



**Diseño de una solución para la predicción de radiación solar en diferentes escalas temporales para la región de la Comunidad de Castilla y León, España, para la gestión de proyectos de generación fotovoltaica**

Moisés Alfonso Guerrero Jiménez

Andrés Castaño Licona

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

Nombres completos, Título académico más alto

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2023

Cita	(Muñoz Zapata & Martínez Naranjo, 2023)
Referencia	Muñoz Zapata, L., & Martínez Naranjo, J. A. (2023). <i>Título del trabajo</i> Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.
Estilo APA 7 (2020)	



Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Centro de Investigación Ambientales y de Ingeniería (CIA).



Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - [www.udea.edu.co](http://www.udea.edu.co)

Rector: John Jairo Arboleda Céspedes.

Decano: Julio Cesar Saldarriaga Molina

Jefe departamento: Diego José Luis Botia Valderrama

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

**Encabezado**

**Dedicatoria**

Texto de dedicatoria centrado.


**Agradecimientos**

Texto de agradecimientos centrado.

## Tabla de contenido

Resumen	9
Abstract	10
1. Descripción del problema	11
1.1. Problema de negocio	11
1.2. Aproximación desde la analítica de datos	11
1.3. Origen de los datos	11
1.4. Métricas de desempeño	11
2. Objetivos	12
2.1. Objetivo general	12
2.2. Objetivos específicos	12
3. Datos	13
3.1. Datos originales	13
3.2. Datasets	13
3.3. Analítica descriptiva	13
4. Proceso de analítica	14
4.1. Pipeline principal	14
4.2. Preprocesamiento	14
4.3. Modelos	14
4.4. Métricas	14
5. Metodología	15
5.1. Baseline	15
5.2. Validación	15

---

5.3. Iteraciones y evolución	15
5.4 Herramientas	15
6. Resultados y discusión	16
6.1. Métricas	16
6.2. Evaluación cualitativa	16
6.3. Consideraciones de producción	16
7. Conclusiones	20
8. Recomendaciones	21
Referencias	22
Anexos	23
Anexo 1. Autoarchivo en Repositorio y documentos de interés	24
Anexo 2. Gestor de citas y referencias de Microsoft Word 	25
Anexo 3. Citas y referencias de material legal (leyes, decretos, sentencias, etc.)	27
Anexo 4. Ortografía y gramática	30
Anexo 5. Buscar, reemplazar y eliminar espacios (o palabras)	32
Anexo 6. Atajos de teclado útiles en Microsoft Word	33
Anexo 7. Sinónimos y antónimos	34
Anexo 8. Copiar y pegar sin formato	35
Anexo 9. Comparar dos documentos	36
Anexo 10. Control de cambios	37
Anexo 11. Insertar salto de página	39
Anexo 12. Recortar y abreviar direcciones web largas	40

---

### Lista de tablas

<b>Tabla 1</b> Resultados del test PBQ-SF (Personality Belief Questionnaire Short Form)	19
<b>Tabla 2</b> Características demográficas y tipo de tratamiento de hemodiálisis y diálisis peritoneal con la adherencia (SMAQ)	20
<b>Tabla 3</b> Categorías de la investigación	21

### **Lista de figuras**

<b>Figura 1</b> Portada Normas APA séptima edición 2020 en inglés	22
<b>Figura 2</b> Logo Universidad de Antioquia	22

**Siglas, acrónimos y abreviaturas**

<b>APA</b>	American Psychological Association
<b>Cms.</b>	Centímetros
<b>ERIC</b>	Education Resources Information Center
<b>Esp.</b>	Especialista
<b>MP</b>	Magistrado Ponente
<b>MSc</b>	Magister Scientiae
<b>Párr.</b>	Párrafo
<b>PhD</b>	Philosophiae Doctor
<b>PBQ-SF</b>	Personality Belief Questionnaire Short Form
<b>PostDoc</b>	PostDoctor
<b>UdeA</b>	Universidad de Antioquia



## Resumen

La energía solar fotovoltaica es una de las principales propuestas dentro de las energías renovables para mitigar el cambio climático. Sin embargo, la variabilidad de la radiación solar percibida por los paneles solares es un desafío técnico y operacional para la implementación de esta tecnología. La radiación solar percibida es diferente a la radiación solar teórica debido a la interacción de la luz con la atmósfera. En esta monografía se aborda el problema de diseñar una solución para la predicción de la radiación solar, para diferentes escalas temporales, usando información de 37 estaciones meteorológicas de la comunidad de Castilla y León, en España, para el período 2002-2019.

El resultado de las predicciones será evaluado utilizando métricas de machine learning, verificando si responde a las necesidades de costo y tiempo de respuesta, a la vez que se evalúa la radiación solar por unidad de tiempo para realizar la gestión de posibles proyectos de generación de energía fotovoltaica.

*Palabras clave:* radiación solar, aprendizaje profundo, predicciones, energía renovables, memoria de largo y corto plazo.

<https://github.com/MoisesGuerreroUdeA/Seminario-Analitica-CDatos>

### **Abstract**

Photovoltaic solar energy is one of the main proposals within renewable energies to mitigate climate change. However, the variability of solar radiation perceived by solar panels is a technical and operational challenge in the implementation of this technology. The perceived solar radiation is different from theoretical radiation due to the interaction of light with the atmosphere. This monograph addresses the problem of designing a solution for the prediction of solar radiation, for different time scales, using data from 37 weather stations in the community of Castilla y León, in Spain, for the period 2002-2019.

The result of the predictions will be evaluated using machine learning metrics, verifying if it corresponds to the needs of cost and response time, while evaluating solar radiation per unit of time to carry out the management of possible photovoltaic power generation projects.

*Keywords:* solar radiation, deep learning, forecasting, renewable energy, long short term memory.

## **1. Descripción del problema**

La actual crisis climática es uno de los problemas más complejos e importantes a los que se enfrenta la humanidad. Los efectos del cambio climático están teniendo un fuerte impacto en el planeta y en la vida humana: aumento del nivel del mar, intensificación de los fenómenos meteorológicos extremos (lluvias, sequías, oleadas de calor, huracanes), cambios en los patrones de precipitación, son solo algunos de los ejemplos más nombrados (Jentsch & Beierkuhnlein, 2008).

