumo energético. El dataset final integra información de tres fuentes principales:

- Datos Meteorológicos: Proporcionados por la AEMET, incluyen variables como temperatura diaria (media, mínima y máxima), precipitaciones, velocidad del viento,

presión atmosférica y horas de luz. Estos datos fueron recolectados desde aproximadamente 947 estaciones meteorológicas distribuidas por toda España, y se consolidaron por provincia y fecha.

- Precios de Electricidad: Extraídos de la Red Eléctrica Española, contienen información horaria y diaria de los precios de la electricidad en €/MWh durante los últimos 10 años. Incluyen precios por hora, promedio diario, y diferentes modalidades tarifarias.

- Días Festivos en España: Archivo que recopila los festivos nacionales y autonómicos, permitiendo evaluar cómo afectan estos días al comportamiento del consumo y los precios eléctricos.

Estos datasets fueron estructurados en formato CSV, con columnas normalizadas y compatibles entre sí, para facilitar su unión y análisis. El resultado es un dataset rico en variables tanto numéricas como categóricas, representativo a nivel provincial y diario, y preparado para alimentar modelos de machine learning.

Las variables de los datos meteorológicos son las siguientes:

- **FECHA**: object (representa una fecha)
- **ALTITUD**: float64 (numérica)
- **TMEDIA**: float64 (numérica, temperatura media)
- **TMIN**: float64 (numérica, temperatura mínima)
- **TMAX**: float64 (numérica, temperatura máxima)
- **DIR**: float64 (numérica, representa dirección del viento)
- **VELMEDIA**: float64 (numérica, velocidad media del viento)
- **RACHA**: float64 (numérica, racha máxima de viento)
- **SOL**: float64 (numérica, horas de sol)
- **PRESMAX**: float64 (numérica, presión máxima)
- **PRESMIN**: float64 (numérica, presión mínima)

En cuanto a las variables de los precios de la electricidad nos encontramos las siguientes:

- **Fecha**: object (representa una fecha)
- **Precio medio diario sin impuestos (€/MWh)**: float64 (numérica)
- **Impuesto eléctrico (€/MWh)**: float64 (numérica)
- **IVA (€/MWh)**: float64 (numérica)
- **Precio total con impuestos (€/MWh)**: float64 (numérica)

Por último, en los días festivos teníamos las variables siguientes:

- **Fecha**: object (texto, representa una fecha)
- **Provincia**: object (texto, nombre de la provincia)

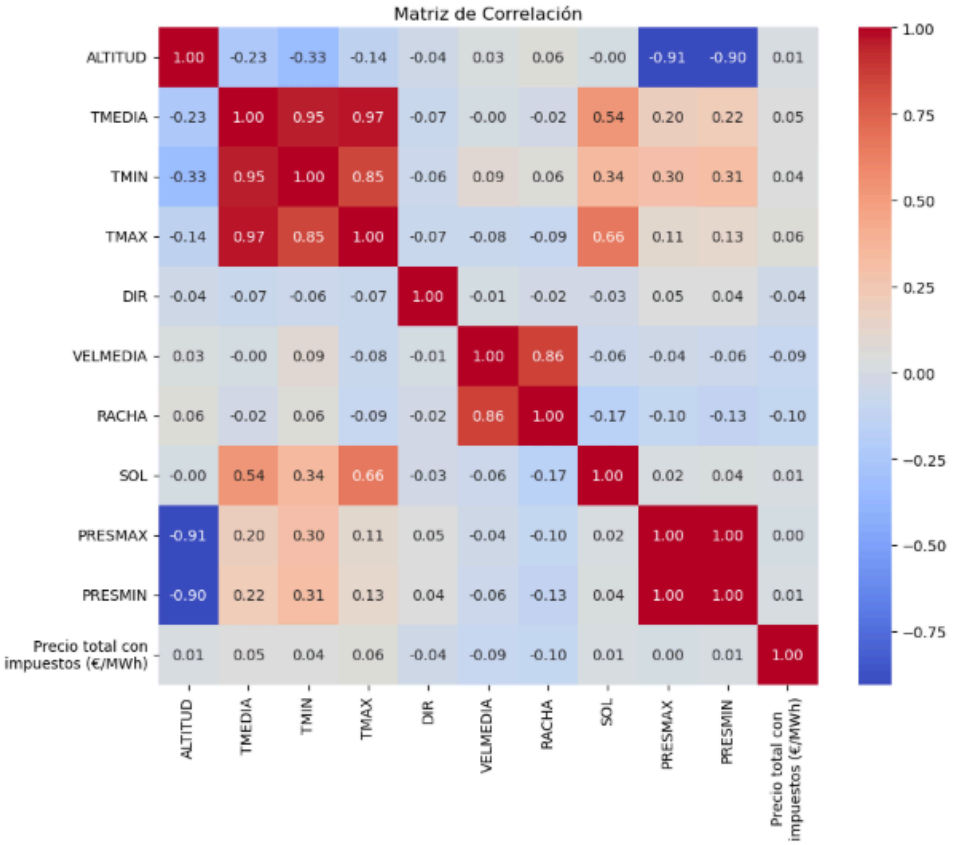
- **Festividad**: object (texto, nombre o tipo de festivo)

