# Predicción de la energía solar producida por una granja solar

## Proyecto de Machine Learning

### Pablo Barbero De La Orden

Contenido

[Predicción de la energía solar producida por una granja solar 1](#_Toc121056062)

[Proyecto de Machine Learning 1](#_Toc121056063)

[Pablo Barbero De La Orden 1](#_Toc121056064)

[Objetivo y contexto del trabajo 3](#_Toc121056065)

[Datos empleados 3](#_Toc121056066)

[Limpieza de datos 4](#_Toc121056067)

[Reestructuración de los datos 4](#_Toc121056068)

[EDA 5](#_Toc121056069)

[Modelos 6](#_Toc121056070)

[Conclusión 7](#_Toc121056071)

## Objetivo y contexto del trabajo

Durante el último año se ha observado como las noticias acerca del precio de la luz abordaban de manera continua las portadas de los periódicos y las entradas de los informativos. Pero realmente quién fija el precio de la luz.

En primer lugar es importante determinar que la energía eléctrica no se puede almacenar, es decir ha de producirse bajo demanda. Una vez claro este concepto, se realiza una subasta marginalista. En la que las diferentes empresas energéticas acuden a la misma con la previsión de energía que podrán proporcionar. De esta manera se irá asignando la energía que se consumirá en función de aspectos monetarios y medioambientales, cuando el cupo se cumpla a todas las energías que han entrado en el cupo se les pagará el precio de la última.

Toda esta situación lleva a plantearse la importancia que tiene la predicción de la cantidad de energía que se va a poder aportar por parte de las empresas energéticas, porque recordemos que esta energía no se puede almacenar y también que se ha de ser capaz de cumplir con la cantidad de energía demandada para no dejar a un sector de la población sin electricidad.

Tras reflexionar un poco sobre las distintas fuentes de energía que se disponen, me llevo a intentar construir un modelo capaz de predecir la cantidad de energía solar producida.

## Datos empleados

Tras realizar una búsqueda exhaustiva de datos en relación a la problemática que me había propuesto, encontré datos sobre una competición realizada en Kaggle durante los años 2013 y 2014.

Esta competición disponía de datos sobre la energía solar diaria total entrante en 98 sitios de Oklahoma Mesonet, que servían como "granjas solares" para el concurso. Datos de predicción meteorológica numérica de entrada para el concurso procedentes de la versión 2 del Sistema Global de Predicción Ensemble (GEFS) de la NOAA/ESRL, recogidos de manera diaria y separados en los intervalos de tiempo de previsión 12, 15, 18, 21 y 24. Los datos de entrenamiento procederán de 1994-2007.

Las variables meteorológicas recogidas son las siguientes:

|  |  |  |
| --- | --- | --- |
| Variable | Descripción | Unidad |
| apcp\_sfc | Precipitación acumulada en 3 horas en la superficie | kg m-2 |
| dlwrf\_sfc | Media del flujo radiativo de onda larga descendente en la superficie | W m-2 |
| dswrf\_sfc | Media del flujo radiativo de onda corta descendente en la superficie | W m-2 |
| pres\_msl | Presión atmosférica al nivel medio del mar | Pa |
| pwat\_eatm | Agua precipitable en toda la profundidad de la atmósfera | kg m-2 |
| spfh\_2m | Humedad específica a 2 m sobre el suelo | kg kg-1 |
| tcdc\_eatm | Nubosidad total en toda la profundidad de la atmósfera | % |
| tcolc\_eatm | Condensación total integrada en la columna en toda la atmósfera | kg m-2 |
| tmax\_2m | Temperatura máxima de las últimas 3 horas a 2 m sobre el suelo | K |
| tmin\_2m | Temperatura mínima de las últimas 3 horas a 2 m sobre el suelo | K |
| tmp\_2m | Temperatura actual a 2 m sobre el suelo | K |
| tmp\_sfc | Temperatura de la superficie | K |
| ulwrf\_sfc | Radiación de onda larga ascendente en la superficie | W m-2 |
| ulwrf\_tatm | Radiación de onda larga ascendente en la parte superior de la atmósfera | W m-2 |
| uswrf\_sfc | Radiación ascendente de onda corta en la superficie | W m-2 |

Y la variable objetivo contiene la energía solar diaria total entrante en (J m-2).

