

**Clasificación de tipos de cielo basado en mediciones de  
radiación solar**

**Giovanni Gamaliel López Padilla**  
**Adriana Ipiña Hernandez**  
**Oscar Dalmau Cedeño**

**Resumen**

La radiación solar participa en múltiples procesos biológicos y atmosféricos indispensables para la vida en la Tierra. La energía solar desencadena un gran número de reacciones y en las grandes ciudades interviene en la fotoquímica del smog. La variación de la intensidad solar depende en general de la composición atmosférica, ubicación geográfica, época del año y de la hora del día. Determinar la irradiancia solar global nos permite conocer su disponibilidad para beneficio humano, extendiendo su uso a través de la conversión y almacenamiento e incluso se puede utilizar en tratamientos médicos. En este contexto, identificar y caracterizar los elementos que la atenúan es esencial. Las nubes juegan un rol fundamental en el balance radiativo y dependiendo de su altura y estructura pueden dispersar, absorber o reflejar los fotones del sol. La cobertura de nubes es clave para cualquier tipo de pronóstico meteorológico, siendo la radiación solar un detector natural de las mismas. En las últimas décadas se han desarrollado una variedad de modelos para clasificar los diferentes tipos de cielo según las condiciones de nubes o porcentaje de nubosidad. Estos modelos son comparados con mediciones in situ o datos satelitales. Los niveles de complejidad y parámetros de entrada para estimar la irradiancia solar global en cielo despejado, con frecuencia se basan en expresiones empíricas. En otros casos, se requieren parámetros geométricos como el ángulo cenital o parámetros meteorológicos básicos como horas de sol, humedad relativa, presión, índice de claridad y temperatura. Algunos modelos incluyen parámetros de las condiciones atmosféricas como la profundidad óptica del aerosol, cantidad de agua precipitable y columna de ozono. Los datos de cielo despejado suelen extraerse del conjunto de datos de radiación solar medidos en todos los tipos de cielo, a menudo mediante el uso de algoritmos que se basan en otros parámetros meteorológicos medidos. Los procedimientos actuales para la extracción de datos de cielo despejado se han examinado y comparado entre sí para determinar su confiabilidad y dependencia de la ubicación. Se implementaron una serie de modelos de clasificación que tienen como base teórica en geometría, topología de datos y redes neuronales. Los modelos estiman si un día presenta las condiciones de un cielo despejado, parcialmente nublado ó nublado. Los mdoelos fueron entrenados usando vectores diarios de las diferencias o proporciones entre las mediciones in situ y los modelos de irradiancia solar extraterrestre y Robledo-Soler. Se obtuvo que los modelos de redes convolucionales y bosques aleatorios entrancados con las proporciones de las mediciones diarias y el modelo Robledo-Soler estiman de manera correcta con mayor regularidad la condición del cielo en el día a predecir.

## 1. Introducción

La irradiancia solar a nivel de suelo esta influenciada por las condiciones de cielo. El índice de claridad de cielo (sky clearness index) es la proporción entre la radiación global en la superficie y la radiación solar extraterrestre.<sup>1</sup> Este índice puede ser un gran estimador para cuantificar una observación a lo largo del tiempo. Se ha exhibido que existe una caracterización basada en el índice de claridad cuando se tiene una resolución de los datos menor a 5 minutos,<sup>2-4</sup> la cual es capaz de distinguir de manera precisa la condición de cielo binaria (despejado o nublado). Para la clasificación diaria basada en radiación solar a nivel de suelo existen tres clases: despejado, parcialmente nublado y nublado. Maafi<sup>5</sup> empleó la dimensión fractal aplicada a señales y Harrouni<sup>6</sup> implementó un análisis fraccional con una resolución de 10 minutos para realizar la clasificación diaria de las condiciones de cielo. En este trabajo se empleará una resolución de 1 hora para mediciones de irradiancia global a nivel de suelo implementando modelos basados en geometría, topología de los datos y redes neuronales para realizar la clasificación de las condiciones de cielo diarias.

