

# Predicción de la generación energética basada en variables meteorológicas. Parte I: Análisis Descriptivo\*

## Prediction of energy generation based on meteorological variables. Part I: Descriptive Analysis

Recibido: enero 24 de 2026 - Evaluado: marzo 15 de 2026 - Aceptado: abril 12 de 2026

Jorge Miguel Meriño-Muriele \*\*  
Francisco Burgos-Flórez \*\*\*  
Juan Pablo Hoyos-Sánchez \*\*\*\*

### Para citar este artículo / To cite this Article

Jorge Miguel Meriño-Muriele, Francisco Burgos-Flórez, Juan Pablo Hoyos-Sánchez, “Predicción de la generación energética basada en variables meteorológicas. Parte I: Análisis Descriptivo” Revista de Ingenierías Interfaces, vol. 9, no.2, pp.1-20, 2026.

### Resumen

La generación de energía solar fotovoltaica depende directamente de variables meteorológicas como la radiación solar, la temperatura, la humedad relativa y la velocidad del viento, las cuales presentan una alta variabilidad temporal. En este contexto, el presente estudio correspondiente a la Parte I del proyecto SOLAB que tiene como objetivo realizar un análisis descriptivo y exploratorio de las principales variables climáticas asociadas a la generación de energía solar en el municipio de La Paz, Cesar (Colombia). El propósito es caracterizar su comportamiento estadístico y sus relaciones internas, con el fin de establecer una base sólida para el desarrollo de modelos predictivos en fases posteriores. Para ello, se aplicaron técnicas de estadística descriptiva, análisis de correlación y visualizaciones gráficas con curvas de ajuste LOESS y polinómicas, que permitieron identificar tendencias lineales y no lineales en los datos. Los resultados mostraron que la radiación solar es la variable con mayor influencia directa sobre la energía generada, mientras que la temperatura se mantiene en un rango estable entre 24 °C y 36 °C. La prueba de Shapiro–Wilk confirmó que las variables no presentan distribución normal. En conjunto, este análisis proporciona una comprensión integral del comportamiento meteorológico local y constituye un insumo esencial para la construcción de modelos predictivos más precisos en la Parte II del proyecto SOLAB.

**Palabras clave:** Energía solar fotovoltaica; análisis descriptivo; variables meteorológicas; radiación solar; correlación estadística; series de tiempo; prueba de normalidad.

---

\*Artículo inédito: “Predicción de la generación energética basada en variables meteorológicas. Parte I: Análisis Descriptivo”.

\*\* Ingeniería Mecatrónica, Correo electrónico: jmerinom@unal.edu.co, Universidad Nacional de Colombia, La Paz, ORCID:0009-0004-2639-2847, Cesar, Colombia.

\*\*\*Doctorado en Ingeniería Mecánica, Ingeniería Mecatrónica, Correo electrónico: fjburos@unal.edu.co, Universidad Nacional de Colombia, La Paz, ORCID:0000-0002-5381-1398, Cesar, Colombia.

\*\*\*\*Doctor en Ciencias de la Electrónica, Ingeniero en Electrónica y Telecomunicaciones, Correo electrónico: jhoyoss@unal.edu.co, Universidad Nacional de Colombia, La Paz, ORCID: 0002-1825-0097, Cesar, Colombia.

## Abstract

Photovoltaic solar energy generation depends directly on meteorological variables such as solar radiation, temperature, relative humidity, and wind speed, which exhibit high temporal variability. In this context, the present study, corresponding to Part I of the SOLAB project, aims to conduct a descriptive and exploratory analysis of the main climatic variables associated with solar energy generation in the municipality of La Paz, Cesar (Colombia). The purpose is to characterize their statistical behavior and internal relationships in order to establish a solid basis for the development of predictive models in later phases. To this end, descriptive statistics techniques, correlation analysis, and graphical visualizations with LOESS and polynomial fitting curves were applied, which made it possible to identify linear and nonlinear trends in the data. The results showed that solar radiation is the variable with the greatest direct influence on the energy generated, while temperature remains within a stable range between 24 °C and 36 °C. The Shapiro–Wilk test confirmed that the variables do not follow a normal distribution. Taken together, this analysis provides a comprehensive understanding of local meteorological behavior and constitutes an essential input for building more accurate predictive models in Part II of the SOLAB project.

**Keywords:** Photovoltaic solar energy; descriptive analysis; meteorological variables; solar radiation; statistical correlation; time series; normality test.

## 1. Introducción

En las últimas décadas, la transición hacia fuentes de energía renovables ha adquirido una importancia estratégica a nivel mundial. Entre estas, la energía solar fotovoltaica se destaca por su capacidad de contribuir a la descarbonización del sistema eléctrico, reducir la dependencia de combustibles fósiles y promover la sostenibilidad ambiental. Sin embargo, su aprovechamiento eficiente se ve limitado por la intermitencia natural de la radiación solar y su dependencia de variables meteorológicas como la temperatura, la humedad relativa y la velocidad del viento, las cuales influyen directamente en la cantidad de energía que puede ser generada [1].

La variabilidad climática representa un desafío técnico para la planificación y operación de sistemas fotovoltaicos, especialmente en regiones tropicales donde las condiciones atmosféricas pueden cambiar bruscamente. Por ello, comprender el comportamiento estadístico y temporal de las variables meteorológicas es un paso esencial antes de aplicar cualquier técnica de modelado o predicción. Este tipo de análisis descriptivo permite identificar patrones de comportamiento, relaciones entre variables, valores extremos y posibles anomalías, aportando información clave para la calibración de futuros modelos predictivos y la toma de decisiones en la gestión energética [2], [3].

Diversos estudios han evidenciado que la caracterización de variables meteorológicas mediante análisis exploratorio constituye una base fundamental para mejorar la precisión en la predicción de generación solar. Por ejemplo, Yılmaz y Şahin [4] demostraron que el uso combinado de radiación solar, temperatura, humedad y velocidad del viento aumenta significativamente la fiabilidad de las estimaciones energéticas. De igual manera,

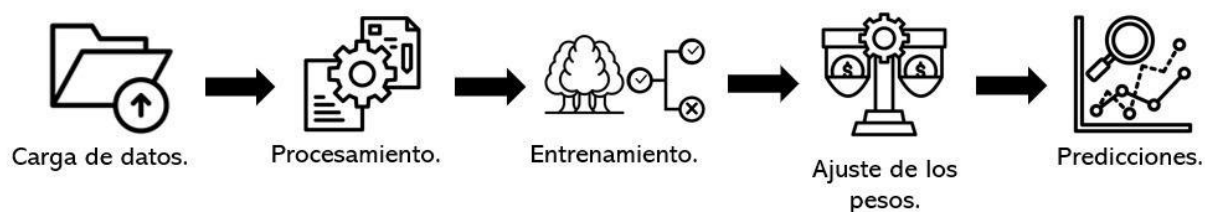
investigaciones recientes destacan la importancia de analizar previamente la distribución, correlación y comportamiento temporal de estos factores para optimizar el diseño de modelos de aprendizaje automático [5]–[7].

