Assesment ensemble learning

ML II



Grupo 2

Javier Álvarez Martínez

David Cocero Quintanilla

Raúl González Gómez

abril de 2024

Índice

[1. Introducción 2](#_Toc163297361)

[1.1 Contexto 2](#_Toc163297362)

# Introducción

## 1.1. Contexto

Los ensembles, también llamados métodos de ensamblaje, son una técnica en el aprendizaje automático. En lugar de depender de un solo modelo, los ensembles combinan las predicciones de varios modelos para obtener resultados más precisos y robustos. Esto se basa en la idea de que juntar las decisiones de múltiples modelos puede superar las limitaciones individuales, mejorando la precisión y evitando el sobreajuste.

Los ensembles se inspiran en la idea de que las decisiones tomadas en grupo suelen ser más acertadas que las individuales. En el aprendizaje automático, esto significa que un conjunto de modelos "votando" o combinando sus predicciones probablemente será más preciso que un solo modelo, especialmente en tareas complejas o con datos difíciles de modelar.

Existen diferentes métodos para crear y combinar estos modelos en un ensemble. Algunos de los más populares son:

* **Bagging:** Se generan múltiples conjuntos de datos de entrenamiento mediante muestreo con reemplazo y se entrena un modelo en cada uno. La decisión final se toma por mayoría o promedio.
* **Boosting**: Los modelos se entrenan de manera secuencial, cada uno intentando corregir los errores del anterior. La predicción final se combina ponderando la precisión de cada modelo.
* **Stacking**: Se entrenan múltiples modelos diferentes con los mismos datos y luego se combinan sus predicciones para aprender a optimizar la combinación de las predicciones individuales para mejorar el rendimiento global.

## 1.2. Descripción del problema

El problema abordado en este documento se centra en la estimación de las utilizaciones horarias de la energía solar fotovoltaica para un día específico, empleando un conjunto de variables disponibles y aplicando técnicas de aprendizaje automático, enfocándonos en los métodos de ensamblaje. Este desafío se enmarca en el contexto de la gestión y optimización de la energía solar, una fuente de energía renovable crucial para la transición energética global. La precisión en la predicción de la utilización de la energía solar fotovoltaica es esencial para mejorar la eficiencia en la generación y distribución de energía, así como para la planificación y operación de las redes eléctricas.

El objetivo principal es estimar las utilizaciones horarias de energía solar para un día determinado, basándose en variables disponibles. Esto implica predecir cuánta energía solar fotovoltaica se utilizará en cada hora del día, lo cual es un desafío complejo debido a las numerosas variables que pueden influir en la generación de energía solar, como la irradiación solar, la hora del día, la estacionalidad y las condiciones meteorológicas, entre otras.

Antes de proceder con la modelación, se requiere realizar un análisis exploratorio de los datos disponibles. Este análisis es esencial para entender las características de los datos, identificar patrones, anomalías, correlaciones y cualquier otro insight que pueda ser relevante para la modelación posterior.

El trabajo se centrará en el uso de técnicas de ensamblaje para abordar este problema, comparándolas con métodos más directos. Los ensembles, al combinar las predicciones de múltiples modelos, pueden ofrecer mejoras significativas en la precisión y robustez de las predicciones.

La validación del modelo es fundamental para asegurar que las predicciones sean confiables y aplicables a datos nuevos o no vistos. Se solicita realizar una validación honesta, lo que implica aplicar técnicas de validación cruzada o división de datos en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo de manera justa y sin sesgos.

## Análisis exploratorio

## 2.1. ETL de los datos

En este primer apartado lo que se busca es entender los datos, saber decir que hay en ellos, errores que encontremos, dificultades o insights que nos puedan venir bien a la hora de introducir los datos en nuestro modelo.

Partimos de 2 ficheros csv que contienen datos de irradiación, una media cada 3 horas y por otro lado otro archivo csv que contiene los datos de utilización, es decir, una media cada 3 horas del valor que se genera de la potencia que se tiene instalada.

El primer archivo tiene la siguiente estructura:

Calendario

Descripción generada automáticamente con confianza media

Ilustración 1. Inspección del dataframe de irradiación

El archivo de utilización tiene la siguiente estructura:

Imagen en blanco y negro

Descripción generada automáticamente con confianza media

Ilustración 2. Inspección del dataframe de utilización

Una vez cargados los archivos, lo primero que se ha hecho es una descripción de las variables, para que nos den información previa de la distribución de los datos en las columnas numéricas, lo que puede ser útil para identificar posibles problemas en los datos, como valores atípicos o errores, así como para comprender mejor la naturaleza de los datos y cómo están distribuidos. Además, también se ha querido observar que no hay valores nulos o valores duplicados y una vez que es así se ha querido preparar los dataframes para su unión. Para no duplicar columnas, primeramente, se ha borrado del dataframe de irradiación las columnas de año, mes y día de la semana para que al unir los dataframes por Fecha no se nos generen columnas con la misma información. Una vez realizado esto se ha hecho un inner join de los 2 dataframes y se ha generado el archivo csv que será posteriormente del que partamos para nuestros modelos.