## 2. Modelos de irradiancia solar

Los modelos de irradiancia solar pueden estimar el valor con respecto a las condiciones de cielo despejado.<sup>7-9</sup> Las estimaciones obtenidas por los modelos son usadas para realizar comparaciones sobre mediciones de una locación, las cuales pueden contener datos equivocados o faltantes. Existen diversos modelos basados en redes neuronales donde a partir de parámetros geoespaciales ó datos meteorológicos estiman el promedio por hora, dia, mes o año.<sup>10-13</sup> Estos modelos requieren que se les alimente con una gran cantidad de información, lo cual representa un problema. Aunado a esto, los modelos están delimitados por la precisión que tienen sus estimaciones.<sup>14</sup> En este trabajo se propone el uso de modelos que puedan adaptarse a las mediciones in situ con la suficiente precisión para estimar la irradiancia global bajo condiciones de cielo despejado.

### 2.1. Declinación y ángulo solar

La dependencia del tiempo de los modelos de irradiancia solar se introduce por medio de la declinación solar y el ángulo solar (ecuación 1 y 2 respectivamente).

$$\delta = 24.45 \sin(\gamma) \quad (1)$$

$$\omega = 15(h_{LTC} - 12) \quad (2)$$

donde  $h_{LTC}$  es la hora local y  $\gamma$  es la fracción de rotación de la tierra con respecto al sol. El ángulo cenital está definido como:

$$z = \cos^{-1}(\cos(\phi)\cos(\delta)\cos(\omega) + \sin(\phi)\sin(\delta))$$

Donde  $\phi$  es la latitud de la locación.

### 2.2. Irradiancia solar extraterrestre

El modelo de irradiancia solar extraterrestre ( $GHI_0$ ) está definida como:<sup>1</sup>

$$GHI_0 = I_{SC} \left[ 1 - 0.033 \cos\left(\frac{360n}{365}\right) \right] \cos(z) \quad (3)$$

Donde  $I_{SC}$  es la constante solar con valor de  $1367 \text{ W/m}^2$ ,  $n$  es el día consecutivo del año ( $n=1$  y  $365$  son el primero y último día del año respectivamente, para años bisiestos el denominador cambia a  $366$  y el último día se toma como  $366$ ).

### 2.3. Irradiancia solar global horizontal

Kwarikunda et al 2021<sup>15</sup> menciona que realizó comparaciones entre los modelos Berger-Duffle (BD), Adnot-Bouges-Campana-Gicquel (ABCG) y Robledo-Soler (RS). El modelo RS estima con más cercanía a las mediciones a nivel de suelo en diferentes locaciones. El modelo RS se define como:

$$GHI_{RS} = a(\cos z)^b \exp(-c(90 - z)) \quad (4)$$

donde  $\cos z$  es el angulo cenital. En la tabla 1 se encuentran los parámetros  $a$ ,  $b$ ,  $c$  usados.

Parametro	a	b	c
Valor	1119	1.19	$1 \times 10^{-6}$

Tabla 1: Parámetros del modelo RS.

### 3. Sistema de Monitoreo Ambiental

El Área Metropolitana de Monterrey (AMM) se ubica en una región montañosa donde se realizan extracciones de material para la construcción (pedreras) a la par de actividades industriales y alto flujo vehicular. El Sistema de Monitoreo Ambiental (SIMA) tiene como objetivo evaluar la calidad del aire, monitoreando las concentraciones de contaminantes atmosféricos a las que se encuentra expuesta la población del AMM. Cuando el AMM se encuentra con altos índices de contaminación atmosférica que representa un peligro para el ser humano, el SIMA es el responsable de reportar estos periodos. El SIMA se compone de 13 estaciones de monitoreo repartidas a lo largo del AMM (figura 1).