Para simplificar el proyecto, se va a centrar en intentar predecir la cantidad de energía producida por una de las 98 granjas solares de Oklahoma, de esta manera no tenemos en cuenta la latitud y longitud geográfica de las granjas solares ya que para una misma granja solar es constante, pero estas dos variables pueden ser un factor influyente en la generación de energía.

### Limpieza de datos

En este proyecto la limpieza de datos se ha tratado sobre todo en que toda la información meteorológica provenía en formato netCDF4, con el consiguiente tratamiento de este tipo de dato para transformarlo en un pandas data\_frame.

### Reestructuración de los datos

Este apartado ha resultado clave dentro del proyecto. En primer lugar se tuvo que realizar una transformación en las coordenadas de la localización de las granjas solares, de manera que siguieran la misma unidad que las coordenadas de los netCDF4.

Con esta transformación se fue capaz de identificar la estación meteorológica más cercana a nuestra granja solar seleccionada nombrada como ‘APAC’ de manera que una vez agrupada por día la información meteorológia, mediante unas métricas de variables medias por día (recordemos que se disponía de 5 intervalos de tiempo de previsión), se pudo combinar ambos data\_frames con los que obteníamos combinados la información meteorológica y la cantidad de energía solar producida.

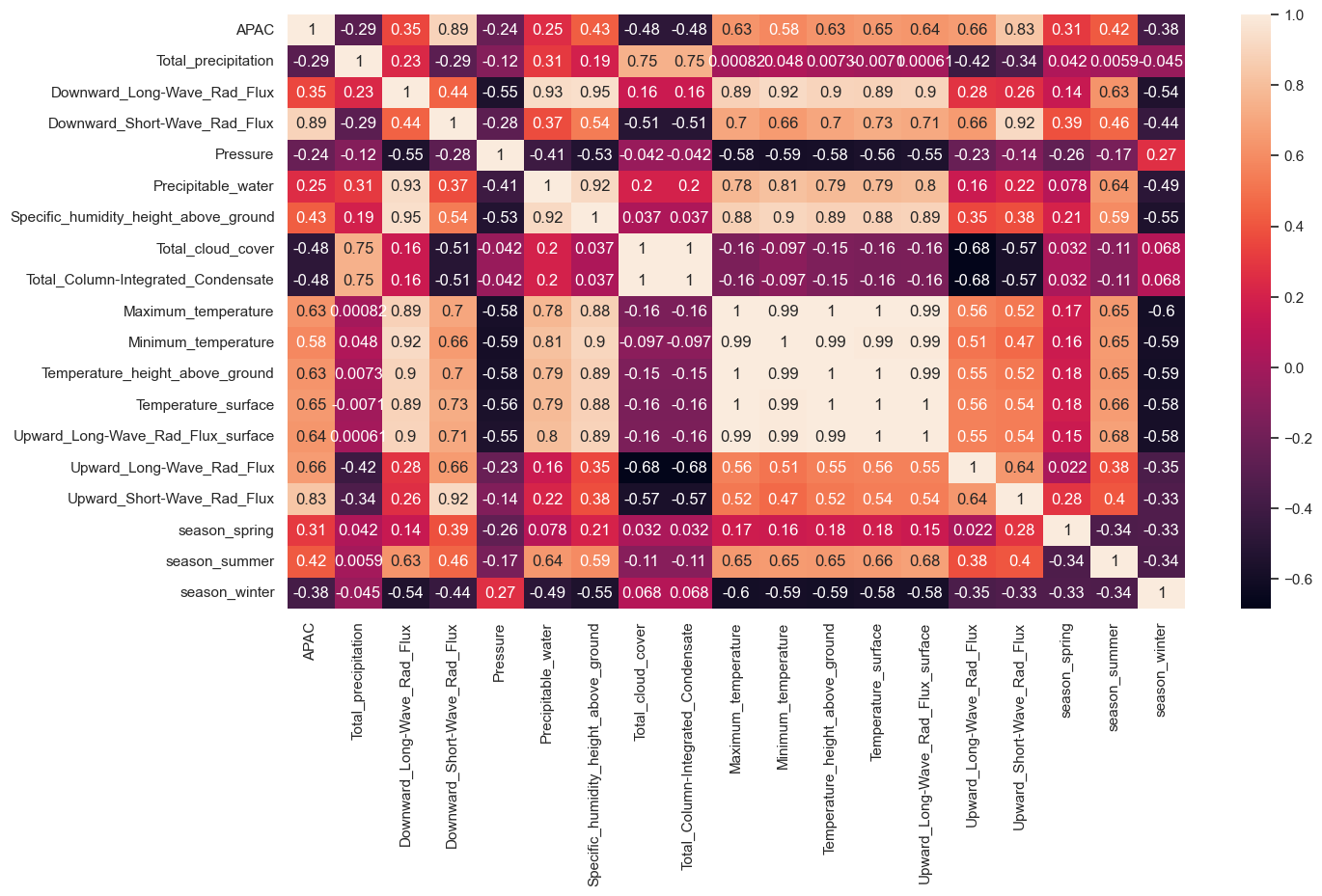
Además se creó una variable artificial que después de codifico utilizando un one\_hot\_encoder, a través de las fechas que indicaba la estación del año en la que fue recogida la información, de manera que quedara implícita las horas de luz que había en el día.

