Documentación

Arquitectura de datos

Big Data II TerraWatt

Alma Gutierrez
Rafael Borge
Deniz Alcobendas
Íñigo Pérez

Índice

Introducción

La base de cualquier modelo predictivo sólido está en la calidad y relevancia de los datos utilizados, y este proyecto no es la excepción. Para predecir el precio medio diario de la electricidad en España, se optó por un enfoque centrado en los datos, recopilando y combinando información clave que influye directamente en cómo varía el precio de la luz día a día.

La solución propuesta parte de integrar múltiples fuentes de datos que reflejan distintos aspectos del comportamiento del mercado eléctrico. Se recopilaron datos meteorológicos detallados de la AEMET, precios históricos extraídos de la Red Eléctrica Española y un calendario completo de días festivos a nivel nacional y autonómico. Cada uno de estos conjuntos de datos aporta una pieza fundamental: el clima puede afectar tanto la producción como el consumo de energía, los precios pasados ayudan a identificar tendencias, y los festivos alteran los hábitos de consumo habituales.

Todos los datos fueron procesados, limpiados y consolidados en un único archivo listo para alimentar el modelo de predicción. Este enfoque permite al modelo captar no solo patrones históricos, sino también el impacto de factores externos en tiempo real, como olas de calor o puentes festivos, lo que mejora considerablemente la precisión de las estimaciones.

Fuentes de datos

Para desarrollar un modelo que prediga el precio de la electricidad de forma precisa, fue fundamental recopilar datos relevantes que expliquen las causas detrás de sus variaciones. El objetivo principal de esta etapa fue reunir información que permita entender cómo influyen distintos factores (como el clima, el comportamiento del mercado y los eventos sociales) en el precio diario de la electricidad en el mercado mayorista español.

Las fuentes de datos utilizadas fueron:

1. **Datos meteorológicos (AEMET):**
   Se descargaron registros diarios procedentes de aproximadamente 947 estaciones meteorológicas repartidas por toda España. Estos datos, en formato CSV, incluyen variables como la temperatura (media, mínima y máxima), precipitaciones, velocidad del viento, presión atmosférica y horas de sol. Estas variables son clave porque el clima influye tanto en la generación (especialmente renovable) como en el consumo de electricidad. Por ejemplo, en días muy calurosos o fríos, el uso de aire acondicionado o calefacción dispara la demanda eléctrica.

2. **Precios históricos de la electricidad (Red Eléctrica Española):**
   A través de su API, se extrajo un archivo con los precios horarios de los últimos 10 años. Esta información permite analizar la evolución de los precios a lo largo del tiempo y detectar patrones como horas punta, variaciones estacionales y cambios por tipo de tarifa. También se calculó el precio medio diario, que es la variable objetivo del modelo. Gracias a su granularidad horaria y cobertura temporal, estos

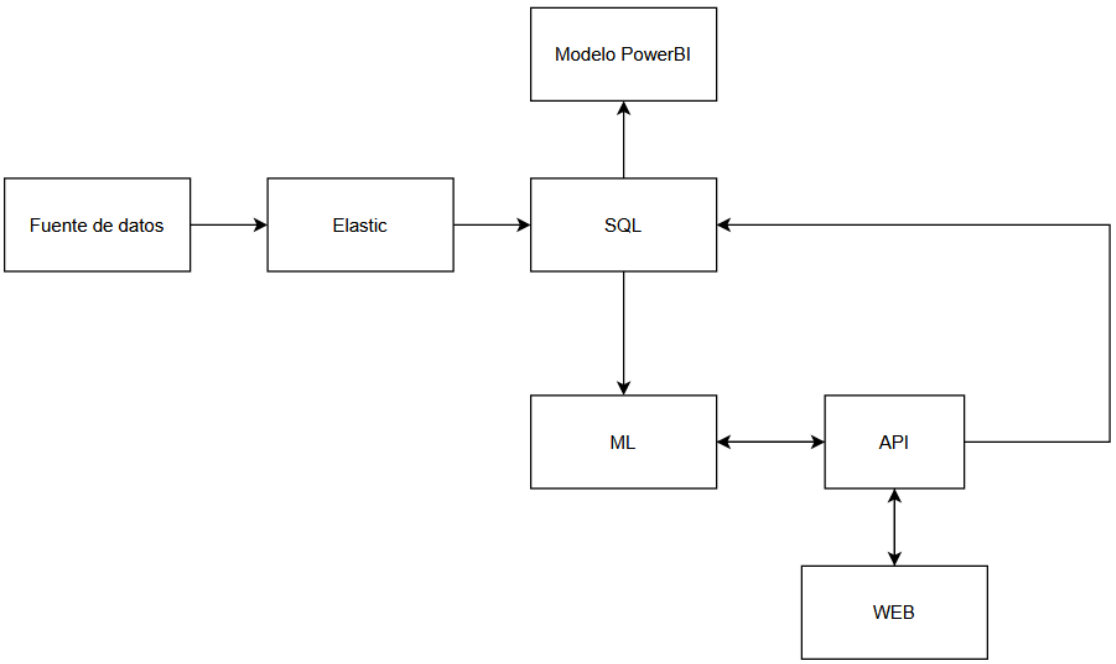
datos son fundamentales para entrenar el modelo con precisión y ajustar sus predicciones a la realidad del mercado eléctrico.