En este contexto, el presente estudio corresponde a la Parte I del proyecto SOLAB, enfocado exclusivamente en el análisis descriptivo y exploratorio de las variables meteorológicas asociadas a la generación de energía solar en el municipio de La Paz (Cesar, Colombia). Se analizan datos horarios recolectados entre febrero y mayo de 2025, con el propósito de identificar la estructura, distribución y relaciones estadísticas entre las variables de radiación solar, temperatura, humedad relativa, velocidad del viento y energía solar estimada.

El objetivo principal de esta fase es caracterizar de manera descriptiva el comportamiento de las variables meteorológicas y su relación con la energía solar generada, como base metodológica para el desarrollo posterior de modelos predictivos (Parte II).

## 2. Materiales y Métodos

En esta investigación se busca predecir la generación de energía solar a partir de variables meteorológicas como la temperatura, la velocidad del viento, la humedad relativa, con el tiempo como variable de referencia en la serie temporal. Para ello, se emplea una metodología que consta de dos partes: **(I)** combina análisis exploratorio y visualización, **(II)** modelado predictivo con enfoque en series de tiempo. En la Figura 1, se presenta la estructura general del flujo de trabajo propuesto en el artículo para la generación de los modelos de predicción (véase la Figura 1).



**Figura 1.** Flujo de trabajo propuesto para la construcción de los modelos de predicción.

En este artículo, dada la extensión del análisis, se presenta la parte I, la cual se fundamenta en los conceptos del marco teórico expuesto en la introducción, y se determina la relación entre el clima y el comportamiento de los sistemas solares fotovoltaicos. En la parte II, se crean y evalúan los modelos predictivos.

## 2.1. Fuente de datos y construcción de la base

La base de datos fue obtenida por medio de la API de Open-Meteo [8], que ofrece datos climáticos gratuitos y actualizados. Se seleccionó el municipio de La Paz, Cesar (Colombia) como zona de estudio, y se recolectaron datos horarios entre el 1 de febrero y el 31 de mayo de 2025. Para descargar y organizar los datos se utilizó Python 3 con las bibliotecas *requests* y *pandas*.

A partir de los datos de radiación solar inicialmente disponibles, se estimó la energía generada por un sistema solar residencial de 1.6 m<sup>2</sup>, con una eficiencia neta del 17 %. Sin embargo, en la fase de modelado se decidió no utilizar directamente esta variable, con el objetivo de construir un modelo más autónomo y generalizable basado únicamente en condiciones meteorológicas accesibles. La estimación de energía se mantuvo como variable objetivo del modelo, calculada con la siguiente fórmula:

$$\text{Energía (kWh)} = \frac{\text{Radiación (W/m}^2\text{)} \times \text{Área (m}^2\text{)} \times \text{Eficiencia neta}}{1000}$$

Posteriormente, se construyó una base de datos unificada que incluye observaciones de tres municipios del Cesar: La Paz, San Diego y Valledupar. Esta expansión de la base permite evaluar el comportamiento de los modelos en un entorno multizona, incorporando una nueva variable categórica llamada zona, la cual identifica el origen geográfico de cada observación. Aunque no se usó como entrada directa en los modelos, esta variable fue clave para estructurar el grafo secuencial aplicado en el modelo GCN multizona.

## 2.2. Justificación del diseño experimental

Para este proyecto se trabajó con datos horarios de energía solar entre el 1 de febrero y el 31 de mayo de 2025, en el municipio de La Paz, Cesar. Se eligió este período porque abarca varios meses del primer semestre del año, en los que se presentan cambios importantes en el clima, como el paso de la temporada seca al inicio de las lluvias. Esto permite observar cómo varía la generación de energía solar en diferentes condiciones meteorológicas.

La elección de La Paz se basó en su potencial solar, dado que es una zona donde el sol brilla la mayor parte del año, con una intensidad promedio de radiación solar estimada en aproximadamente 5,0 kWh/m<sup>2</sup>/día, según el *Atlas de Radiación Solar de Colombia* [9]. Este valor refleja condiciones favorables para el aprovechamiento de sistemas fotovoltaicos. En este contexto, el uso de energías limpias representa una solución importante, especialmente en áreas rurales. De esta forma, los datos recolectados resultan útiles y reflejan situaciones reales que podrían aplicarse a proyectos similares en la región.

Adicionalmente, el trabajo con datos por hora permite identificar patrones dentro del mismo día, ya que la energía solar cambia significativamente a lo largo de la jornada. Esta información, junto con las variables de series de tiempo construidas, proporciona una base adecuada para entrenar modelos con mayor capacidad predictiva.

## 2.2. Variables utilizadas

El conjunto de datos quedó conformado por un grupo de variables meteorológicas, temporales y agregadas, diseñadas para capturar el comportamiento estacional y la dependencia temporal de la generación de energía solar. Estas variables se construyeron a partir de los datos originales descargados desde la API de Open-Meteo, enriqueciendo la base con transformaciones útiles para el modelado:

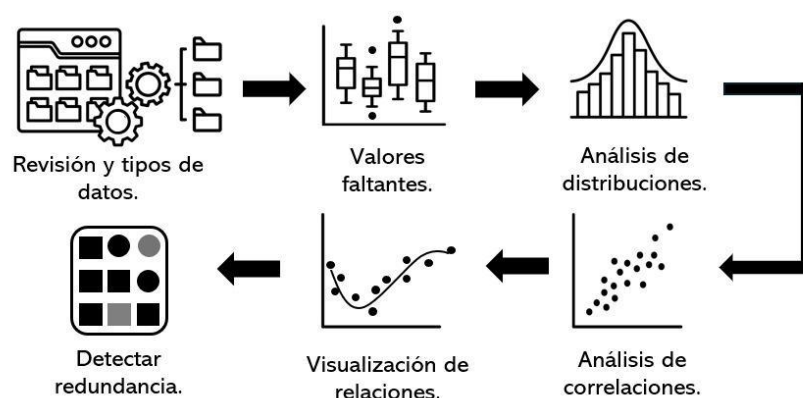
- Time: Fecha y hora de cada registro, utilizada como base para generar variables temporales adicionales.
- Temperatura 2m: Temperatura del aire a 2 metros del suelo.
- Velocidad del viento 10m: Velocidad del viento a 10 metros de altura.
- Humedad relativa 2m: Porcentaje de humedad en el aire a 2 metros.
- hora, día, mes, día del año, día de la semana: Variables extraídas directamente del campo *time* mediante funciones de manejo de fechas en Python (datetime). Son útiles para capturar la estacionalidad horaria, diaria y semanal de la energía solar.
- Energía lag 1 y lag 24: Variables calculadas con la función `shift()` de pandas, que permite crear columnas desplazadas temporalmente. *Lag 1* representa la energía solar generada una hora antes, mientras que *lag 24* captura la energía generada en la misma hora del día anterior, permitiendo detectar repeticiones diarias.
- Promedio móvil 3 Horas: Construido con la función `rolling(window=3).mean()` sobre la serie de energía solar, para capturar suavemente la tendencia reciente en intervalos cortos.
- Zona: Variable categórica que identifica el municipio al que pertenece cada observación (La Paz, San Diego o Valledupar). Fue utilizada exclusivamente en el experimento multizona con GCN para estructurar la secuencia de datos por zona, sin ser una entrada directa del modelo.
- Energía solar kWh: Energía estimada, utilizada como variable objetivo de los modelos.