Nuestros datos, almacenados en un archivo CSV con 193,500 filas y 15 columnas, incluyen un conjunto completo de variables sin valores nulos. Entre las variables numéricas se encuentran la altitud, las temperaturas (media, máxima y mínima), las horas de sol y el precio total con impuestos (€/MWh). Las variables categóricas incluyen las provincias, días festivos y si corresponde a entre semana. La amplitud y variedad de los datos proporcionan una base sólida para analizar tendencias, relaciones y patrones.

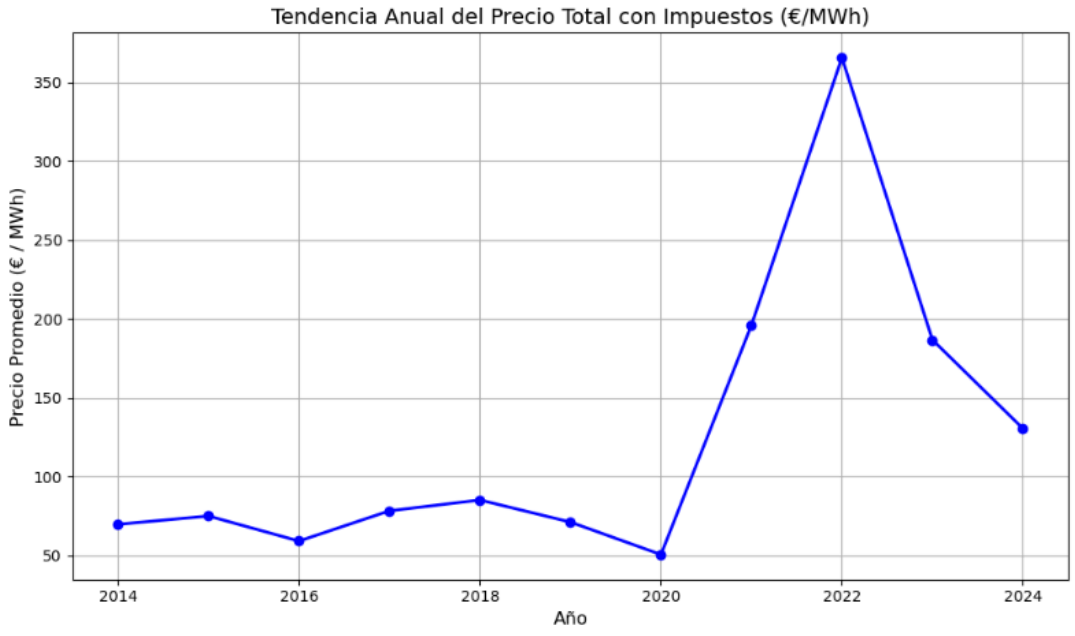
La altitud promedio es de 561.6 metros, mientras que las temperaturas tienen una media de 15.3°C, con valores que oscilan entre -13.8°C y 41.7°C. Las horas de sol promedian 7.4 horas diarias, y la velocidad media del viento es de 2.7 m/s, aunque se registran ráfagas de hasta 31.6 m/s. En cuanto a los precios energéticos, presentan una media de 125.6 €/MWh, pero con una alta dispersión, desde 2.2 €/MWh hasta 989.8 €/MWh, reflejando la alta variabilidad climática y de mercado en los datos.