3. **Calendario de días festivos en España:**
   Se utilizó un archivo CSV que incluye tanto festivos nacionales como autonómicos. Esta información es importante porque en días festivos cambia el patrón de consumo: suele bajar la demanda industrial y aumentar el consumo doméstico, lo que puede afectar significativamente el precio. Además, al incluir festivos locales, se pueden identificar posibles variaciones regionales en la demanda y en los precios diarios.

El objetivo de integrar estas fuentes es construir un conjunto de datos completo y equilibrado que permita al modelo identificar cómo interactúan estos factores y cómo impactan en el precio de la electricidad. Con esta base de datos unificada, se facilita el análisis predictivo y se mejora la capacidad del modelo para anticipar fluctuaciones del mercado con mayor exactitud.

Arquitectura de datos

Diagrama de bloques



Almacenamiento

El almacenamiento de datos se realiza en una base de datos MySQL alojada en Clever Cloud. La base de datos contiene una tabla llamada PREDICCIONES_CLIENTES, que almacena tanto la información introducida por el usuario como los resultados generados por el sistema. Esta tabla registra cada predicción de forma estructurada, con los siguientes campos:

POTENCIA: Potencia contratada (kW).

NRESIDENTES: Número de residentes en la vivienda.

TIPOVIVIENDA: Tipo de vivienda seleccionada (Piso, Adosado, etc.).

PROVINCIA: Provincia para la cual se realiza la predicción.

MES: Mes para el cual se desea obtener una estimación.

PREDICCION_CONSUMO: Consumo energético predicho en kWh.

PREDICCION_PRECIO: Precio medio de la electricidad predicho en €/kWh.

COSTE_POTENCIA: Coste estimado por potencia contratada.

COSTE_ESTIMADO: Coste total estimado de la factura.

Este sistema de almacenamiento permite llevar un registro histórico de todas las solicitudes realizadas por los usuarios, lo que facilita el análisis posterior y la trazabilidad de resultados.

Ingesta de datos

La ingesta de datos comienza en la interfaz web, donde el usuario introduce los siguientes datos mediante un formulario:

- Potencia contratada
- Número de residentes
- Tipo de vivienda
- Provincia
- Mes

Al enviar el formulario, los datos se envían al backend mediante una petición HTTP POST al endpoint /transformar de una API desarrollada con FastAPI. La comunicación se realiza en formato JSON.

La función enviarDatos() en JavaScript es la encargada de:

- Capturar los datos del formulario.
- Convertirlos al formato adecuado (números, cadenas).
- Enviarlos al backend usando fetch.

Este mecanismo permite una integración sencilla y eficiente entre el cliente web y el servidor de predicción.

Procesamiento de datos

Una vez que los datos llegan al backend, se inicia el procesamiento:

1. Categorización y transformación:

Se generan variables dummies según el tipo de vivienda.
Se filtran y agregan datos meteorológicos históricos de la provincia y mes seleccionados.

2. Predicción de consumo:

Se utiliza un modelo de machine learning previamente entrenado (guardado en formato .pkl) para predecir el consumo energético mensual esperado en función de los datos del usuario y las condiciones meteorológicas.

3. Predicción del precio:

Se simulan los precios de la electricidad día a día para el mes seleccionado, utilizando otro modelo ML.
Se calcula el precio medio mensual estimado en €/MWh, que luego se transforma a €/kWh.

4. Cálculo de costes:

Se calcula el coste de la potencia contratada (potencia * precio_kwh * 30).
Se estima el coste por consumo (consumo * precio_kwh).
Se suman ambos para obtener el coste total estimado de la factura.

5. Almacenamiento final:

Toda la información se guarda automáticamente en la base de datos MySQL a través de una conexión directa con mysql.connector.

El proceso está completamente automatizado, permitiendo al usuario obtener resultados instantáneos y al mismo tiempo registrar la información en una base de datos centralizada

Retos y limitaciones

Heterogeneidad y volumen de los datos meteorológicos: El procesamiento de datos provenientes de 947 estaciones meteorológicas, con diferentes formatos y periodos temporales, supuso un reto considerable a la hora de normalizar y consolidar la información a nivel provincial.

Normalización geográfica compleja: La necesidad de unificar criterios de nomenclatura (por ejemplo, nombres de provincias con variaciones como "GERONA" vs. "GIRONA") implicó un trabajo manual y automatizado para evitar errores de integración entre datasets.

Alta variabilidad regional y temporal: Los precios de la electricidad y el consumo energético varían mucho según la provincia y la época del año. Esto obligó a entrenar modelos distintos por provincia, lo cual aumentó la complejidad del sistema y el tiempo de entrenamiento.

Ausencia de actualización en tiempo real: Los modelos utilizan datos históricos y no se reentrenan automáticamente con nueva información, lo que puede afectar su precisión con el tiempo si cambian las condiciones del mercado o del clima.

Posibles mejoras

Automatización del pipeline de datos y modelos: Implementar flujos de trabajo automáticos que permitan actualizar los datos y reentrenar los modelos de forma periódica, sin intervención manual. Esto garantizaría que las predicciones se mantengan actualizadas frente a nuevos patrones de consumo o precios.

Predicciones por franjas horarias: Desarrollar modelos específicos que trabajen a nivel horario en lugar de diario o mensual, permitiendo a los usuarios optimizar su consumo en función de las horas más económicas del día (horas valle, punta y llano).

Integración con datos en tiempo real: Añadir la capacidad de recibir y procesar datos conforme se generan, como los precios de la electricidad en directo. Esto permitiría que las predicciones y sugerencias del sistema se actualicen al momento, reflejando las condiciones reales y actuales del mercado o del consumo doméstico.