Estas variables agregadas permitieron transformar el conjunto original de datos en una base más rica para el aprendizaje supervisado, especialmente al trabajar con series de tiempo,

donde la información pasada y los patrones horarios son esenciales para mejorar la precisión predictiva de los modelos.

### 2.3. Análisis exploratorio de datos

Para la construcción de los modelos predictivos, se llevó a cabo un análisis exploratorio con el objetivo de comprender la estructura de los datos, detectar posibles anomalías y evaluar la viabilidad de las variables incluidas. Este proceso permite identificar patrones relevantes y ajustar el tratamiento de los datos previo al modelado.



**Figura 2.** Etapas del análisis exploratorio de datos.

La Figura 2 resume las principales etapas del análisis exploratorio de datos. En primer lugar, se revisaron los tipos y las estructuras de las variables para garantizar la correcta interpretación y el tratamiento estadístico. Posteriormente, se verificó la existencia de valores faltantes mediante mapas de calor, permitiendo evaluar su impacto y definir estrategias de imputación si fuera necesario.

Luego, se analizaron las distribuciones de las variables con histogramas, estimaciones de densidad y la prueba de normalidad de Shapiro-Wilk, con el propósito de identificar sesgos o asimetrías. A continuación, se calcularon correlaciones —tanto lineales como no lineales— junto con análisis de distancia para evaluar asociaciones entre variables.

Este proceso se complementa con visualizaciones de relaciones mediante diagramas de dispersión, curvas LOESS y ajustes polinómicos. Finalmente, se evaluó la redundancia entre variables a través del cálculo del Factor de Inflación de la Varianza (VIF) y el examen de la matriz de correlaciones.

Este enfoque sistemático permitió detectar posibles fuentes de ruido y redundancia, así como confirmar que las variables derivadas del tiempo y del historial energético son particularmente relevantes, incluso en ausencia directa de la variable de radiación solar.

### 3. Resultados

El análisis descriptivo ayuda a entender cómo están organizados los datos, qué tipo de variables tenemos y si están completas. Esta revisión es fundamental para que la información esté lista antes de aplicar modelos o técnicas estadísticas tradicionales o de machine learning.

#### 3.1. Análisis de la Estructura de los Datos

La base de datos empleada para este análisis consta de un total de 2880 observaciones y 6 variables. Cada observación representa una hora del período comprendido entre el 1 de febrero y el 31 de mayo de 2025, para el municipio de La Paz, Cesar. Las variables incluyen tanto medidas meteorológicas como la estimación horaria de energía solar generada, lo que permite realizar un análisis predictivo robusto con base en condiciones ambientales.

**Identificación de Tipos de Variables:** Las variables presentes en la base de datos se clasifican de acuerdo con su tipo y naturaleza estadística, lo cual es fundamental para seleccionar los métodos adecuados de análisis y modelado. A continuación, se detallan los tipos identificados:

- **Variables numéricas continuas:** *shortwave radiation*, *temperature 2m*, *wind speed 10m*, *relative humidity 2m* y *energía solar kWh*. Estas variables representan mediciones físicas que pueden tomar un amplio rango de valores reales. Son fundamentales como predictores (*features*) y como variable objetivo (*target*) en modelos de regresión.
- Variable categórica ordinal: *time*. Aunque esta variable se encuentra en formato de fecha-hora, se considera categórica ordinal, ya que se puede descomponer en componentes temporales como hora, día o mes, y permite ordenar cronológicamente los datos. Es esencial en análisis de series temporales.
- Identificación de claves únicas: *time* es la clave única, ya que cada registro horario es distinto y no presenta duplicados.

#### 3.2. Estadísticos Descriptivos Básicos

**Medidas de tendencia central:** Los resultados obtenidos en la Tabla I indican que, aunque la *shortwave radiation* tiene una media de 249.02 W/m<sup>2</sup>, su mediana es solo 16, lo que revela una alta concentración de valores bajos con algunos picos altos durante el día. Esto mismo se refleja en *energía solar kWh*, donde la media es 0.068 pero la mediana es apenas 0.004, lo que indica que la mayoría del tiempo se genera muy poca energía, probablemente

en horas sin sol. La temperatura y la humedad presentan valores medios estables, lo que sugiere un clima relativamente constante, ideal para analizar patrones energéticos.

**Tabla 1.** Medidas de tendencia central.

| Variable             | Media   | Mediana |
|----------------------|---------|---------|
| shortwave_radiation  | 249.016 | 16.000  |
| temperatute_2m       | 28.758  | 27.900  |
| wind_speed_10m       | 12.383  | 12.800  |
| relative_humudity_2m | 63.140  | 65.000  |
| energia_solar_kwh    | 0.068   | 0.004   |

**Medidas de dispersión:** La radiación solar presenta una variación muy grande, con un valor de SD es de 322.3, como se aprecia en la Tabla II. Este comportamiento es esperable, ya que hay horas del día con mucha luz solar y otras donde no la hay. Esta diferencia también se refleja en la energía solar generada, que depende directamente de cuánta radiación haya en cada momento. De la Tabla III se observa que las variables como la temperatura, velocidad del viento y humedad relativa muestran cambios más moderados - una SD baja, lo que indica que el clima en La Paz es relativamente estable en esos aspectos. En resumen, la radiación solar es la que más cambia, afectando fuertemente cuánto se puede generar de energía solar.

**Tabla II.** Medidas de Dispersión.

| Variable             | Desviación Estándar ( $\sigma$ /SD) | Rango (R) | Rango Intercuartílico (IQR) |
|----------------------|-------------------------------------|-----------|-----------------------------|
| Shortwave_Radiation  | 322.3                               | 995       | 518.5                       |
| Temperatute_2m       | 3.4                                 | 14.8      | 5.5                         |
| Wind_Speed_10m       | 6.5                                 | 30.3      | 10.4                        |
| Relative_Humudity_2m | 15.1                                | 75.0      | 21.0                        |
| Energia_Solar_Kwh    | 0.08                                | 0.27      | 0.14                        |

**Valores mínimos y máximos:** Al analizar los valores extremos, se observa que tanto la radiación solar como la energía generada pueden alcanzar valores de cero (ver Tabla III), situación normal durante la noche cuando no hay incidencia solar. La temperatura oscila entre 22.7 °C y 37.5 °C, lo que corresponde a un clima cálido característico del municipio de La Paz. La velocidad del viento, por su parte, puede ser nula en ciertos momentos y llegar hasta 30.3 m/s, evidenciando la presencia de ráfagas ocasionales. En cuanto a la

humedad relativa, sus valores van del 25 % al 100 %, lo que refleja la alternancia entre ambientes secos y condiciones de alta humedad, posiblemente asociadas a episodios de lluvia.