La columna "Provincia" incluye 50 valores únicos, representando regiones tanto peninsulares como insulares. Esto permite realizar análisis específicos por ubicación. En el caso de las temperaturas, el análisis de boxplots muestra distribuciones razonables, aunque con valores atípicos en los extremos inferiores de la temperatura mínima y superiores en la máxima, que podrían reflejar eventos climáticos extremos.

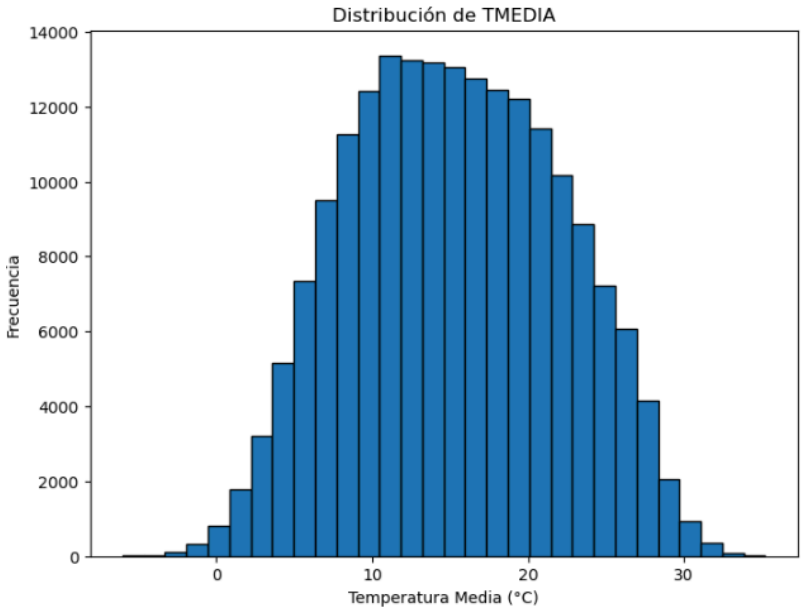
La matriz de correlación destaca relaciones fuertes entre variables climáticas. Las temperaturas están correlacionadas entre sí, y las horas de sol tienen una relación moderada con ellas. La presión atmosférica muestra una relación inversa con la altitud (-0.91), mientras que las ráfagas están relacionadas con la velocidad media del viento (0.86). Sin embargo, no hay correlaciones significativas entre las variables climáticas o geográficas y los precios de la energía, sugiriendo que estos últimos están influenciados por factores externos al clima.



En el análisis temporal, el precio promedio anual muestra un aumento drástico a partir de 2021, alcanzando un máximo en 2022, probablemente reflejando crisis de suministro o aumentos en la demanda. Aunque en 2024 se observa una ligera estabilización, los precios siguen siendo notablemente altos en comparación con años anteriores.



Finalmente, la distribución de la temperatura media sigue un patrón aproximadamente normal, con la mayoría de los valores entre 10°C y 20°C, reflejando las condiciones promedio en las regiones analizadas. Estos análisis iniciales subrayan la riqueza del conjunto de datos y la necesidad de explorar factores externos que puedan influir en los patrones observados.



Proceso de preparación y limpieza de datos

El proceso de preparación y limpieza de datos del proyecto TerraWatt fue exhaustivo y fundamental para garantizar la calidad y coherencia del dataset integrado, el cual combina información meteorológica, precios de electricidad y días festivos.

1. Revisión de archivos meteorológicos:

- Se identificó un archivo defectuoso (`1111X-20120301-20241103.csv`) que contenía una línea con formato incorrecto. Esta fue eliminada mediante un script, permitiendo una correcta lectura del archivo.