Uno de los principales motores del cambio climático son los gases de efecto invernadero (GEI) que se emiten de manera más acelerada en la atmósfera desde la era industrial. Una de las fuentes de emisiones de GEI son las centrales térmicas, que utilizan combustibles fósiles para la generación de energía (Stern & Kaufmann, 2013).

En este contexto, las energías renovables, como la energía solar, eólica e hidráulica, son una alternativa que de adoptarse a gran escala podrían contribuir a la mitigación del cambio climático (Panwar et al., 2011). Estas, comparadas con las energías convencionales, son más limpias, renovables y, en muchos casos, más asequibles (Verbruggen et al., 2010).

En particular la transición o mayor implementación de energía renovable a partir de la radiación solar, también llamada energía solar fotovoltaica, supone un gran desafío en aspectos técnicos y operacionales. Uno de ellos está relacionado con la variabilidad de la radiación solar que perciben los paneles solares, puesto que la radiación teórica es diferente a la radiación solar percibida, afectada principalmente por la interacción de la luz con la atmósfera (Sehrawat et al., 2023).

### **Radiación solar teórica**

Según Whiteman & Allwine (1986) la radiación solar extraterrestre es el término que recibe la radiación electromagnética emitida por el sol y que llega a la tierra. La radiación que percibe un punto en la tierra depende de varios factores, incluyendo la latitud, longitud y día del año. Por ejemplo, los puntos más cercanos al ecuador geográfico reciben en promedio más radiación solar

extraterrestre que los puntos más cercanos a los polos, esto debido a la inclinación de la tierra sobre su eje de rotación. La longitud no tiene un efecto directo sobre la cantidad de radiación solar que percibe un punto, sin embargo, se utiliza como referencia para estimar la zona horaria y la franja del día en la que se percibe luz solar.

El día del año también influye en la cantidad de radiación solar que percibe un punto, por lo menos de dos maneras. En primer lugar, la tierra describe una órbita alrededor del sol en forma elíptica, por lo que la distancia entre el sol y la tierra varía a lo largo del año. La tierra está más cerca del sol en el perihelio (el punto de la órbita que es más cercano al sol) y más lejos del sol en el afelio (el punto de la órbita que es más lejano del sol). Durante el perihelio la tierra percibe una mayor radiación solar.

### **Radiación solar percibida**

La atmósfera de la tierra está compuesta por diferentes gases, vapor de agua y aerosoles, como partículas de polvo y aerosoles. Estos componentes interactúan con la radiación solar de diversas formas, haciendo que se modifique su intensidad y distribución (Kumari & Toshniwal, 2021b). La capa de ozono, por ejemplo, es capaz de absorber la radiación ultravioleta, mientras que los aerosoles y el vapor de agua pueden absorber la radiación en el espectro infrarrojo. Así, a medida que la luz solar atraviesa la atmósfera, parte de ella puede ser reflejada, dispersada o absorbida por las moléculas de aire (nitrógeno, oxígeno, helio), el vapor de agua, las nubes, los aerosoles, entre otros (Hassan et al., 2016).

De esta forma, la radiación solar percibida en la superficie de la tierra difiere de la radiación solar teórica por las diferentes interacciones que tiene con la atmósfera. Pronosticar este tipo de radiación solar es supremamente complejo pues el sistema climático depende de numerosos aspectos físicos (como la temperatura, vapor de agua, velocidad de viento, presencia de núcleos de condensación), además de la variabilidad extrema introducida por el cambio climático (Hassan et al., 2016). Uno de los principales moderadores de la radiación solar percibida es la nubosidad. Las nubes reflejan y absorben la radiación solar, lo que puede reducir la cantidad de radiación

que llega a la superficie terrestre. De acuerdo con Cess et al. (1995) la cantidad de radiación solar que se refleja o absorbe por las nubes depende de su tipo, espesor y tamaño.

### **1.1 Problema de negocio**

Como se explicó previamente, estimar la radiación solar percibida en algún punto de la tierra, basados en su latitud, longitud y día del año puede ser una tarea compleja y completamente relevante para la implementación de un proyecto de generación de energía fotovoltaica.

Para cualquier sistema de generación de energía debe existir un correcto balance entre la demanda y la oferta. Las predicciones de estos dos componentes permiten a los integrantes del sistema (generadores, distribuidores, operadores, reguladores, entre otros) establecer las dinámicas del mercado energético, como la compra y venta de energía. Para un generador de energía fotovoltaica es completamente relevante establecer en el corto, mediano y largo plazo su capacidad de generación para ofrecer energía en firme y determinar si puede cumplir con sus obligaciones. Por tanto, grandes desviaciones en las predicciones tienen un alto costo en los actores del sistema energético (Moreno-Munoz et al., 2008; Paredes et al., 2017).

Por tal motivo, es supremamente deseable poder realizar una predicción de la radiación solar, en diferentes escalas temporales, sobre todo si se tiene en cuenta que la implementación de proyectos fotovoltaicos ha ido en aumento en diferentes lugares del planeta. Estas predicciones tienen amplias aplicaciones: determinar los mejores lugares de instalación de parques solares, estimar el balance entre oferta y demanda, establecer la franja intra horaria para el despacho de energía al operador del sistema de energía, administrar el almacenamiento de energía, planear los ciclos de mantenimiento y de las transacciones en el mercado energético (Hammer et al., 1999).

La predicción de radiación solar percibida para un punto de la Tierra basado en su latitud, longitud y día del año se puede realizar utilizando modelos físicos o estadísticos. Los modelos físicos utilizan leyes físicas para calcular la cantidad de radiación solar que llega a la superficie de la tierra, combinando la radiación solar teórica y las estimaciones de las interacciones con la

atmósfera. Los modelos estadísticos utilizan datos históricos de radiación solar para predecir la radiación solar futura.

En términos generales, los modelos físicos se consideran más precisos que los modelos estadísticos, pero son más complejos y requieren de muchos más datos. En contraposición, los modelos estadísticos son más simples y requieren una menor cantidad de datos, pero pueden resultar menos precisos que los modelos físicos (Gautier et al., 1980). Además, existen dificultades intrínsecas e instrumentales para la medición y predicción de la radiación solar. En primer lugar, la radiación solar es una onda electromagnética que se propaga en todas las direcciones, por lo que dificulta su medición en una superficie pequeña, como la de un piranómetro, el instrumento usado para medir la radiación solar. En segundo lugar, los instrumentos de medida de las variables climatológicas tienen limitaciones y pueden estar sujetos a errores de calibración y a la degradación de los componentes con el tiempo.