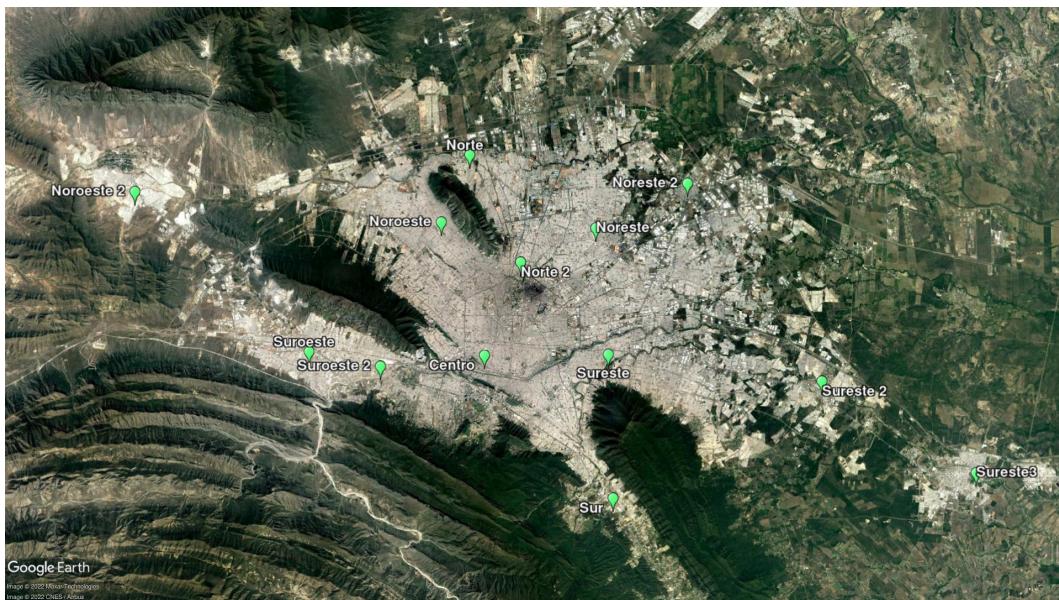


Figura 1: Ubicación geográfica de las estaciones meteorológicas del SIMA en el AMM.

En la tabla 2 se muestra la información geográfica de las estaciones del SIMA en el AMM.

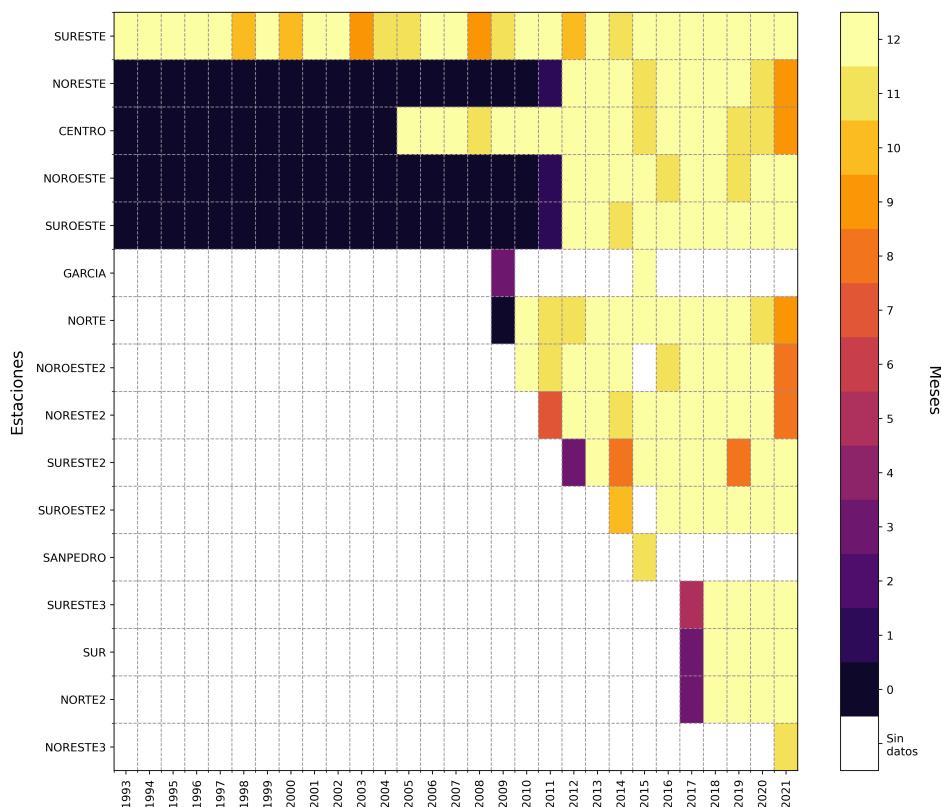
Ciudad	Nombre	Elevación (m s. n. m.)	Latitud (°N)	Longitud (°O)
Guadalupe	Sureste	492	25.67	-100.25
Monterrey	Centro	560	25.67	-100.34
Monterrey	Noroeste	571	25.76	-100.37
San Nicolas de los Garza	Noreste	476	25.75	-100.26
Santa Catarina	Suroeste	694	25.68	-100.46
Garcia	Noroeste2	716	25.78	-100.59
Escobedo	Norte	528	25.80	-100.34
Apodaca	Noreste2	432	25.78	-100.19
Juarez	Sureste2	387	25.65	-100.10
San Pedro Garza Garcia	Suroeste2	636	25.66	-100.41
Cadereyta de Jimenez	Sureste3	340	25.58	-99.99
Monterrey	Sur	630	25.57	-100.25
San Nicolas de los Garza	Norte2	520	25.73	-100.31

Tabla 2: Información de la localización geográfica de las estaciones del SIMA en el AMM.