- Se eliminaron columnas irrelevantes como `INDICATIVO` y `NOMBRE` de las estaciones meteorológicas, ya que el análisis se realiza a nivel de provincia y estas columnas no aportaban valor en ese contexto.

- Se normalizó la columna "PROVINCIA", corrigiendo inconsistencias y duplicaciones en los nombres, como por ejemplo:

  - `GERONA` → `GIRONA`
  - `ORENSE` → `OURENSE`
  - `VIZCAYA` → `BIZKAIA`

- Se realizó un filtrado temporal para restringir los datos desde el 1 de abril de 2014, alineando el dataset con la disponibilidad de precios eléctricos.

2. Agrupación de datos meteorológicos:

- Se agruparon las mediciones por provincia y fecha, calculando promedios diarios de las variables meteorológicas, lo que permitió una consolidación eficiente de los datos y su posterior uso analítico.
- Cada provincia cuenta con un archivo CSV independiente con estos promedios, simplificando la gestión y uso del dataset.

3. Limpieza del dataset de precios de electricidad:

- Se eliminaron filas completamente vacías.
- Los valores faltantes en columnas numéricas se completaron con la media de la columna correspondiente.
- Se eliminaron columnas innecesarias como impuestos específicos, dejando solo aquellas relevantes para el modelo.
- Se corrigió el delimitador del archivo para asegurar una correcta lectura en el análisis.

4. Revisión de nombres de provincias en todos los archivos:

- Se normalizaron las variaciones y errores tipográficos mediante un diccionario de mapeo en Python, lo cual evitó problemas al unir datasets.

5. Unión de datos meteorológicos, de precios y de festivos:

- Se cargaron todos los archivos de meteorología y precios, ajustando fechas al mismo formato `YYYY-MM-DD`.
- Se creó una clave compuesta por “Fecha” y “Provincia” para facilitar la combinación.

- Se añadieron dos columnas nuevas:
    - "Festivo": indica si la fecha es festiva según el archivo `Festivos.csv`.
    - "Entre Semana": clasifica los días como laborales (lunes a viernes) o no laborales (sábado y domingo).

6. Exportación del archivo final:

- El resultado se consolidó en un archivo final `Modelo_Precios_Met_Fest.csv`, el cual incluye:
    - Datos meteorológicos diarios por provincia.
    - Precios diarios de electricidad.
    - Indicadores de si el día era festivo y si era entre semana.

Métricas del proceso

El proceso de consolidación y preparación de los datos meteorológicos generó una serie de transformaciones significativas en el volumen y estructura del dataset. Las principales métricas e impactos derivados de estas acciones son las siguientes:

1. Volumen inicial de datos:
Se partió de aproximadamente 947 archivos CSV, cada uno correspondiente a una estación meteorológica diferente en España. Estos archivos contenían registros diarios que, en algunos casos, abarcaban desde el año 1920 hasta 2024. Este volumen de información, aunque valioso históricamente, era excesivo y contenía datos irrelevantes para el objetivo del proyecto.

2. Filtrado temporal aplicado:
Para reducir el volumen y garantizar la compatibilidad con los datos de precios eléctricos (disponibles únicamente desde el 1 de abril de 2014), se aplicó un filtro que eliminó todos los registros anteriores a esa fecha. Esta decisión eliminó registros de más de 90 años en muchos archivos, reduciendo el volumen del dataset meteorológico en más de un 60%.

3. Registros inválidos y datos corruptos:
Durante la revisión de los archivos meteorológicos, se identificó un archivo defectuoso:

- **Archivo:** `1111X-20120301-20241103.csv`
- **Problema:** contenía una línea con un formato incorrecto que impedía la lectura del archivo completo.

Se procedió a eliminar dicha línea mediante un script automatizado, lo que permitió conservar el resto del archivo sin pérdida significativa de información.

4. Eliminación de columnas irrelevantes:
Los archivos originales incluían columnas como:

- `INDICATIVO`: código numérico de cada estación.
- `NOMBRE`: nombre específico de la estación meteorológica.