Encontrar una solución para la predicción de la radiación solar a partir de los datos medidos en estaciones meteorológicas es una tarea obligatoria después de la fase de factibilidad de un proyecto fotovoltaico. En esta monografía se aborda este problema, buscando una alternativa para la toma de decisiones respecto a la implementación y operación de un sistema de generación de energía fotovoltaica, partiendo de información disponible de estaciones meteorológicas que permitan generalizar la solución y que pueda ser ajustada para cualquier proyecto específico.

### **1.1. Aproximación desde la analítica de datos**

Según lo expuesto en la sección anterior, existen dos formas de determinar la radiación solar percibida, bien sea a través de cálculos físicos o mediante métodos estadísticos. La aproximación que se aborda en este trabajo está relacionada con la analítica de datos, mediante métodos estadísticos y machine learning para predecir la radiación solar en diferentes escalas temporales.

Para ello se realiza un análisis estadístico con el fin de identificar dependencias entre predictores usados (variables meteorológicas). El tratamiento de los datos y el abordaje de los modelos será

híbrido, combinando modelos de aprendizaje supervisado para predecir la radiación solar a partir de otras variables meteorológicas, incluida la radiación solar teórica, y modelos diseñados para el análisis de series de tiempo, en los que los datos son dependientes de los valores inmediatamente anteriores y los muestreos no son aleatorios.

Se seguirá un enfoque de aprendizaje conjunto presentado en Sehrawat et al. (2023), integrando predicciones de los modelos desarrollados, permitiendo combinar las fortalezas de diferentes enfoques de modelado para mejorar la precisión de las predicciones.

### **Factores que influyen en la predicción**

En la actualidad, la predicción de la radiación solar depende del horizonte de tiempo de la predicción. Este horizonte hace referencia al período de tiempo en el futuro en el que la predicción es requerida o el paso de tiempo entre el momento actual y el momento para el cual se realiza la predicción. Dentro de la literatura, este horizonte de tiempo se suele dividir en tres escalas: corto, mediano y largo plazo (Kumari & Toshniwal, 2021b).

La escala de corto plazo, generalmente entre 30 minutos a 72 horas, tiene una alta aplicabilidad en el mercado eléctrico: balance entre la oferta y la demanda, decisiones de despacho de energía, planeación de las transacciones en el mercado energético, por ejemplo (Ibrahim & Khatib, 2017).

La escala de mediano plazo, usualmente de unos días, semanas e incluso meses, es útil para la planificación de mantenimientos (Olatomiwa et al., 2015). Mientras que la escala de largo plazo es usada para la planificación de proyectos exitosos de generación de energía fotovoltaica. Esta escala va desde varios meses hasta algunos años (Mishra et al., 2008). Usualmente, entre más alta sea la escala habrá una menor precisión en el pronóstico, donde es más difícil capturar la variabilidad climática asociada a la radiación percibida.

Otro factor determinante para la predicción de la radiación solar es el clima del lugar en el cual se realiza la predicción, tal como lo indican Nann & Riordan (1991). Como ya se mencionó, las

condiciones meteorológicas como la presión atmosférica, temperatura, humedad relativa, cantidad de aerosoles en la atmósfera, velocidad del viento y nubosidad, modifican la cantidad de radiación solar percibida. Según la literatura existente, la clasificación del clima del sitio de predicción es uno de los factores que más afecta la mejora de los modelos, por lo que su inclusión es necesaria para una adecuada estimación de la radiación solar percibida.

### **Métodos de aprendizaje profundo para la predicción de energía solar**

Los métodos más usados y altamente recomendados para el pronóstico de la radiación solar son LSTM (*long short-term memory*), redes de creencia profunda (*deep belief network, DBN*), redes neuronales convolucionales (*convolutional neural network, CNN*), redes de estado de eco (*echo state network, ESN*), redes neuronales recurrentes (*recurrent neural network, RNN*), unidades recurrentes cerradas (*gated recurrent network, GRU*) y sus híbridos. (Kumari & Toshniwal, 2021b) explica de manera concisa la diferencia entre estos métodos.

En la literatura se encuentran numerosas referencias a modelos realizados utilizando estos métodos, que son usados normalmente de forma híbrida, en conjunto con otras técnicas, para optimizarlos y comparar su desempeño (Kumari & Toshniwal, 2021b). Usualmente, su evaluación se realiza dividiendo la predicción para las diferentes estaciones del año, primavera, verano, otoño e invierno, cuando hacen parte de la caracterización del lugar de estudio (Kumari & Toshniwal, 2021a). A continuación, se indican brevemente algunos resultados del desempeño y comparaciones de diferentes autores que han realizado aportes en la predicción de la radiación solar.

Los modelos de LTSM son uno de los más usados en este campo. Srivastava y Lessmann (2018) utilizaron esta técnica para predecir la radiación diaria de varias estaciones en Europa y Estados Unidos. Su modelo mejoró la predicción en un 52% respecto a los modelos de persistencia inteligente, un modelo físico que utiliza datos históricos para calcular el índice de cielo despejado y estimar la radiación percibida (Liu et al., 2021). Resultados similares a los obtenidos por Yu et al. (2019) que les permitió predecir la radiación solar horaria para diferentes lugares de Estados



Unidos con un RMSE de 41,37 W/m<sup>2</sup>. Algunos autores han agregado mecanismos adicionales a los modelos LSTM para mejorarlos. Brahma y Wadhvani (2020), por ejemplo utilizaron mecanismos de cerrado (*gating mechanism*) y celdas de memoria para desarrollar modelos a los que llamaron *BiLSTM* y *attention-based LSTM* para hacerlos más eficientes en el aprendizaje a largo plazo de las relaciones entre las variables de entrada, demostrando una alta precisión en la predicción. Estos modelos de LSTM son computacionalmente más demandantes, pero tienen muy buenos resultados.

Menos populares son los modelos GRU. Wojtkiewicz et al (2019), desarrollaron un modelo basado en GRU en que integraron las técnicas de puerta de entrada y de olvido (*forget and input gates*) que hace más robusta la predicción. Como resultado obtuvieron un modelo que es computacionalmente menos demandante porque entrenan menos parámetros y usan una menor cantidad de memoria, por lo que ejecutan más rápido.