La base de datos que contiene mediciones de irradiancia solar por hora en las estaciones del SIMA en el periodo 1993-2021. Se realizó un filtro de los datos que cumplen las siguientes condiciones:

- Un dato diario es válido cuando contiene al menos 10 mediciones entre las 8 a las 19 horas.
- Un mes es considerado cuando contiene al 21 datos diarios validados.

En la figura 2 se muestra la distribución de los meses en las estaciones del SIMA que cumplen las condiciones de selección.



**Figura 2:** Distribución de los meses validos para las estaciones meteorológicas del SIMA en el periodo 1993-2021.

## 4. Creación de la base de datos

Se seleccionaron las estaciones noroeste, noreste, sureste2 y suroeste en el periodo 2019-2021. Estas estaciones fueron elegidas debido a que la topografía es representativa y se encuentra en zonas estratégicas de diferentes fuentes de emisión. Adicionalmente cuentan con un gran número de mediciones dentro del periodo seleccionado. Con base a las mediciones diarias de cada estación se clasificaron de acuerdo al comportamiento de su intensidad solar en función de las horas del día. Las condiciones de cielo contempladas son: despejado, parcialmente nublado y nublado.

### 4.1. Criterios para las condiciones de cielo despejado

Los criterios para la clasificación de cielo despejado de manera predeterminada de las mediciones diarias son las siguientes:

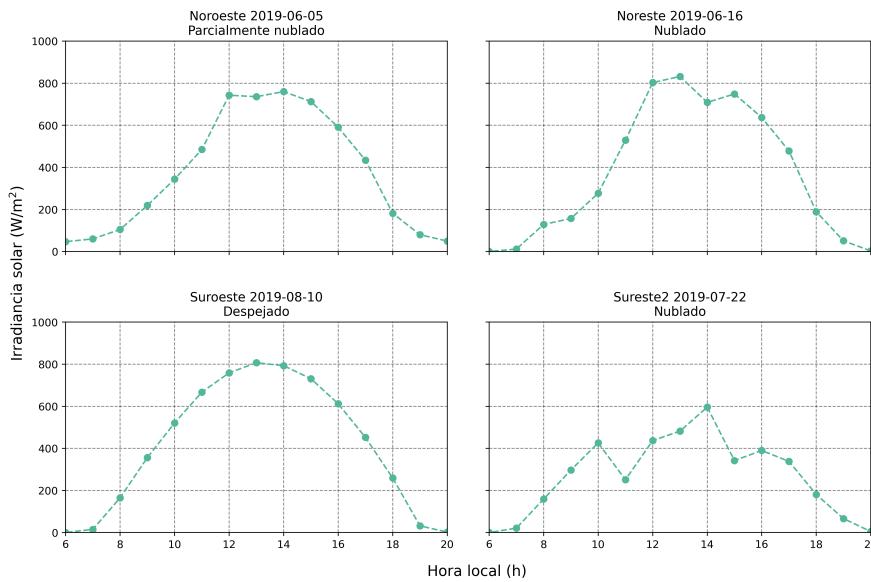
- Cielo despejado
  - Un día de cielo despejado se caracteriza por tener un comportamiento gaussiano a lo largo del día, teniendo como máximo el mediodía solar. Para el AMM, el mediodía solar debe encontrarse alrededor de las 12:30-14:30 horas. Las mediciones deben registrar un valor diferente a cero entre las 6 a las 20 horas.
- Cielo parcialmente nublado

- Un día de cielo parcialmente nublado se caracteriza por presentar el comportamiento de un día con cielo despejado pero en ciertos intervalos de tiempo. Esto puede ocurrir en solo una hora, o en varias. Si el día contiene entre uno y cinco mediciones que caracterizan a un día despejado, entonces el día será clasificado como parcialmente nublado.

■ Cielo nublado

- Un día nublado se caracteriza por presentar un comportamiento caótico o un comportamiento gaussiano con variaciones abruptas en intervalos de tiempo cortos.