Una vez dispuesto el data\_frame definitvo con el que se iba a trabajar se realizó un sencillo y rápido análisis exploratorio de datos.

## EDA

Se comprobó que no se dispusieran ni de datos faltas ni de columnas constantes.

También se comprobó las correlaciones entre variables, obteniendo el siguiente gráfico.

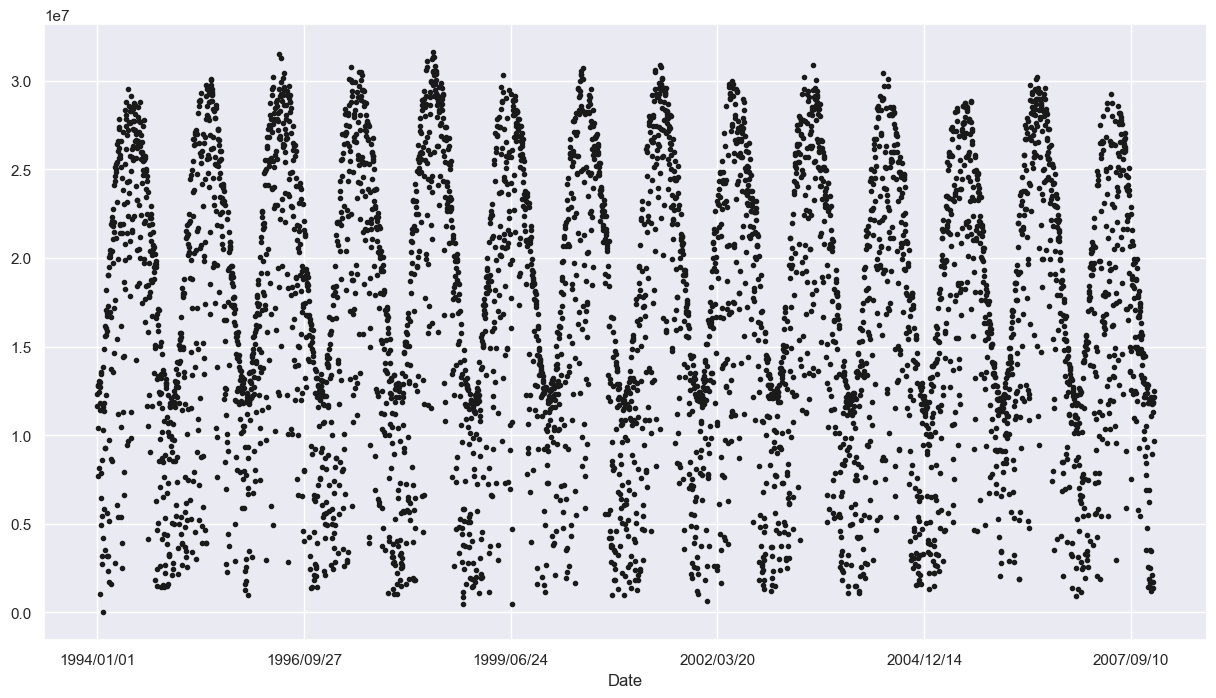


De manera que se observan correlaciones muy altas entre variables.

Sin embargo, tras haber pensado en eliminar aquellas columnas que tienen una correlacion superior al 0.85 con otras, quedándose con aquellas que estén más correlacionadas con el target. Es decir eliminando: Total\_Column-Integrated\_Condensate, Maximum\_temperature, Minimum\_temperature,Temperature\_height\_above\_ground, Downward\_Long-Wave\_Rad\_Flux, Upward\_Short-Wave\_Rad\_Flux, Precipitable\_water.

Se demuestra que no mejora la predicción de nuestro modelo por eso se siguen considerando todas las variables

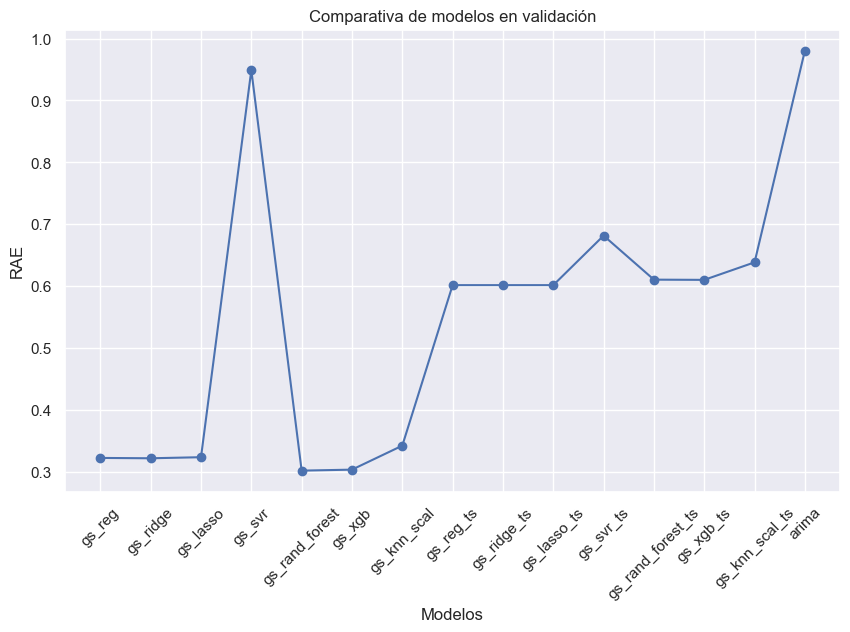
También se graficó la variable objetivo, obteniendo el siguiente gráfico:



El cuál me sugerió que quizá sería interesante estudiar también la predictibilidad de una serie temporal. Recordemos que nuestro principal objetivo es obtener una predicción precisa. Por ello en el proyecto se valorará tanto un problema de regresión como uno de series temporales.