Usando CNN también se han realizado predicciones de radiación solar con un enfoque en análisis de imágenes. De esta manera se utilizan imágenes en tierra de la cobertura de nubes para estimar características espacio temporales. Así, se logró un modelo que mejoró la predicción en un 17.06% respecto a modelos de persistencia ( Zhao et al., 2019). Los modelos CNN tienen como ventaja que pueden ser usados tanto si se disponen datos basados en imágenes o si disponen datos dispuestos de manera bidimensional.

Los modelos DBN, que tienen como estructura base las máquinas de Boltzman restringidas, un tipo de redes neuronales artificiales estocásticas que pueden aprender la distribución de probabilidad de su conjunto de datos de entrada. De esta manera, los modelos DBN pueden usarse para realizar predicciones de radiación solar si las características típicas en los datos no son fácilmente detectables (Zang, Cheng, et al., 2020).

Otro grupo importante de modelos son los RNN en los que su principal característica radica en que las neuronas tienen métodos de retroalimentación y conexión hacia adelante (*feed forward*) y hacia atrás (*feed back*), simulando funciones de memoria lo que las hace ideales para el

procesamiento de series de tiempo de radiación solar. Usando este enfoque, Mishra y Palanisamy (2018) propusieron un modelo predictivo para diferentes horizontes de tiempo, en el cual obtuvieron un RMSE de 18,57 W/m<sup>2</sup>.

En un esfuerzo adicional para mejorar la predicción de la radiación solar, numerosos investigadores han desarrollado modelos híbridos que integran diferentes modelos de aprendizaje profundo. Por ejemplo se han realizado modelos híbridos que integran métodos CNN y LSTM (Zang, Liu, et al., 2020), algoritmos genéticos para optimizar redes neuronales profundas: GRU, LSTM y RNN (Bendali et al., 2020). Los resultados obtenidos con modelos híbridos sugieren que ensamblar múltiples modelos juntos mejoran la predicción de la radiación solar. Las ventajas y desventajas de cada uno de los modelos son los que se relacionan en la Tabla 1 (Kumari & Toshniwal, 2021b).

Modelo	Ventaja	Desventaja	Escenarios de aplicación
LSTM	Capaz de manejar relaciones de dependencia de largo plazo en series de tiempo	Alto costo computacional	Existen registros de series de tiempo de radiación solar
GRU	Diseño poco complejo, menor consumo de memoria, rápida ejecución, menor costo computacional	lenta convergencia y baja eficiencia de aprendizaje	Los recursos computacionales son limitados
CNN	Es capaz de procesar datasets de imágenes y puede extraer características espaciales	Alto costo computacional, las características deben ser predeterminadas	El dataset de radiación solar tiene imágenes (nubosidad) o los datos pueden ser convertidos a arreglos bidimensionales
DBN	Es capaz de extraer características no supervisadas y es menor en costo computacional	Incapaz de procesar datos meteorológicos multidimensionales	No se consideran múltiples variables.
RNN	Es capaz de procesar series de tiempo y es eficiente	Incapaz de determinar características	Existen registros de series de tiempo

---

	computacionalmente	eficientemente	de radiación solar
Híbrido	Capaces de extraer diferentes tipos de características de los datos y son altamente precisos	Son costosos computacionalmente	Los datos de radiación contienen diferentes características, tales como información espacial y temporal

## 1.2. Origen de los datos

Los datos usados hacen parte de un repositorio abierto con licencia *Creative Commons Attribution*, que contiene mediciones de radiación horizontal global solar (GHI), para el período 2002-2019, con frecuencia de 30 minutos, para 37 estaciones de Castilla y León, en España.

Los datos fueron publicados por integrantes del Departamento de Ingeniería Topográfica y Cartográfica de la Escuela Técnica Superior de Ingenieros en Topografía, Geodesia y Cartografía, Universidad Politécnica de Madrid, y se encuentran disponibles para su uso con fines de investigación en un repositorio público. Los datos contienen información de marca de tiempo, datos meteorológicos (precipitación, temperatura, humedad relativa, radiación solar, velocidad y dirección del viento) y geoespaciales (latitud y longitud)

## 1.3. Métricas de desempeño

### Métricas de machine learning (desempeño de los modelos)

Las métricas que se usarán para evaluar los modelos de machine learning son las estándar, según lo refiere la literatura (Kumari & Toshniwal, 2021b), a saber:

*Error absoluto medio (MAE)*: se calcula como el promedio de la diferencia absoluta entre el valor real y el valor predicho. Esta métrica es útil para penalizar el modelo tanto si realiza sobreestimaciones como subestimaciones. Matemáticamente, el MAE se calcula como:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|$$

donde  $y_i$  es el valor real en la  $i$ ésima observación,  $\hat{y}_i$  es el valor predicho y  $N$  es el número total de observaciones.

El MAE tiene un valor entre 0 y infinito, donde 0 indica que las predicciones del modelo son perfectas e infinito indica que las predicciones del modelo son completamente inexactas. Es decir, un valor bajo del MAE indica que el modelo es preciso y un valor alto del MAE indica que el modelo es impreciso.

*Error cuadrático medio (MSE)*: se calcula como el promedio del cuadrado de las diferencias entre el valor predicho y el valor real. Esta métrica penaliza los casos en los que las diferencias entre ambos valores son mayores. Matemáticamente, el MSE se calcula como:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2$$

Un valor bajo del MSE indica que el modelo es preciso y un valor alto del MSE indica que el modelo es impreciso.

El MSE es una métrica de error que es sensible a los valores extremos. Esto significa que los valores extremos pueden afectar significativamente el valor del MSE.

*Raíz del error cuadrático medio (RMSE)*: se calcula como la raíz cuadrada del promedio del cuadrado de las diferencias entre el valor de radiación predicho y el valor real. Esta métrica suele ser una de las más utilizadas para evaluar el desempeño de un modelo. También ayuda a identificar y eliminar outliers en los datos. Matemáticamente, el RMSE se calcula como la raíz del MSE.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} = \sqrt{MSE}$$

Un valor bajo de la RMSE indica que el modelo es preciso, mientras que un valor alto indica que el modelo es impreciso.

La RMSE es una métrica de error que es menos sensible a los valores extremos que el MSE en vista de que la RMSE se basa en la raíz cuadrada de los errores cuadráticos, lo que suaviza los efectos de los valores extremos.