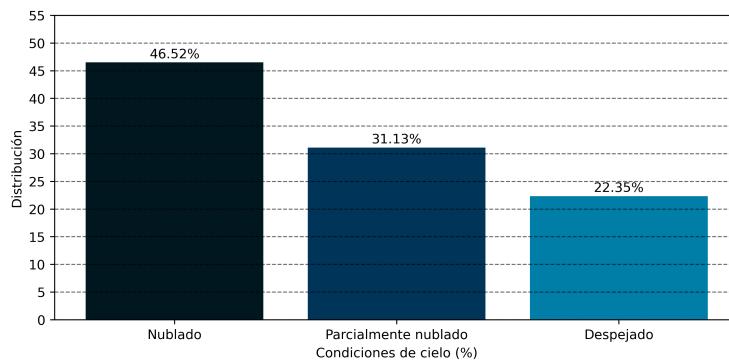
En la figura 3 se presentan diferentes mediciones donde se presentan las diferentes condiciones de cielo.



**Figura 3:** Ejemplos de las clasificaciones de las condiciones de cielo a partir de mediciones diarias de cada una de las estaciones del SIMA.

#### 4.2. Distribución de los datos

En la figura 4 se muestra la distribución de las condiciones de cielo clasificadas de manera predeterminada, en esta se observa que existe una mayor cantidad de días nublados en comparación de los días categorizados como despejado o parcialmente nublado.



**Figura 4:** Distribución de las clasificaciones de las condiciones de cielo en la base de datos.

### 4.3. Operaciones de comparación

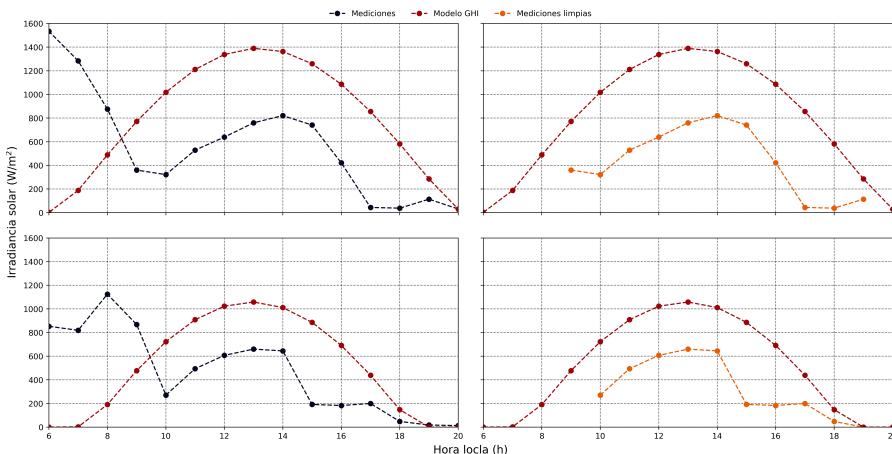
Se implementaron dos operaciones para comparar las diferencias y proporciones (ecuación 5 y 6 respectivamente) entre las mediciones y los modelos GHI<sub>0</sub> y RS.

$$d_t = \begin{cases} \text{Modelo} - \text{Medición} & \text{si Modelo} \neq 0 \\ 0 & \text{si Modelo} = 0 \end{cases} \quad (5)$$

$$k_t = \begin{cases} \frac{\text{Medición}}{\text{Modelo}} & \text{si Modelo} \neq 0 \\ 0 & \text{si Modelo} = 0 \end{cases} \quad (6)$$

### 4.4. Datos atípicos

Los datos de las estaciones del SIMA pueden contener ruido o mediciones que físicamente no son posibles, a estos datos los denominamos como atípicos. Se implementó un filtrado de datos automática. El filtro de datos consiste en realizar una operación de comparación para cada medición (ecuación 6) con respecto al modelo GHI, si para alguna hora se obtiene un valor mayor a 0.9, entonces la medición corresponde a un dato atípico y se eliminará de la base de datos. Si el modelo GHI<sub>0</sub> es igual a 0, entonces se sobreescribe la medición con el valor 0, esto con el propósito de eliminar el ruido que puede tener el radiómetro de la estación analizada. En la figura 5 se visualiza el proceso de filtrado de datos atípicos en dos fechas diferentes. Los valores atípicos ocurren a menudo al inicio o al final del día solar.