## 2.2. Insights de los datos

A continuación, lo que se busca es dar una información previa de los datos, con un mayor detalle que en el apartado anterior, obteniendo información que puede ser valiosa para la posterior creación y validación de los modelos.

En primer lugar, se va a representar La distribución de la irradiación y de la utilización para las principales horas del día, de 9.00 a 18.00

Gráfico, Histograma

Descripción generada automáticamente

Ilustración 3. Distribución de la irradiación para las principales horas del día

Gráfico, Histograma

Descripción generada automáticamente

Ilustración 4. Distribución de la utilización solar para las principales horas del día

Lo que podemos observar de las 2 gráficas anteriores es lo siguiente:  
De 15 a 18 horas parece el periodo de estos tres donde menos radiación se recoge, y esto correlaciona bien con la utilización obtenida de las placas. Sin embargo, por ejemplo, para la franja de 9 a 12 horas la utilización tiene una distribución distinta a la radiación, parece más sesgada a la derecha. Lo mismo pasa con la franja de 12 a 15 horas. Otra cosa interesante es que a pesar de que la radiación de 12 a 15 es mayor, la distribución de la utilización en ese periodo es algo menor que la de 9 a 12. Es curioso que la utilización es mayor en horas en las que la irradiación es menor, en teoría hay menos sol, sin embargo, entendemos que la temperatura afecta, es decir, que cuanto mayor es la temperatura que tenemos en nuestra placa menos es la utilización que podemos obtener de la misma ya que los paneles están preparados para funcionar a una temperatura de trabajo óptima alrededor de 25ºC, en el momento que se sobrepasa, la eficiencia de conversión disminuye ya que se aumenta la resistencia eléctrica interna del panel.

Para medidas entre 0 y 6 de la mañana la utilización es cercana a, además los registros de irradiación también están a 0 todos los días. De cara a nuestro modelo hay que tener cuidado y tomar decisiones para que no pueda afectar en exceso al MSE.

Por último, queremos observar la relación y correlación entre las variables. En primer lugar, se va a representar la relación entre las variables de irradiación:

Diagrama

Descripción generada automáticamente

Ilustración 5. Relación entre variables de irradiación.

El gráfico pairplot sugiere correlaciones significativas y patrones entre ciertas variables, con algunas distribuciones mostrando sesgo o multimodalidad. Los datos parecen tener valores atípicos y distintas escalas que podrían requerir normalización. La concentración de datos en regiones específicas indica rangos operativos comunes o subgrupos dentro de las variables. Estas visualizaciones proporcionan una visión valiosa para el preprocesamiento y la selección de características en el modelado predictivo.

A continuación, se va a representar la relación de las variables para la utilidad.

Diagrama

Descripción generada automáticamente con confianza media

Ilustración 6. Relación entre variables de utilización solar.

El pairplot presentado muestra diversas relaciones entre múltiples variables, indicando patrones lineales y no lineales así como agrupaciones de datos que podrían sugerir dependencias. Las distribuciones marginales en la diagonal varían desde distribuciones concentradas hasta multimodales, señalando posibles subpoblaciones. Valores atípicos y clusters son visibles, lo cual podría influir en modelos predictivos. Estas visualizaciones son clave para identificar características relevantes y guiar la estrategia de modelado.

Finalmente se quiere realizar una representación de la correlación de las variables y para ello se ha representado en el siguiente heatmap:

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Ilustración 7. Correlación entre variables

La información que queremos destacar es que hay una correlación importante de las variables de Irradiación y de Utilización solar. Además, no hay correlación de los outputs con variables de día y día de la semana por lo que de cara al modelado podemos prescindir de ellas. Sin embargo, con año y mes hay algo de correlación (negativa), aunque no es demasiado grande y vamos a decidir utilizarlas en nuestro modelo.

# Predicción de la irradiación

## 3.1. Linear Regression

La regresión lineal es un método de predicción utilizado en estadística y aprendizaje automático para modelar la relación entre una variable dependiente (target) y una o más variables independientes (features). En el caso de la regresión lineal simple, se modela esta relación mediante una línea recta, mientras que, en la regresión lineal múltiple, se utiliza un plano o hiperplano, dependiendo del número de variables independientes.