*Coefficiente de determinación ( $R^2$ ):* el  $R^2$  es una métrica que se puede utilizar para comparar modelos con diferentes números de variables. Es una medida de la bondad de ajuste de un modelo de regresión lineal, muy importante para evaluar el rendimiento de un modelo, ya que indica qué proporción de la variabilidad de la variable dependiente se explica por el modelo.

Sin embargo, el  $R^2$  puede ser engañoso si el modelo tiene muchas variables. Esto se debe a que el  $R^2$  puede aumentar simplemente agregando variables al modelo, incluso si esas variables no son significativas.

Matemáticamente, el coeficiente de determinación se calcula como:

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

donde  $\bar{y}$  es la media de la variable  $y$ .

*Error porcentual medio (MAPE):* es una métrica de error que se utiliza para evaluar la precisión de las predicciones de un modelo. Se calcula como la media del error absoluto, expresado como porcentaje de los valores reales. Matemáticamente, el MAPE se calcula como:

$$MAPE = 100 * \frac{\sum_{i=1}^N |\hat{y}_i - y_i|}{\sum_{i=1}^N |y_i|}$$

Ejemplificando, si el MAPE de un modelo es 10%, entonces el modelo comete un error del 10% en promedio. Esto significa que las predicciones del modelo están alejadas de los valores reales en un promedio de 10%. Un MAPE bajo indica que el modelo es preciso y uno alto indica que el modelo es impreciso.

El MAPE es una métrica útil para evaluar el rendimiento de un modelo en una variedad de problemas. Sin embargo, es importante tener en cuenta que el MAPE puede verse afectado por la escala de los datos. Por ejemplo, si los datos están en una escala pequeña, entonces un MAPE de 10% puede ser significativo. Sin embargo, si los datos están en una escala grande, entonces un MAPE de 10% puede ser insignificante.

En general, el MAPE es una métrica útil para evaluar el rendimiento de un modelo. Sin embargo, es importante interpretar el MAPE con precaución, teniendo en cuenta la escala de los datos.

*Error cuadrático medio porcentual (RMSPE)*: es una métrica de error que se utiliza para evaluar la precisión de las predicciones de un modelo. Se calcula como la raíz cuadrada de la media del error cuadrático, expresado como porcentaje de los valores reales. Matemáticamente, el RMSPE se calcula como:

$$RMSPE = 100 * \sqrt{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N |y_i|}}$$

Si, por ejemplo, el RMSPE de un modelo es 10%, entonces el modelo comete un error del 10% en promedio, expresado en términos de la raíz cuadrada del error cuadrático. Esto significa que las predicciones del modelo están alejadas de los valores reales en un promedio de  $\sqrt{10} = 3.16\%$ .

Un RMSPE bajo indica que el modelo es preciso. Un RMSPE alto indica que el modelo es impreciso.

El RMSPE es una métrica similar al MAPE, pero tiene algunas ventajas. En primer lugar, el RMSPE es menos sensible a la escala de los datos que el MAPE. Esto significa que el RMSPE es una métrica más comparable para conjuntos de datos con diferentes escalas. En segundo lugar, el RMSPE es una métrica más robusta a los outliers que el MAPE. Esto significa que el RMSPE es menos probable que se vea afectado por valores atípicos en los datos.

*Raíz del error cuadrático medio de la predicción:* la raíz del error cuadrático medio de la predicción (RMSEP) es una medida de la precisión de las predicciones de un modelo. Se calcula como la raíz cuadrada de la media de los errores cuadráticos de las predicciones. Matemáticamente, la RMSEP se puede calcular como:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{2}}$$

Un valor bajo de la RMSEP indica que el modelo es preciso. Un valor alto de la RMSEP indica que el modelo es impreciso.

La RMSEP es una métrica de error que es menos sensible a los valores extremos que otras métricas de error, como el error cuadrático medio (RMSE). Esto se debe a que la RMSEP se basa en la raíz cuadrada de los errores cuadráticos, lo que suaviza los efectos de los valores extremos.

### **Métricas de negocio**

En vista de que el objetivo del presente trabajo está relacionado con la predicción de la radiación solar percibida con el fin de ser usada para la evaluación, implementación u operación de un sistema de generación de energía fotovoltaico, las métricas de negocio están orientadas a los agentes responsables de dicho sistema. Estas métricas son:

*Tiempo de respuesta:* el tiempo de respuesta se refiere al tiempo que tarda la solución propuesta para realizar la predicción de la energía solar percibida. Es importante que el tiempo de respuesta sea lo suficientemente rápido para que la predicción pueda utilizarse en tiempo real. Por ejemplo,

si se pretende predecir la radiación solar en escala intra horaria para un período de un día, no tiene sentido que la solución tarde horas en realizar la predicción.

*Costo:* el costo se refiere a los recursos necesarios para generar una predicción. Incluye el costo de los datos, el costo del hardware y el costo del software. El responsable del negocio deberá determinar si el costo de la predicción, sumado a los demás costos del proyecto, viabilizan el caso de negocio.

*Radiación solar percibida por unidad de tiempo:* esta métrica es importante para los negocios que utilizan la predicción de la energía solar para tomar decisiones. Por ejemplo, un operador de una planta de energía solar fotovoltaica puede utilizar la predicción de la energía solar para planificar la producción y el almacenamiento de energía. Un desarrollador de proyectos de energía solar puede utilizar la predicción de la energía solar para evaluar el potencial de un proyecto. La radiación solar se mide en vatios por metro cuadrado ( $\text{W/m}^2$ ). Esta unidad de medida representa la cantidad de energía solar que llega a un área de un metro cuadrado. Se dispondrá la radiación solar percibida por unidad de tiempo como la radiación percibida en dicho período, por ejemplo la radiación solar por hora o por día.

### **Métricas de aplicación**

Según Voyant et al. (2017), en una revisión que realizaron de aplicación de diferentes modelos para la predicción de radiación solar en varios lugares del planeta, la métrica de RMSE promedio fue de 13% para los modelos con predicciones de un horizonte de 1 hora. Por lo tanto, una estimación razonable para este proyecto será realizar predicciones con un horizonte de tiempo de 1 hora, un RMSE del 15% para el modelo con mejor desempeño, el cual deberá poder realizar la estimación (tiempo de respuesta) en un tiempo inferior al horizonte seleccionado (1 hora) y que en el que el costo de implementación, asociado a pago de infraestructura, sea cero, es decir, que pueda implementarse con los recursos disponibles como estudiantes, sin ningún tipo de pago adicional.