Dado que el análisis se realiza a nivel provincial, estos campos no aportan valor y solo incrementan el tamaño del dataset. Su eliminación ayudó a simplificar la estructura de los datos y facilitó su posterior agregación.

5. Normalización de nombres de provincias:
Se unificaron los nombres de 6 provincias que presentaban variantes inconsistentes, lo que evitó errores de duplicación y permitió una agregación precisa:

- `ARABA/ÁLAVA` → `ÁLAVA`
- `STA. CRUZ DE TENERIFE` → `SANTA CRUZ DE TENERIFE`
- `BALEARES` → `ILLES BALEARS`
- `GERONA` → `GIRONA`
- `ORENSE` → `OURENSE`
- `VIZCAYA` → `BIZKAIA`

Estos ajustes aseguraron una codificación única por provincia, lo que evitó errores de duplicación o agrupaciones incorrectas al unir con otros datasets.

6. Agrupación y reducción de granularidad:
Para facilitar el análisis y reducir la complejidad del dataset, los datos se agruparon por provincia y fecha, calculando el promedio diario de todas las variables meteorológicas disponibles (temperatura, viento, presión, etc.). Esto transformó los datos de múltiples estaciones en un único registro representativo por día y provincia.

7. Generación del dataset final:
Como resultado del proceso, se generaron 52 archivos CSV, uno por cada provincia, conteniendo únicamente los valores promedio diarios. Estos archivos fueron diseñados para integrarse directamente con los datasets de precios de la electricidad y festivos nacionales, permitiendo así la creación de un dataset único y coherente para el entrenamiento de los modelos.

Modelos de ML

Justificar el modelo elegido

El proyecto TerraWatt aborda dos problemas distintos mediante enfoques de aprendizaje automático adaptados a la naturaleza de cada uno:

Predicción del Precio de la Electricidad – Modelo LSTM (Long Short-Term Memory)

Se optó por un modelo de red neuronal LSTM para predecir los precios de la electricidad, debido a que se trata de un problema de series temporales donde los valores futuros dependen en gran medida de patrones históricos. Las LSTM son especialmente adecuadas para este tipo de tareas porque:

- Son capaces de capturar dependencias a largo plazo y relaciones no lineales.
- Evitan problemas comunes en redes neuronales recurrentes tradicionales, como la desaparición del gradiente.
- Pueden adaptarse bien a datos volátiles y ruidosos, como los precios del mercado eléctrico.

En este contexto, las LSTM permiten modelar la estacionalidad, tendencias y picos de precios con mayor precisión que otros modelos más simples o estáticos.

Predicción del Consumo Energético por Provincia – Modelos de Regresión Avanzada

Para la estimación del consumo energético se evaluaron y compararon diversos modelos de regresión en cada provincia por separado, aplicando el mismo conjunto de modelos en todos los casos. Posteriormente, se seleccionó el modelo con mejor desempeño en cada provincia. Los modelos considerados fueron:

- **Regresión Lineal:** útil por su simplicidad y capacidad interpretativa, aunque limitada en presencia de relaciones no lineales.
- **Random Forest Regressor:** robusto frente a datos ruidosos y capaz de modelar interacciones complejas sin asumir linealidad.
- **Gradient Boosting Regressor:** mejora la precisión ajustando errores de predicción secuencialmente.
- **XGBoost:** versión optimizada del Gradient Boosting, con alta eficiencia computacional, manejo automático de valores faltantes y regularización para prevenir sobreajuste.

Para determinar qué modelo ofrecía las mejores predicciones para cada provincia, implementamos una evaluación automática basada en métricas de desempeño. Se calcularon las siguientes métricas:

- **Error Cuadrático Medio (MSE - Mean Squared Error)**: Esta métrica penaliza los errores grandes en la predicción y es útil para evaluar la precisión del modelo en términos absolutos. Se utilizó como referencia principal para minimizar las desviaciones de las predicciones.

- **Coeficiente de Determinación (R^2 Score)**: Mide qué porcentaje de la variabilidad de los datos es explicada por el modelo. Un valor de R^2 cercano a 1 indica un ajuste muy bueno, mientras que valores cercanos a 0 indican que el modelo no logra explicar la variabilidad en los datos.