**Tabla III.** Mínimos y Máximos por Variable.

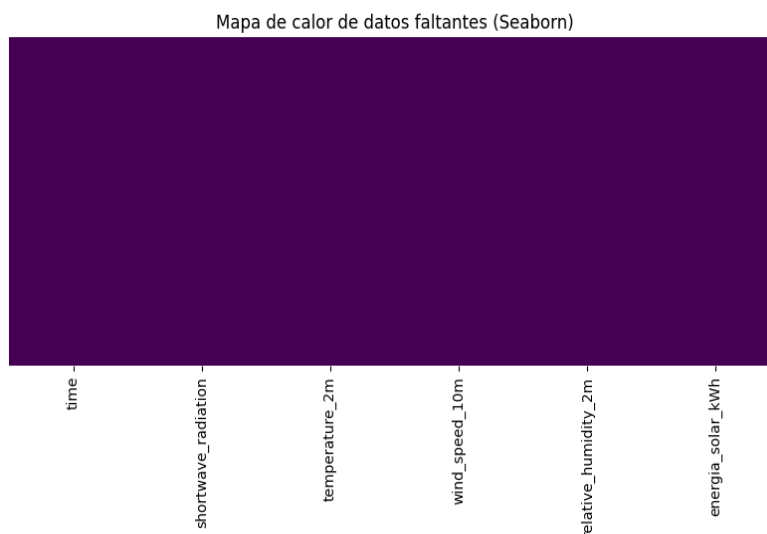
| Variable             | Mín  | Máx  |
|----------------------|------|------|
| shortwave_radiation  | 0.0  | 995  |
| temperatute_2m       | 22.7 | 37.5 |
| wind_speed_10m       | 0.0  | 30.3 |
| relative_humudity_2m | 25.0 | 100  |
| energia_solar_kwh    | 0.0  | 0.27 |

**Percentiles:** Al analizar los percentiles, se aprecia que muchas horas no hubo radiación ni generación solar (valores bajos en P25 y P50), pero en otras horas sí se alcanzaron niveles altos (P75). La temperatura y humedad fueron estables, con valores concentrados en rangos estrechos. La velocidad del viento varió más, lo que puede influir en la eficiencia de los paneles solares.

**Tabla IV.** Percentiles por variable.

| Variable             | P25  | P50   | P75   |
|----------------------|------|-------|-------|
| shortwave_radiation  | 0.0  | 16.0  | 518.5 |
| temperatute_2m       | 25.9 | 27.9  | 31.4  |
| wind_speed_10m       | 6.9  | 12.8  | 17.32 |
| relative_humudity_2m | 53   | 65.0  | 74.0  |
| energia_solar_kwh    | 0.0  | 0.004 | 0.14  |

**3.3. Análisis de Valores Faltantes:** Como se aprecia en la Figura 7, no se encontraron valores faltantes en las variables del conjunto de datos. Esto es positivo para el análisis, ya que garantiza que toda la información está completa y lista para ser utilizada en la construcción del modelo de predicción de energía solar sin necesidad de aplicar técnicas de imputación.



**Figura 3.** Mapa de calor de los valores faltantes por cada variable.

### 3.4. Análisis de Outliers

#### Detección mediante métodos estadísticos (IQR, z-scores)

Para identificar posibles valores atípicos en las variables numéricas del conjunto de datos, fueron aplicados dos métodos estadísticos clásicos: el método del rango intercuartílico (IQR) y el método del puntaje z (z-score).

- Con el método IQR, un valor se considera atípico si cae fuera del rango definido por:

$$\text{Límite inferior} = Q1 - 1.5 \times \text{IQR}$$

$$\text{Límite superior} = Q3 + 1.5 \times \text{IQR}, \text{ donde } \text{IQR} = Q3 - Q1.$$

- Con el método z-score, un valor es considerado atípico si su puntuación estandarizada supera 3 en valor absoluto:

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

y se considera outlier si  $z > 3$ . Después de aplicar ambos métodos a las variables radiación solar, temperatura 2m, velocidad viento 10m, humedad relativa 2m y energía solar kWh, no se detectaron valores atípicos. Esto indica que los datos están bien distribuidos y no presentan valores extremos que puedan distorsionar el análisis. Este resultado se interpreta positivamente, ya que facilita el uso directo de estos datos en modelos predictivos sin necesidad de transformaciones adicionales o eliminación de registros. Esta condición ofrece

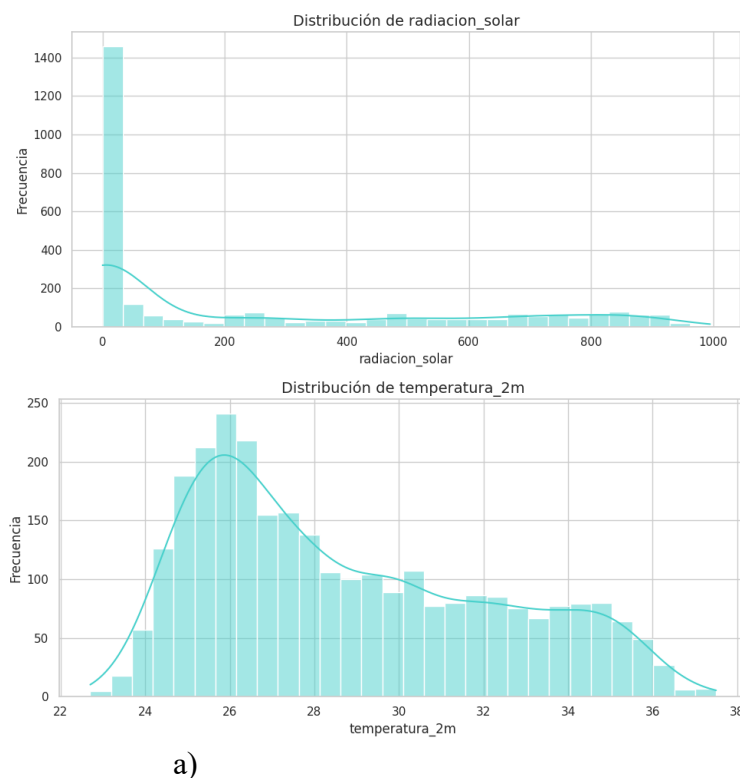
mayor confianza para continuar con el modelado de la generación de energía solar sin preocupación por efectos indeseados causados por outliers.

### 3.5. Análisis de Distribuciones

Hacer el análisis de distribuciones resulta útil para visualizar el comportamiento de los datos a usar para una predicción de la energía solar. Por ejemplo, podemos ver si hay muchas horas con poca radiación o si la temperatura casi siempre es parecida. Esto nos sirve para entender mejor los datos y preparar bien la información antes de entrenar y seleccionar el modelo que va a hacer las predicciones.