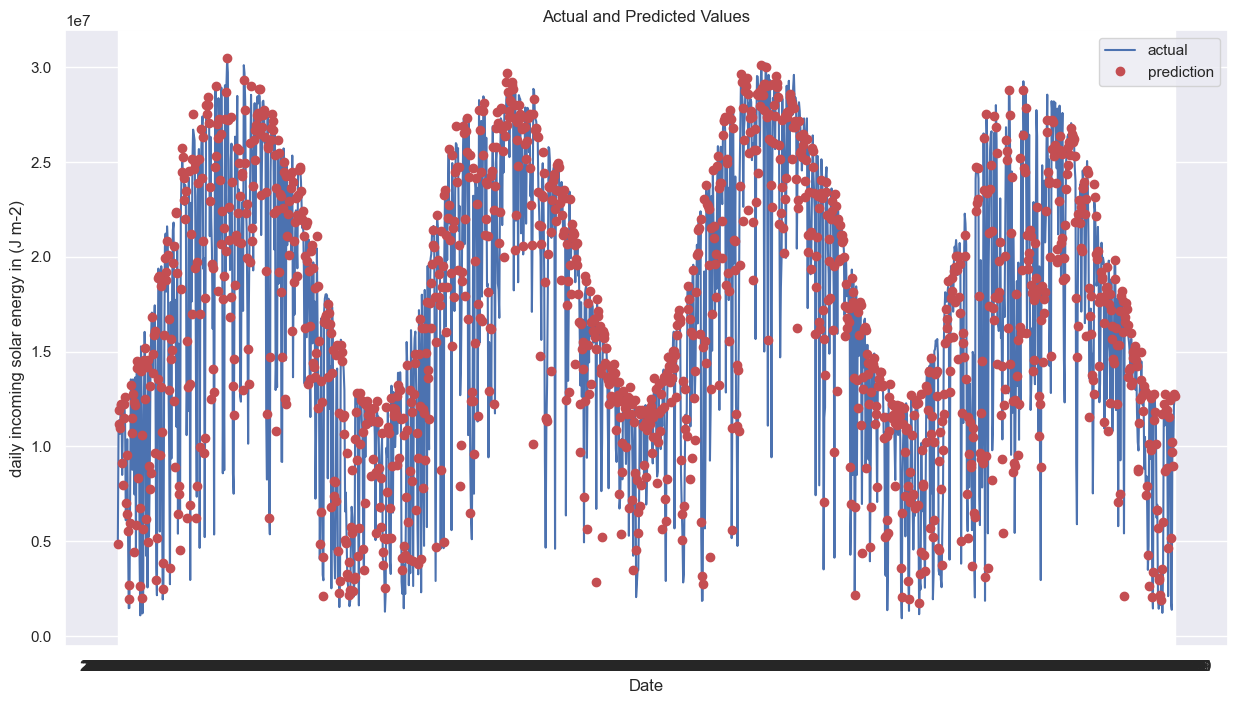
## Modelos

Debido a que el modelo está sujeto a la variable tiempo, se ha decidido dividir el conjunto de datos desde 1994 al 2000 como conjunto de entrenamiento, del 2001 al 2003 como validación y del 2004 al 2007 como conjunto de test. De manera que la aleatorización no influirá en nuestro modelo y respetará la variable implícita del tiempo.



Tras evaluar diferentes modelos mediante la métrica de error absoluto relativo ('RAE'), se determina que el mejor modelo corresponde con un random forest de n\_estimators = 1500 y un max\_features = 5.

Reentrenamos todo el modelo ya con la partición de train completa y realizamos las predicciones para el conjunto de test, obteniendo el siguiente resultado.



El modelo dispone de un error relativo absoluto de 0.32, es decir el modelo mejora en un 68% al modelo trivial, predecir todo con la media del entrenamiento.

## Conclusión

Para mi sorpresa el modelo que realiza las mejores predicciones corresponde con un problema de regresión y no de series temporales, si bien es cierto que podría optimizarse en mejor medida la parte de series temporales, considero que para la dificultad de predicción del problema y los datos empleados se alcanza una métrica aceptable.