## **2. Objetivos**

### **2.1.Objetivo general**

Desarrollar una solución para la predicción de radiación solar en diferentes escalas temporales para diferentes estaciones meteorológicas de la Comunidad de Castilla y León, España, que permita la toma de decisiones alrededor de proyectos de generación fotovoltaica:

### **2.2. Objetivos específicos**

- Seleccionar y analizar datos de variables meteorológicas, incluida radiación solar, de diferentes estaciones de la Comunidad de Castilla y León, España, para detectar patrones, tendencias y relaciones entre las variables a través de un análisis estadístico.
- Procesar y transformar los datos relevantes para realizar una predicción de la radiación solar de cada una de las estaciones meteorológicas seleccionadas.
- Desarrollar modelos de aprendizaje automático que permitan predicción de la radiación solar a partir de los *datasets* generados para las diferentes estaciones meteorológicas.
- Evaluar y comparar el desempeño de los diferentes modelos implementados para determinar si son una solución viable al problema de negocio.

### 3. Datos

#### 3.1. Datos originales

Los datos usados hacen parte de un repositorio abierto con licencia *Creative Commons Attribution*, que contiene mediciones de radiación horizontal global solar (GHI), recopiladas durante un período de 18 años, desde el enero de 2002 hasta diciembre de 2019, con una frecuencia de 30 minutos, que fueron medidos en 37 estaciones ubicadas en la región española de Castilla y León. El conjunto de datos no solo cuenta con datos crudos, sino también con datos refinados y etiquetados por estaciones de medición, contando además con variables meteorológicas y geográficas que complementan el valor de los datos de radiación.

Los datos fueron publicados por integrantes del Departamento de Ingeniería Topográfica y Cartográfica de la Escuela Técnica Superior de Ingenieros en Topografía, Geodesia y Cartografía, Universidad Politécnica de Madrid, y se encuentran disponibles para su uso con fines de investigación en un repositorio público con nombre “CyL\_GHI” disponible en el sitio web Zenodo (<https://zenodo.org/records/7404167>), y además cuentan con una documentación detallada (Cesar et al., 2022), que describe de manera precisa el preprocesamiento, estandarización y limpieza realizado por los autores.

Los archivos se encuentran almacenados en formatos *csv* y *zip* y se encuentran ordenados de acuerdo a lo descrito en la siguiente tabla:

Nombre del Archivo	Descripción	Datos	Tamaño
CyL_raw.zip	Datos sin procesar descargados, sin aplicar operaciones de refinamiento	18 carpetas (nombradas con el número de año). Cada carpeta contiene los datos en su formato sin procesar guardados a nivel diario.	232.6 MB
CyL_GHI_ast.csv	Datos GHI	Datos de todas las	281.9 MB

	combinados con variables astronómicas.	estaciones combinados en un único archivo csv, incluyendo variables astronómicas como la elevación solar, azimut solar, entre otras.	
CyL_meteo.csv	Datos meteorológicos para el periodo de tiempo considerado.	Datos de todas las estaciones combinados en un único archivo csv, incluyendo variables meteorológicas de cada estación de medición.	313.2 MB
CyL_geo.csv	Datos geográficos para localizar a las 37 estaciones	Un único archivo csv con la ubicación geográfica de las 37 estaciones.	2.4 kB
CyL_by_stations.zip	Datos refinados separados por estaciones en archivos csv individuales.	37 archivos csv, uno para cada estación meteorológica, nombrados con el código de estación correspondiente.	263.2 MB

Dado que se cuenta tanto con datos crudos, como con datos adecuados y transformados previamente por los autores del dataset, el objetivo es realizar un proceso de extracción y transformación propio, a partir de los datos “raw” relacionados, que posteriormente podrán ser comparados y verificados con los datos ya preparados por los autores. Por tanto se hace uso del dataset CyL\_raw.zip en conjunto con el archivo CyL\_geo.csv para complementar la información con datos geográficos de las estaciones meteorológicas.

La información del archivo **CyL\_raw.zip** se encuentra almacenada en subcarpetas referentes a cada año de mediciones realizadas entre el 2002 y el 2019. Cada subdirectorío, cuenta con una serie de archivos en formato *zip* cuyos nombres siguen un patrón con la estructura

**YYYYMMDD\_RedClimaITACYL\_Horario.zip**, donde **YYYYMMDD** corresponde al día de mediciones para el año de cada subcarpeta.

Cada documento comprimido cuenta con un archivo en formato *csv* que mantiene el mismo patrón de nombre **YYYYMMDD\_RedClimaITACYL\_Horario.csv**, y cuentan con un encabezado con 10 columnas y usan como separador punto y coma (;), las columnas incluidas en todos los *csv* corresponden a:

- **Código:** Código único de la estación en la cual se realizaron las mediciones.
- **Ubicación:** Indica el nombre de la estación de medición de acuerdo a su ubicación.
- **Fecha (AAAA-MM-DD):** Hace referencia a la fecha en la cual se realizaron las mediciones (UTC).
- **Hora (HHMM):** Indica la hora en la cual se realizó la medición para la fecha correspondiente (UTC).
- **Precipitación (mm):** Precipitaciones medidas en milímetros en la ubicación de la estación de medición.
- **Temperatura (°C):** Temperatura en grados celsius medida en la ubicación de la estación meteorológica.
- **Humedad relativa (%):** Indica el porcentaje de humedad medido.
- **Radiación (W/m<sup>2</sup>):** Medición de la radiación solar global horizontal (GHI).
- **Vel. viento (m/s):** Velocidad del viento medida en *m/s*.
- **Dir. viento (°):** Dirección del viento en grados.

Por su parte el archivo **CyL\_geo.csv** contiene información sobre latitud, longitud y altitud de cada estación meteorológica, así como su código y nombre. El archivo cuenta con un encabezado de 5 columnas separados por comas (,) que hace referencia a:

- **station\_code:** Código único por estación
- **name:** Nombre de la estación meteorológica
- **latitude:** Latitud geográfica
- **longitude:** Longitud geográfica
- **height:** Altitud en metros sobre el nivel del mar

El acceso a los datos y la descarga de los mismos se puede realizar directamente desde el sitio web de Zenodo como CyL\_GHI, accediendo a los enlaces de descarga directa.