#### Histogramas y gráficos de densidad

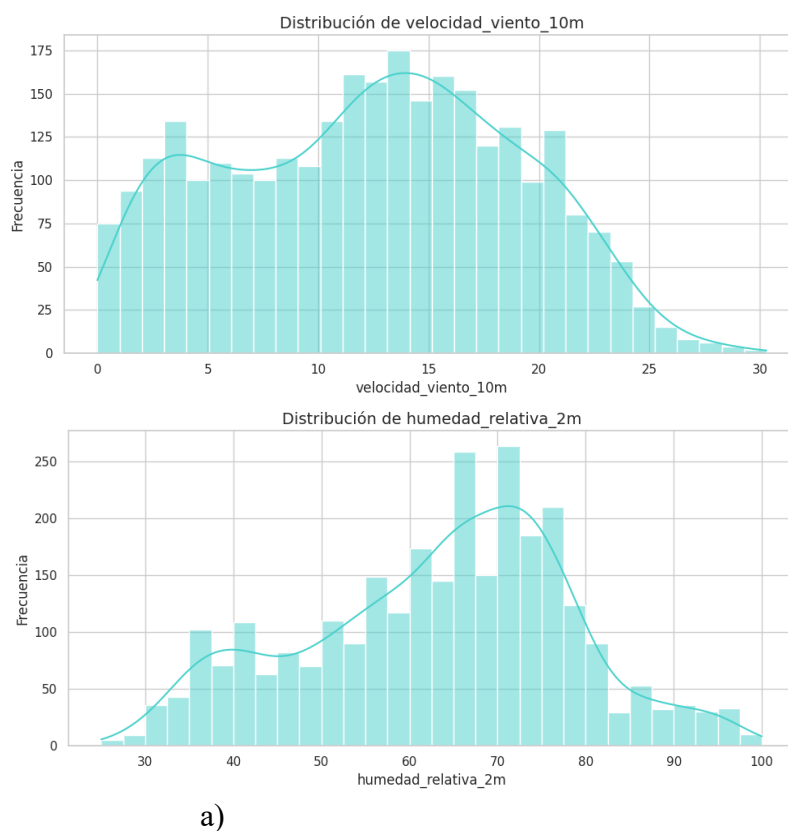
A continuación, se muestran las distribuciones de las variables numéricas para analizar su comportamiento y posibles patrones que influyen en la generación de energía solar. De la Figura 4a se observa que, en la distribución de la radiación solar, la mayoría de los valores están cerca de 0, y hay pocos datos con radiación alta, lo que nos muestra una distribución sesgada a la derecha. Esto significa que en muchas horas del día la radiación es baja (como en la madrugada o en días nublados), lo cual puede influir en que se genere poca energía solar durante esos momentos.



**Figura 4:** a) Distribución de la radiación solar y b) distribución de temperatura 2m.

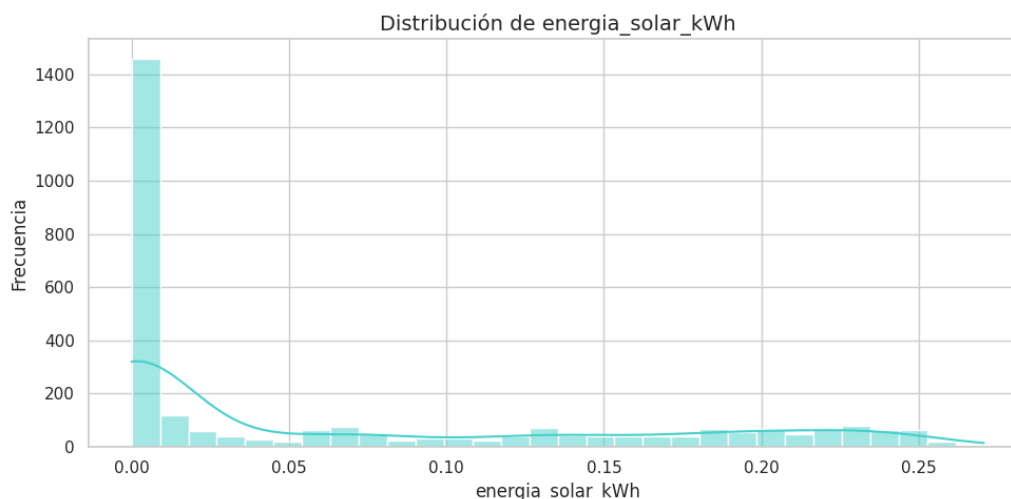
Por otra parte, la distribución de temperatura (ver Figura 4b), se aprecia que tiene una forma parecida a una campana, aunque con una pequeña inclinación a la derecha. Esto indica que la temperatura se mantiene generalmente entre 24°C y 36°C, lo cual es útil para el diseño de modelos predictivos porque la temperatura puede afectar cuánta energía solar se genera, aunque no tenga una relación lineal.

La distribución de la velocidad del viento mostrada en la Figura 5a, se observan varios picos en diferentes valores, lo que muestra una distribución multimodal. Esto quiere decir que hay varios rangos de velocidad del viento que se repiten con frecuencia, lo cual puede influir en el clima y, en algunos casos, en la eficiencia de los paneles solares.



**Figura 5:** a) Distribución de la Velocidad del viento 10m y b) distribución de la humedad relativa 2m.

Por otra parte, en la Figura 5b se aprecia que en la mayoría de los datos están entre 57% y 77%, pero también valores más bajos y más altos. Esta gráfica muestra una distribución asimétrica, un poco más dispersa, lo que indica que el clima suele ser húmedo, lo cual puede afectar la entrada de luz solar a los paneles.



**Figura 6:** Distribución de la Energía solar (kWh).

Por último, la distribución de la energía solar mostrada en la Figura 6, casi todos los valores están muy cerca de cero, y solo unos pocos son más altos. Esto representa una distribución muy sesgada a la derecha, lo que significa que se genera poca energía en la mayoría de los momentos, posiblemente por las condiciones del clima (como nubes o lluvia).

### Pruebas de normalidad (Shapiro Wilk)

Al aplicar la prueba de Shapiro-Wilk a las variables numéricas, se encontró que todas presentan un p-valor menor a 0,05, lo que indica que ninguna sigue una distribución normal, como se muestra en la Tabla V. Esta característica es importante en el momento de la elección de modelos predictivos de machine learning, ya que algunos modelos asumen que los datos siguen una distribución normal. Es de resaltar que muchos de estos modelos, especialmente aquellos basados en redes neuronales, son capaces de capturar patrones complejos sin depender de supuestos de normalidad.

**Tabla V.** Resultados de la prueba de normalidad de Shapiro-Wilk.

| Variable             | Est  | p-value |
|----------------------|------|---------|
| shortwave_radiation  | 0.74 | 0.0     |
| temperatute_2m       | 0.92 | 0.0     |
| wind_speed_10m       | 0.98 | 0.0     |
| relative_humudity_2m | 0.97 | 0.0     |
| energia_solar_kwh    | 0.74 | 0.0     |

## Asimetría y curtosis

En el análisis de la asimetría de las variables, de la Tabla VI se observa que, en general, los valores no están muy alejados de cero, lo que indica que las distribuciones no presentan una inclinación marcada. Por ejemplo, la variable velocidad viento 10m es casi simétrica, mientras que radiación solar, temperatura 2m y energía solar kWh presentan una ligera asimetría positiva, es decir, la mayoría de los datos se concentran hacia la izquierda y presentan una cola que se extiende hacia la derecha. En contraste, la humedad relativa 2m muestra una leve asimetría negativa.