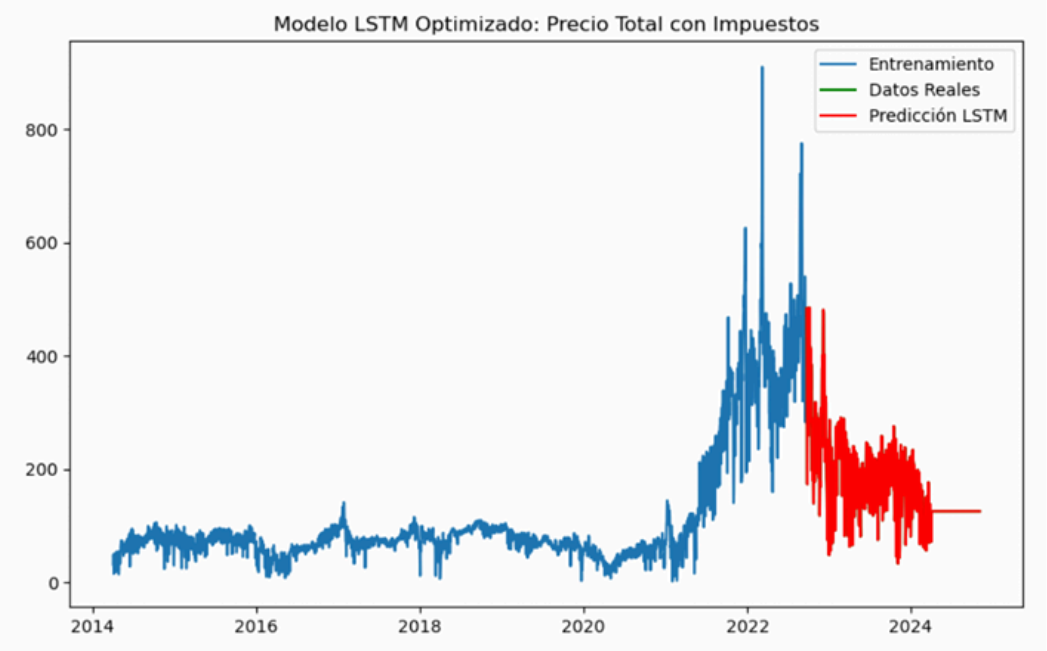
A partir de estas métricas, se seleccionó el modelo con mejor desempeño en cada provincia.

Ajuste de parámetros y entrenamiento

Predicción del Precio de la Electricidad – Modelo LSTM (Long Short-Term Memory)

El primer paso en el entrenamiento del modelo fue la carga y preprocesamiento de los datos. Para ello, se extrajeron los precios históricos de la electricidad desde un archivo CSV y se ordenaron cronológicamente. Luego, se dividió el conjunto de datos en un 80% para entrenamiento y un 20% para prueba, preservando el orden temporal. Posteriormente, se normalizaron los valores con el fin de mejorar la estabilidad del entrenamiento y se estructuraron los datos en el formato requerido por LSTM, con las dimensiones adecuadas de muestras, pasos de tiempo y características.

Durante la fase de entrenamiento, se utilizó el optimizador Adam, que proporciona una convergencia rápida y estable en modelos profundos. La función de pérdida seleccionada fue el error cuadrático medio (Mean Squared Error - MSE), dado que es la más utilizada en problemas de regresión. Para evitar el sobreajuste, se implementó la técnica de "Early Stopping", que detiene el entrenamiento si el error de validación no mejora después de 10 épocas consecutivas. El modelo se entrenó durante 100 épocas con un tamaño de batch de 32, logrando un equilibrio entre precisión y tiempo de cómputo.



Predicción del Consumo Energético por Provincia – Modelos de Regresión Avanzada

El preprocesamiento de datos incluyó la codificación de variables categóricas mediante One-Hot Encoding, en particular para variables como el tipo de vivienda. También se extrajeron características temporales, como el mes y el año, y se normalizaron las variables meteorológicas y de consumo utilizando StandardScaler.