### **3.2. Datasets**

#### ***3.2.1. Descompresión y extracción de los archivos***

Con el objetivo de realizar la construcción de los datasets estandarizados y preparados, se realiza un proceso de extracción, transformación y carga por medio de la herramienta Google Colaboratory. Para lo cual se realizó una descompresión del archivo de datos crudos **CyL\_raw.zip**, el cual contiene un conjunto de subcarpetas correspondientes a los años del 2002 hasta el 2019.

Dado que en cada subdirectorio, la información se encuentra igualmente comprimida en archivos zip, fue necesario realizar un proceso de descompresión automática por medio de una función de Python, ejecutando comandos de sistema operativo sobre cada uno de los archivos de forma iterativa. El proceso consistió en:

1. Realizar un listado de los archivos zip disponibles en cada uno de los sub directorios por año.
2. Crear un directorio temporal, con subdirectorios asociados igualmente con cada año de mediciones, dentro del entorno de Colaboratory.
3. Para cada elemento del listado de archivos obtenidos, realizar la descompresión del archivo por medio de comandos de sistema operativo, indicando como salida el subdirectorio correspondiente por año, dentro del directorio temporal antes mencionado.

#### ***3.2.2. Preprocesamiento y transformación de los datos***

Posteriormente, se realizó una inspección de los archivos extraídos, identificando que estos presentaban una codificación “latin-1”, por lo cual se debía realizar la lectura de los mismos teniendo en cuenta dicha codificación.

Para la lectura automática de todo el conjunto de datos, se desarrolló un script que permite realizar una lectura de cada uno de los archivos extraídos anteriormente de forma automática, y que a su vez estos pudieran ser concatenados a un mismo arreglo de datos, haciendo uso además de una función desarrollada para el preprocesamiento de los datos, que permite renombrar las columnas de acuerdo a las transformaciones indicadas a continuación:

- Código → station\_code
- Ubicación → station\_name
- Fecha (AAAA-MM-DD) → date
- Hora (HHMM) → time
- Precipitación (mm) → precipitation
- Temperatura (°C) → air\_temp
- Humedad relativa (%) → humidity
- Radiación (W/m<sup>2</sup>) → GHI
- Vel. viento (m/s) → wind\_sp
- Dir. viento (°) → wind\_dir

A partir de estas columnas renombradas, se hace uso de las columnas “date” y “time” para la generación de una columna de marca de tiempo (“timestamp”) con un formato YYYY-MM-DD HH:MM:SS.

### ***3.2.3. Distribución de datos por estaciones meteorológicas y adición de radiación teórica***

Posteriormente, se usa el listado de códigos por estación meteorológica disponible en el archivo **CyL\_geo.csv**, para iterativamente filtrar los datos del arreglo general creado en el paso anterior y guardar únicamente los datos asociados con una estación en particular por archivo, siguiendo los pasos:

1. Se crea un nuevo directorio temporal donde se almacenarán 37 archivos en formato csv (para cada una de las estaciones meteorológicas).

2. Se realiza el filtrado de los datos del arreglo general por cada estación meteorológica y se guardan en formato *csv* en el directorio anteriormente creado, ordenados de forma ascendente por la columna “timestamp”.

De acuerdo a lo comentado en (Bautista Carrascosa, 2016), es posible obtener un valor de radiación solar extraterrestre teórica para una ubicación específica en la tierra, solo por medio de la declinación solar  $\delta$  y la latitud del lugar  $\phi$ , así como las horas teóricas de sol  $N_{dia}$ , de acuerdo a la ecuación,

$$S_{dia} = S_0 \cdot 3600 \cdot (\sin(90 - \phi + \delta)) \cdot (2N_{dia}/\pi) J/m^2 día$$

Estos valores pueden ser estimados a partir de la latitud, longitud y longitud del huso horario específicos del lugar y constituyen la radiación solar teórica para un punto en particular. Por tanto, se hizo uso de las latitudes y longitudes específicas de cada estación meteorológica definidas en el archivo de datos geográficos, adicionando una columna a los datos con nombre “**theoretical\_radiation**”.

Finalmente, se realiza la escritura de los datos en un formato *parquet* con el fin de mantener la estructura de los tipos de datos usados para cada estación meteorológica.

### 3.3. Analítica descriptiva

Se realizó una comparación entre la radiación solar medida y la radiación solar teórica, evidenciando diferencias principalmente atribuibles a condiciones meteorológicas de la zona. Se evidencia particularmente que la radiación solar teórica es más alta que la radiación solar medida en las franjas entre las 12 y las 14 horas.

Se realiza la simulación de la radiación solar teórica esperada a lo largo del año por horas del día durante el año 2002.

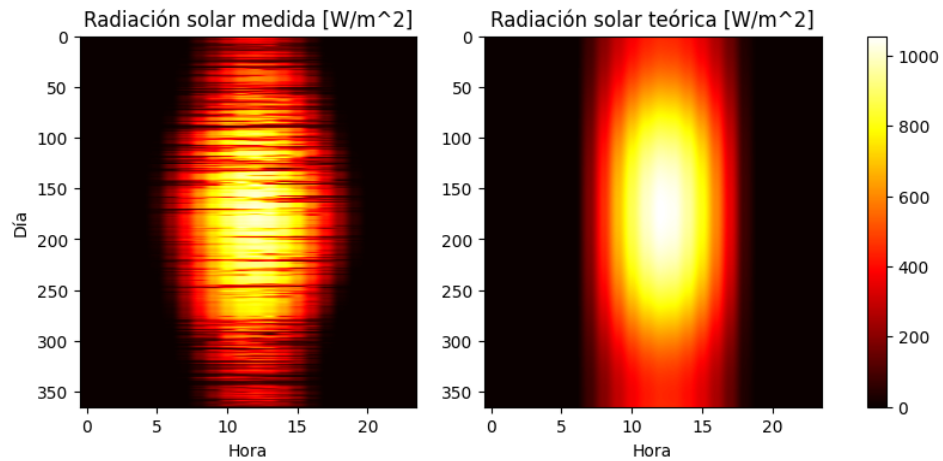


Figura 1. Radiación solar medida y radiación solar extraterrestre teórica

El siguiente gráfico muestra el comportamiento de la radiación solar en W/m<sup>2</sup> a lo largo de los 20 primeros días de enero de 2002 para la estación de medición Nava de Arévalo (AV01), donde se observa la reducción de la irradiación solar en la ubicación de la estación debido a factores meteorológicos adicionales, específicos de la zona, con respecto a la radiación solar teórica esperada a nivel de la atmósfera superior.