**Tabla VI.** Asimetría y curtosis de las variables meteorológicas.

| Variable             | Asimetría | Curtosis |
|----------------------|-----------|----------|
| shortwave_radiation  | 0.8827    | -0.8206  |
| temperatute_2m       | 0.5514    | -0.8437  |
| wind_speed_10m       | -0.0122   | -0.8948  |
| relative_humudity_2m | -0.2377   | -0.4759  |
| energia_solar_kwh    | 0.8827    | -0.8206  |

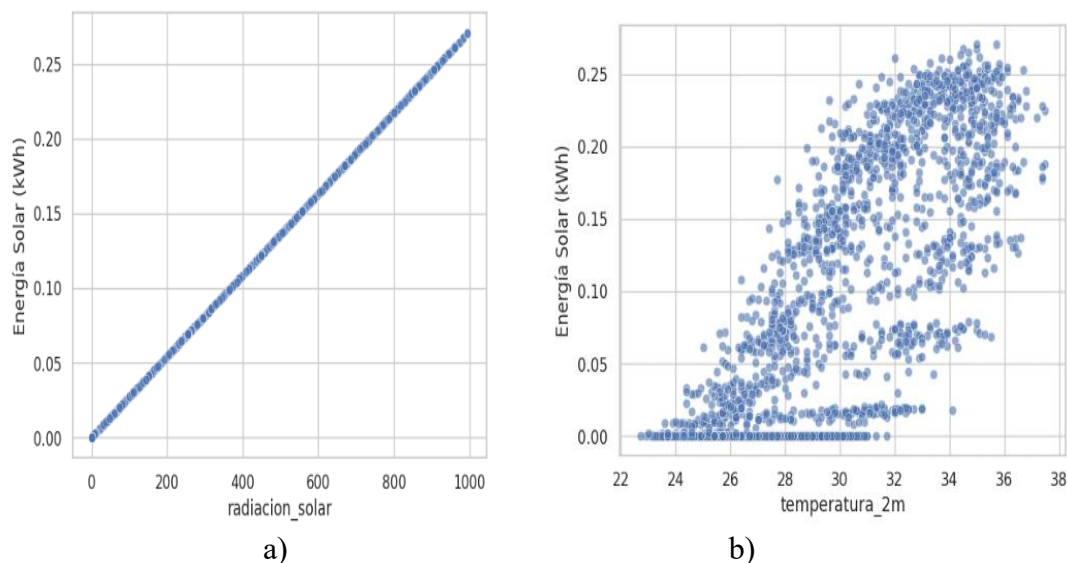
En cuanto a la curtosis, todos los valores obtenidos son negativos, lo que indica que las distribuciones son más planas que una distribución normal. Esto implica que no presentan picos muy marcados y sus colas son más ligeras. Aunque este comportamiento no representa un problema para los modelos de predicción en fases posteriores, sí contribuye a una mejor comprensión de cómo se distribuyen los datos y qué tan concentrados o dispersos pueden estar.

## 3.6. Análisis de Relaciones

El análisis de relaciones es importante porque nos permite identificar qué variables están conectadas entre sí y cuáles pueden ser más útiles para predecir la variable objetivo. Además, permite entender el comportamiento del conjunto de datos y evita problemas si algunas variables son muy parecidas entre sí.

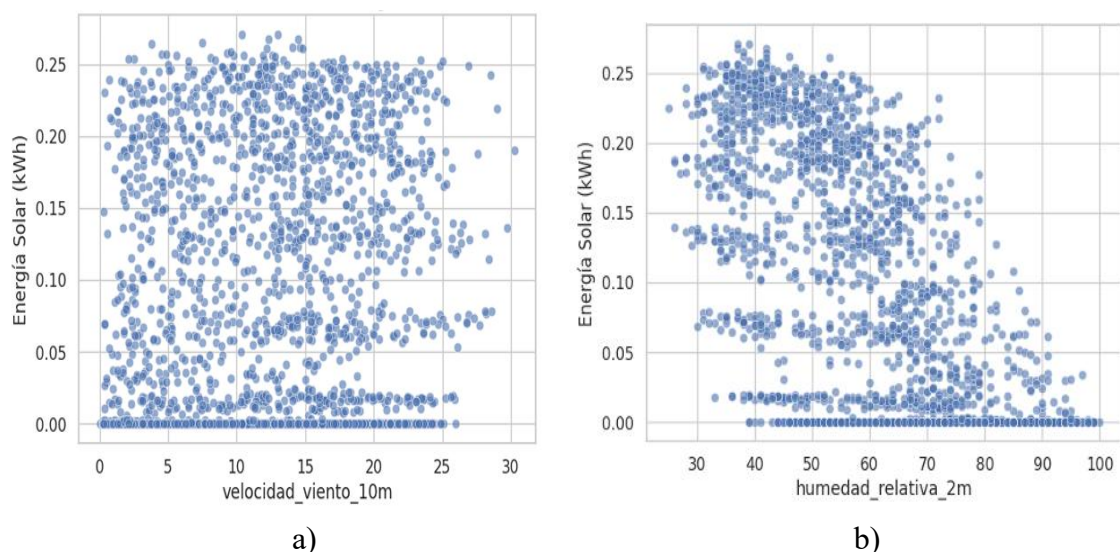
### 3.6.1. Relaciones Lineales

En el gráfico de dispersión entre la Energía Solar y la Radiación Solar presentado en la Figura VIIa, se observa una relación completamente lineal positiva entre estas. Esto resulta coherente, ya que, a mayor cantidad de radiación solar, se produce más energía. Esta relación tan clara confirma que la radiación solar constituye una variable clave en el modelo predictivo.



**Figura VII:** a) Radiación Solar vs Energía Solar y b) Temperatura vs Energía Solar

En el gráfico de la Figura VIIb, se observa que, a medida que la temperatura aumenta, también tiende a incrementarse la energía generada, aunque no de forma tan precisa como en el caso anterior. Se identifica una tendencia positiva, pero con mayor dispersión. Esto indica que la temperatura influye en la generación de energía, aunque no de manera tan directa como lo hace la radiación solar.



**Figura VIII:** a) Velocidad del Viento vs Energía Solar y b) Humedad Relativa vs Energía Solar.

Al analizar la relación entre la velocidad del viento y la energía generada de la Figura VIIa, los puntos se encuentran más dispersos y no se observa una relación clara entre las

variables. Esto sugiere que la variable correspondiente al viento no presenta un impacto directo significativo sobre la generación de energía solar en la zona de estudio.

Por último, de la Figura VIIIb se muestra una tendencia inversa: cuando la humedad es muy alta, la energía generada tiende a ser menor. Aunque hay bastante dispersión, se puede ver que la humedad puede estar relacionada negativamente con la producción de energía solar, tal vez porque los días más húmedos suelen ser más nublados.

### 3.6.2. Detección de Relaciones No Lineales

Este análisis es útil porque nos permite ver si las variables están relacionadas de formas que no son lineales. Así podemos entender mejor cómo factores como la temperatura o la humedad afectan la energía solar, incluso si no lo hacen de forma directa. Esto ayuda a que el modelo predictivo, por ejemplo, modelos basados en árboles como XGBoost, obtengan mejores resultados.