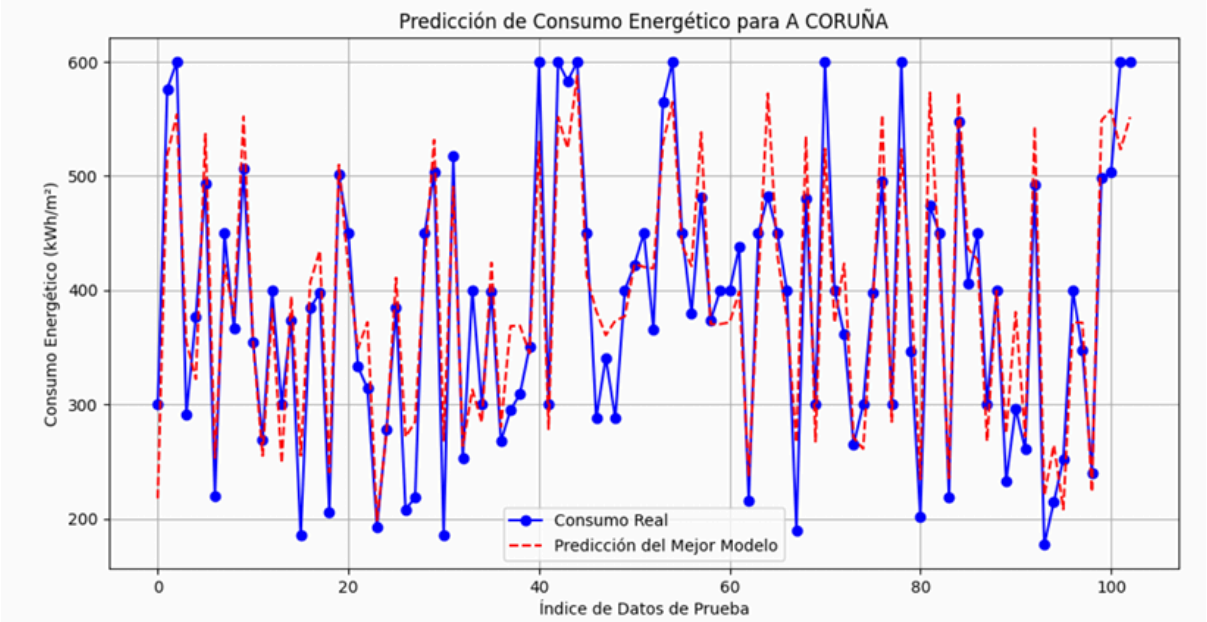
Las variables con las que contamos son las siguientes:

Variable	Descripción
Fecha	Fecha del registro de datos
Provincia	Provincia en la que se realizó la medición
Consumo energético (kWh/m²)	Consumo de energía por metro cuadrado en kilovatios-hora
Media de residentes	Número promedio de personas que habitan la vivienda
Potencia contratada (kW)	Potencia eléctrica contratada en kilovatios
Tipo de vivienda	Clasificación del tipo de vivienda (ej. piso, casa unifamiliar, etc.)
TMEDIA	Temperatura media diaria en grados Celsius
TMIN	Temperatura mínima diaria en grados Celsius

TMAX	Temperatura máxima diaria en grados Celsius
VELMEDIA	Velocidad media del viento en km/h o m/s (según la fuente de datos)
SOL	Horas de sol diarias
PRESMAX	Presión atmosférica máxima del día (hPa)
PRESMIN	Presión atmosférica mínima del día (hPa)

Para entrenar y evaluar los modelos, los datos se dividieron en un 80% para entrenamiento y un 20% para prueba. Se entrenaron los cuatro modelos en cada provincia y se calcularon métricas de desempeño como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2). Finalmente, se seleccionó el modelo con el menor MSE en cada provincia para garantizar la mejor precisión posible.