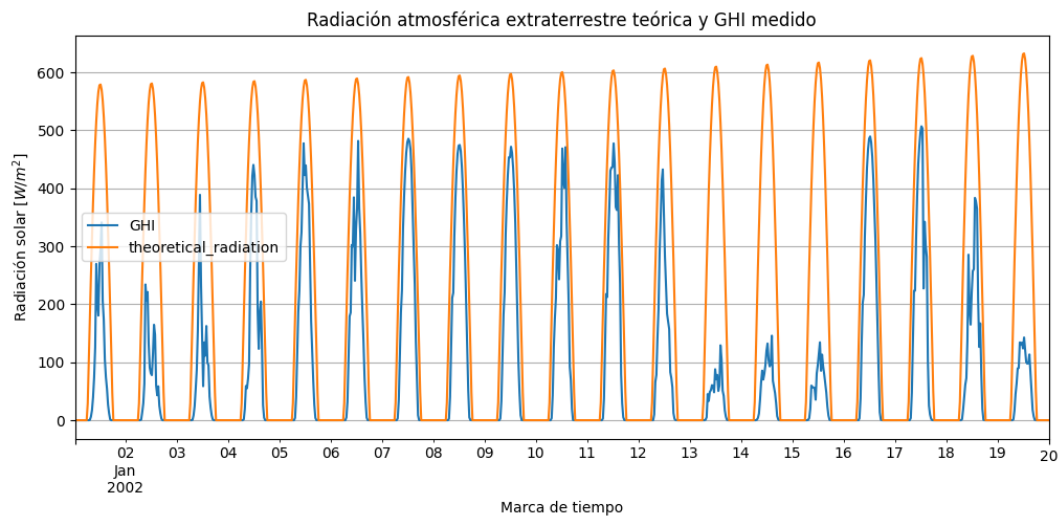


Figura 2. Comportamiento de la radiación solar en W/m<sup>2</sup> a lo largo de 20 días.



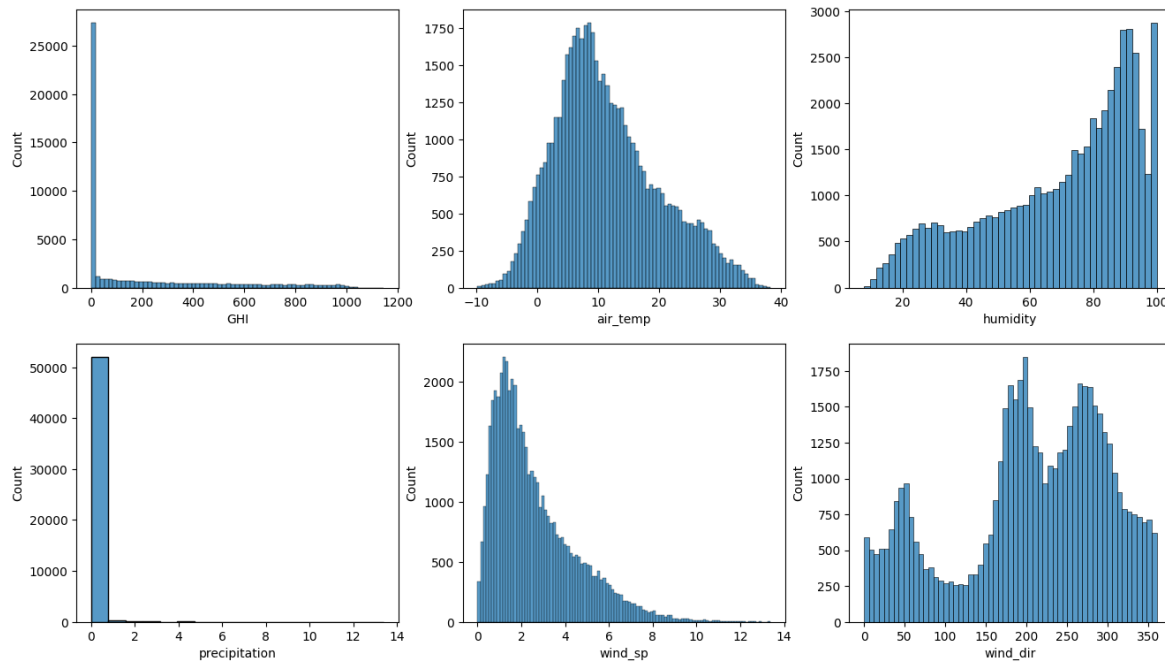


Figura 3. Distribución de variables numéricas para la estación AV01

	GHI	air_temp	humidity	precipitation	wind_sp	wind_dir	theoretical_radiation
count	52603.000000	52603.000000	52498.000000	52603.000000	52603.000000	52603.000000	52603.000000
mean	191.662101	11.444678	69.393970	0.023090	2.585075	206.715517	316.997457
std	280.611273	8.501992	23.674762	0.201364	1.882048	92.790300	408.266442
min	0.000000	-9.770000	7.920000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	5.360000	52.620000	0.000000	1.180000	162.750000	0.000000
50%	6.770000	10.010000	76.100000	0.000000	2.050000	217.200000	0.000000
75%	323.100000	16.620000	89.100000	0.000000	3.570000	278.300000	595.430417
max	1147.000000	38.240000	100.000000	13.400000	13.370000	360.000000	1347.521090

En cuanto a la distribución de los datos por variable se observa que:

- Tal como es de esperarse en cuanto a la radiación solar (GHI), la mayor parte de los datos son valores cercanos a 0 que corresponden a los espacios de tiempo de horario nocturno.
- En cuanto a la temperatura del aire, se observa que los valores varían en su mayoría en un rango entre 5.3 y 16.6°C teniendo momentos en los que alcanza temperaturas superiores a 30°C con un máximo de 38°C medido.
- Se observa cierta tendencia a una humedad relativa alta con una mayor cantidad de mediciones cercanas al 100%

- La dirección del viento presenta muchas variaciones de su ángulo de incidencia con una mayor tendencia a aproximarse con ángulos entre 200 y 300 grados.
- La velocidad media aproximada del viento es cercana 2.5 m/s.