#### Correlaciones No Paramétricas (Spearman y Distance Correlation)

En el análisis de correlación presentado en la Tabla VII, se observa que la variable que presenta una mayor relación con la energía solar generada es la radiación solar, ya que muestra una correlación perfecta (1.0) tanto en Spearman como en distancia. Este resultado es coherente, dado que, a mayor radiación, mayor es la cantidad de energía solar que puede generarse. Asimismo, se identifica que la temperatura posee una relación fuerte (0.76 en Spearman y 0.81 en distancia), lo cual indica que, al aumentar la temperatura, generalmente también se incrementa la energía generada.

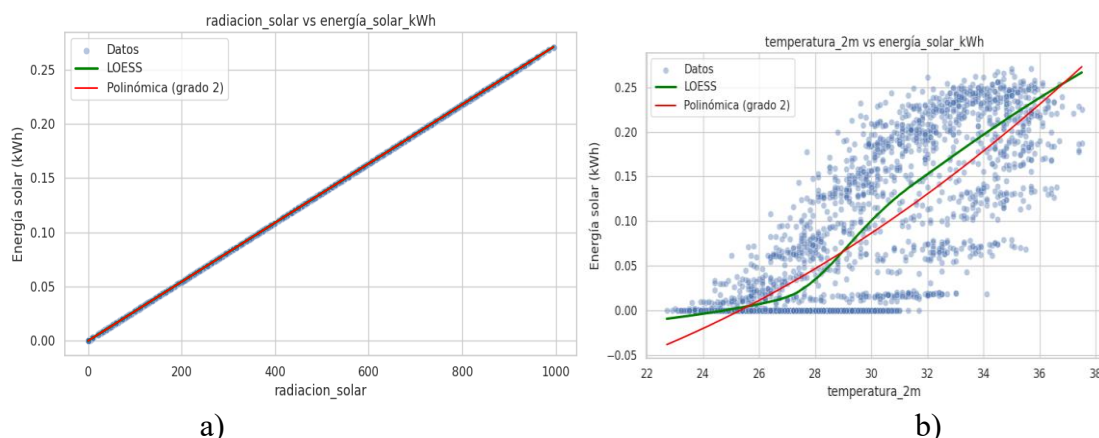
**Tabla VII.** Correlación de Spearman y correlación por distancia respecto a la variable objetivo.

| Variable             | Spearman | Distance |
|----------------------|----------|----------|
| shortwave_radiation  | 1.0      | 1.0      |
| temperatute_2m       | 0.76     | 0.80     |
| wind_speed_10m       | 0.06     | 0.07     |
| relative_humudity_2m | -0.62    | 0.65     |

Por otro lado, la velocidad del viento muestra una relación muy baja con la energía solar, lo que sugiere que esta variable casi no influye en el comportamiento de la energía generada. Finalmente, la humedad relativa tiene una relación negativa moderada en Spearman (-0.63), pero positiva en distancia (0.65), lo que nos dice que su relación con la energía no es lineal: aunque a más humedad se reduce un poco la energía, la conexión no es tan simple y podría tener otros patrones.

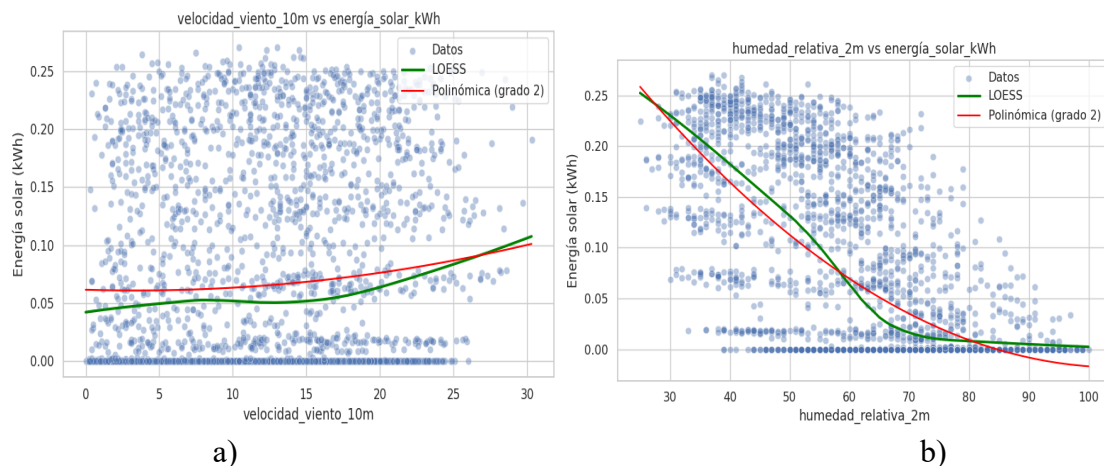
## Scatterplots con curvas de ajuste (LOESS, polinómicas)

Se elaboraron diagramas de dispersión con curvas de ajuste LOESS y polinómicas para analizar la relación entre la energía solar generada y las variables meteorológicas, con el fin de identificar comportamientos lineales y no lineales presentes en los datos. En el caso de la radiación solar presentado en la Figura 9a, se observa una relación claramente directa con la energía generada, a medida que aumenta la radiación, también incrementa la producción de energía. Ambos tipos de ajuste, LOESS y polinómico, se adaptan adecuadamente a la tendencia observada, confirmando que esta variable constituye el principal factor determinante en la generación de energía solar.



**Figura 9:** a) Radiación solar vs Energía Solar y b) Temperatura (2m) vs Energía Solar.

De la Figura 9b se aprecia que la relación de la temperatura con la energía también es positiva, pero no es lineal. A partir de cierta temperatura, la energía solar empieza a aumentar más rápidamente. La curva LOESS muestra una forma curva más flexible, mientras que la polinómica también capta la tendencia ascendente. Esto indica que, a más temperatura, usualmente hay más energía solar, pero no de manera constante.



**Figura 10:** a) Velocidad del Viento (10m) vs Energía Solar y b) Humedad Relativa (2m) vs Energía Solar.

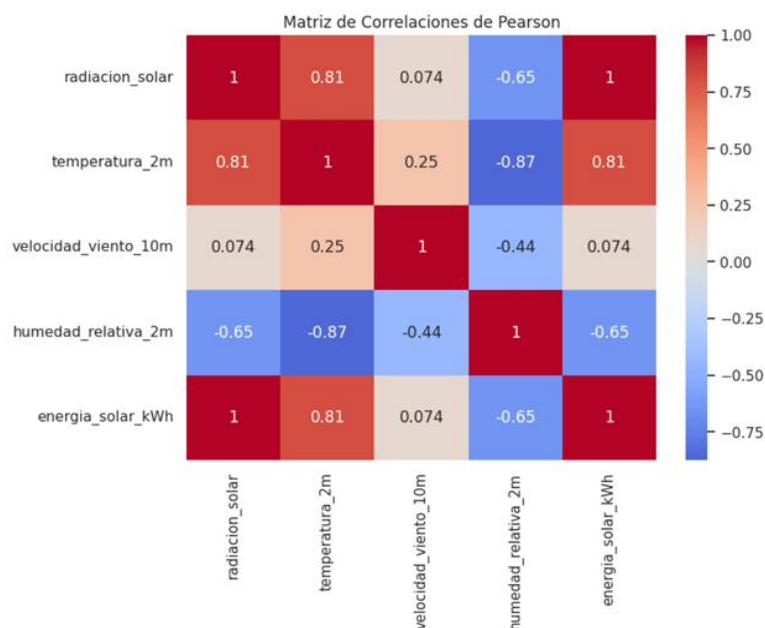
El gráfico de la Figura 10a muestra que la velocidad del viento tiene muy poca relación con la energía solar. Las curvas no tienen una forma definida y los puntos están muy dispersos. Aunque hay una ligera tendencia ascendente en las curvas, no es muy fuerte. En resumen, el viento no influye mucho en la producción de energía solar.