	Model	MSE	R^2
1	Random Forest	1576.981749	0.897034
0	Linear Regression	1609.822065	0.894889
2	Gradient Boosting	1674.268976	0.890681
3	XGBoost	1887.254800	0.876775



Métricas de evaluación del modelo

Predicción del Precio de la Electricidad – Modelo LSTM (Long Short-Term Memory)

Se utilizó un modelo de red neuronal LSTM para predecir los precios de la electricidad, dado que se trata de un problema de series temporales donde los valores futuros dependen fuertemente de patrones históricos. Las LSTM son especialmente adecuadas para esta tarea, ya que capturan dependencias a largo plazo, modelan relaciones no lineales y son robustas frente a datos ruidosos.

Para evaluar su desempeño, se calcularon las siguientes métricas:

- **Error Cuadrático Medio Raíz (RMSE)**: mide la magnitud promedio del error de predicción en las mismas unidades que el precio (€/MWh), penalizando especialmente los errores grandes.

- **Coeficiente de Determinación (R^2 Score)**: refleja qué porcentaje de la variabilidad en los precios reales es explicado por el modelo.

Los resultados mostraron que la red LSTM predijo con buena precisión las fluctuaciones del precio de la electricidad, logrando modelar correctamente tendencias, estacionalidad diaria y dependencias a largo plazo en la serie de precios.

Predicción del Consumo Energético por Provincia – Modelos de Regresión Avanzada

Para la estimación del consumo energético, se aplicaron varios modelos de regresión a cada provincia de forma individual, y se seleccionó el que ofrecía el mejor desempeño según las métricas obtenidas. Los modelos evaluados fueron: Regresión Lineal, Random Forest Regressor, Gradient Boosting Regressor y XGBoost.

La elección del modelo más adecuado varió según las características de los datos provinciales. En algunas provincias, XGBoost destacó por su capacidad de modelar relaciones complejas en grandes volúmenes de datos. En otras, Random Forest ofreció mejores resultados, especialmente en conjuntos de datos más dispersos, es decir, con alta variabilidad del consumo energético entre días o bajo distintas condiciones meteorológicas. En provincias con datos más homogéneos, donde los patrones de consumo eran más estables y regulares, Gradient Boosting superó a los demás modelos.

Las métricas utilizadas para comparar los modelos fueron:

- **Error Cuadrático Medio (MSE)**: mide el promedio de los errores al cuadrado, penalizando desviaciones grandes en las predicciones.

- **Coeficiente de Determinación (R^2 Score)**: indica la proporción de la variabilidad en el consumo energético explicada por el modelo.

Para garantizar un desempeño óptimo, se almacenó el modelo seleccionado de forma independiente para cada provincia, permitiendo así adaptar la predicción a las particularidades de cada conjunto de datos local.

Conclusiones

El proyecto TerraWatt ha demostrado la viabilidad y utilidad del uso de modelos de machine learning para abordar dos problemas clave en la gestión energética: la predicción del precio de la electricidad y la estimación del consumo energético por provincia. A través del desarrollo y entrenamiento de modelos avanzados, como las redes neuronales LSTM y los algoritmos de regresión (XGBoost, Random Forest, Gradient Boosting), se logró construir una herramienta capaz de proporcionar información precisa y útil para la toma de decisiones energéticas informadas.

Por un lado, el modelo LSTM permitió anticipar con precisión las fluctuaciones del precio de la electricidad. Gracias a su capacidad para aprender dependencias temporales, el modelo fue capaz de identificar patrones horarios recurrentes, como los valles de menor demanda y los picos de consumo en horas punta, que se repiten de forma consistente a lo largo de los días. Además, capturó tendencias a más largo plazo y efectos estacionales, mejorando así la precisión de las predicciones.

La preparación y limpieza rigurosa de los datos fue clave para garantizar la calidad del modelo, integrando correctamente variables climáticas, económicas y sociales, como los días festivos y la distribución regional.

En conjunto, los resultados obtenidos refuerzan el potencial del machine learning como herramienta para fomentar un consumo energético más consciente, eficiente y sostenible. Como trabajo futuro, se podría integrar información en tiempo real o explorar modelos híbridos que combinen series temporales con predicciones regionales, mejorando aún más la capacidad de adaptación y respuesta de la herramienta.