Por otra parte, en la Figura 10b se identifica una relación negativa: cuando la humedad es más alta, la energía solar tiende a ser más baja. Las curvas muestran que esta relación no es completamente lineal, pero sí evidente. Este comportamiento resulta lógico, ya que una mayor humedad suele estar asociada con cielos más nublados, lo que reduce la cantidad de radiación solar que llega y, en consecuencia, disminuye la producción de energía.

### 3.7. Análisis de Redundancia y Multicolinealidad

#### Detección de variables altamente correlacionadas

- Matriz de correlaciones (alto valor absoluto)



**Figura 11:** Matriz de Correlaciones de Pearson.

El gráfico de correlación de la Figura 11 nos muestra qué tan conectadas están las variables. Por ejemplo, cuando hay más radiación solar o más temperatura, también suele haber más energía generada. En cambio, cuando la humedad sube, la temperatura y la energía bajan. La velocidad del viento casi no tiene relación con las otras variables. Esto nos ayuda a ver si hay datos muy parecidos que puedan confundir al modelo.

- **Factor de Inflación de Varianza (VIF > 10)**

En el análisis del Factor de Inflación de la Varianza (VIF) no encontramos ninguna variable con valores superiores a 10. Esto significa que no hay evidencia de multicolinealidad grave entre las variables, es decir, ninguna está explicándole de manera excesiva con base en las demás. Podemos decir que las variables se complementan y aportan información distinta.

- **Variables redundantes (correlación > 0.9)**

Al revisar la matriz de correlación, tampoco detectamos pares de variables con una correlación mayor a 0.9, por lo que no existen variables redundantes que estén diciendo prácticamente lo mismo. Esto es bueno porque no hace falta eliminar ninguna por ser repetitiva.

## **Conclusiones**

El análisis descriptivo permitió identificar patrones fundamentales en el comportamiento de las variables meteorológicas asociadas a la generación de energía solar en el municipio de La Paz, Cesar. A partir de la prueba de Shapiro–Wilk, se determinó que las variables no presentan una distribución normal, evidenciando una alta variabilidad y la presencia de asimetrías en los datos, particularmente en la radiación solar y la velocidad del viento. Esta característica estadística resulta clave para la etapa posterior de modelado, pues condiciona la selección de algoritmos y la aplicación de técnicas de normalización adecuadas.

Los resultados mostraron que la temperatura oscila predominantemente entre 24 °C y 36 °C, mientras que la radiación solar exhibe un comportamiento cíclico bien definido, con máximos en horas centrales del día. Estas condiciones reflejan una estabilidad térmica y un régimen de radiación favorable para la generación fotovoltaica en la zona. Los diagramas de dispersión con curvas de ajuste LOESS y polinómicas evidenciaron una relación directa y consistente entre la radiación solar y la energía generada, confirmando su papel como variable de mayor influencia. Por otro lado, la humedad relativa y la velocidad del viento mostraron asociaciones más débiles o no lineales, lo que sugiere efectos indirectos sobre la eficiencia energética.

En términos metodológicos, este análisis exploratorio permitió comprender la estructura, variabilidad y correlaciones entre las variables meteorológicas, sirviendo como base sólida para la fase predictiva (Parte II). Los hallazgos obtenidos orientan la selección de variables relevantes, la detección de outliers y la construcción de conjuntos de datos limpios y representativos, elementos indispensables para entrenar modelos de aprendizaje automático robustos y precisos.

Finalmente, los resultados del presente estudio aportan evidencia empírica útil para la planificación energética local y el diseño de estrategias de integración de energía solar en redes eléctricas. Asimismo, contribuyen a fortalecer la comprensión del comportamiento meteorológico regional, aspecto esencial para mejorar la fiabilidad y sostenibilidad de los sistemas fotovoltaicos en contextos tropicales.

## **Agradecimientos**

Los autores agradecen a la Universidad Nacional de Colombia por el apoyo brindado.

## Referencias

- [1] M. Abdelsattar, A. AbdelMoety, y A. Emad-Eldeen, “Advanced machine learning techniques for predicting power generation and fault detection in solar photovoltaic systems,” *Neural Comput. Appl.*, vol. 37, núm. 15, pp. 8825–8844, 2025. [En línea]. Disponible en: <https://doi.org/10.1007/s00521-025-11035-6>
- [2] A. Alcañiz, D. Grzebyk, H. Ziar, y O. Isabella, “Trends and gaps in photovoltaic power forecasting with machine learning,” *Energy Rep.*, vol. 9, pp. 447–471, 2023. [En línea]. Disponible en: <https://doi.org/10.1016/j.egy.2022.11.208>
- [3] H. I. Aouidad y A. Bouhelal, “Machine learning-based short-term solar power forecasting: a comparison between regression and classification approaches using extensive Australian dataset,” *Sustainable Energy Res.*, vol. 11, núm. 1, 2024. [En línea]. Disponible en: <https://doi.org/10.1186/s40807-024-00115-1>
- [4] H. Yılmaz y M. Şahin, “Solar panel energy production forecasting by machine learning methods and contribution of lifespan to sustainability,” *Int. J. Environ. Sci. Technol. (Tehran)*, vol. 20, núm. 10, pp. 10999–11018, 2023. [En línea]. Disponible en: <https://doi.org/10.1007/s13762-023-05110-5>
- [5] M. Gupta, A. Arya, U. Varshney, J. Mittal, y A. Tomar, “A review of PV power forecasting using machine learning techniques,” *Progress in Engineering Science*, vol. 2, núm. 1, p. 100058, 2025. [En línea]. Disponible en: <https://doi.org/10.1016/j.pes.2025.100058>
- [6] A. Nadeem, A. Naseem, A. Khan, M. Hussain, y F. Khan, “AI-Driven precision in solar forecasting: Breakthroughs in machine learning and deep learning,” *AIMS Geosci.*, vol. 10, núm. 4, pp. 684–734, 2024. [En línea]. Disponible en: <https://doi.org/10.3934/geosci.2024035>
- [7] Ozdemir, M. Kuzlu, y F. O. Catak, “Machine learning insights into forecasting solar power generation with explainable AI,” *Electr. Eng. (Berl., Print)*, 2024. [En línea]. Disponible en: <https://doi.org/10.1007/s00202-024-02933-4>
- [8] “API meteorológica gratuita de código abierto,” *Openmeteo.com*. [En línea]. Disponible en: <https://open-meteo.com/en/docs>
- [9] IDEAM y UPME, *Atlas de Radiación Solar de Colombia*. Bogotá, Colombia: Ministerio de Minas y Energía, 2005. [En línea]. Disponible en: <https://www.ideam.gov.co/web/tiempo-y-clima/radiacion>