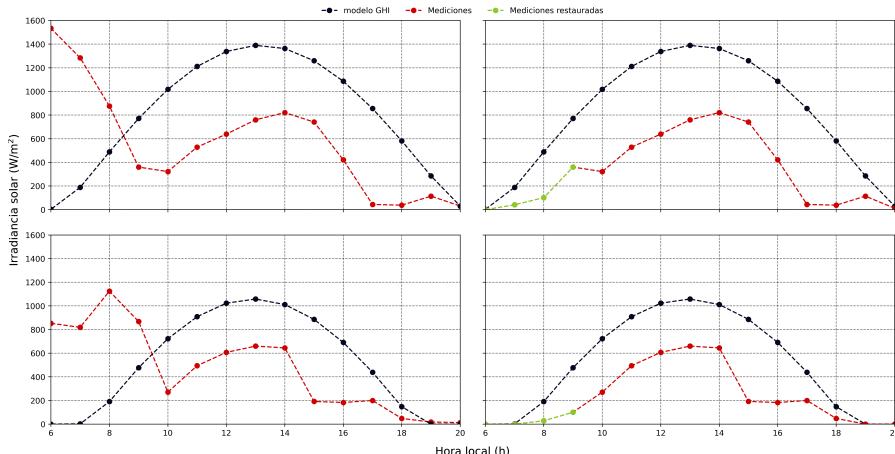
**Figura 5:** Mediciones de irradiancia solar de la estación noroeste originales (izquierda) y sin valores atípicos (derecha).

### 4.5. Reconstrucción

A partir de los datos filtrados, se aplicó un proceso de reconstrucción. El proceso de reconstrucción consiste en asignar el valor del promedio horario de las 10 primeras mediciones que tengan mayor semejanza a la medición a reconstruir. Se toman únicamente las mediciones de la misma estación en una ventana de tres meses (un mes anterior, el mes actual y el siguiente). La semejanza se calcula como:

$$sim(m_i, m_j) = \frac{m_i \cdot m_j}{||m_i|| ||m_j||} \quad (7)$$

En la figura 6 se muestran los datos restaurados para los casos presentados en la figura 5.



**Figura 6:** Restauracion de mediciones por medio de promedios horarios de las 30 mediciones más semejantes al día seleccionado

Con las mediciones restauradas se realizaron las comparaciones (ecuación 5 y ecuación 6) con respecto a los modelos  $GHI_0$  y RS.

## 5. Modelos de clasificación

Los modelos de clasificación basados en geometría, topología de datos y redes neuronales han sido aplicados en el área de la física de la atmósfera, tal como la clasificación de la sustentabilidad de una ciudad,<sup>16</sup> clasificación de contaminantes en el agua por medio de videos<sup>17</sup> y clasificación del efecto tóxico en base a la concentración de contaminantes en peses.<sup>18</sup>

### 5.1. Modelos clásicos

Tradicionalmente la solución a los problemas de clasificación se ha realizado por medio de modelos estadísticos ó geométricos. El rendimiento del modelo dependerá del patrón de correlaciones que mantengan los predictores con la información de entada. En este trabajo se implementaron los modelos Support Vector Machine (SVM), K vecinos más cercanos (KNN), Árbol de decisión, Bosque Aleatorio y Naives Bayes Gaussiano.

**Support Vector Machine** El algoritmo de Support Vector Machine (SVM) es un algoritmo de aprendizaje supervisado que se utiliza en problemas de clasificación y regresión. El objetivo del algoritmo SVM es encontrar un conjunto de hiperplanos que separen de la mejor manera posible a las clases de los datos dados. Cada hiperplano resultante tendrá un margen amplio entre cada clase de datos. El margen se define como la distancia máxima a la región paralela al hiperplano que no contiene datos en su interior. Existen funciones que pueden transformar las características del hiperplano, estas funciones son llamadas funciones kernel. En la tabla 3 se encuentran las diferentes funciones kernel que son mayormente usadas.

Función	Kernel
Gaussiana	$K(x_1, x_2) = \exp\left(-\frac{\ x_1 - x_2\ ^2}{2\sigma^2}\right)$
Lineal	$K(x_1, x_2) = x_1^T x_2$
Polinomial	$K(x_1, x_2) = (x_1^T x_2 + 1)^\rho$
Sigmoide	$K(x_1, x_2) = \tanh(\beta_0 x_1^T x_2 + \beta_1)$

**Tabla 3:** Funciones kernel con los parámetros de cada función.

**KNN** El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un algoritmo de aprendizaje supervisado no paramétrico. El algoritmo usa la proximidad para realizar una

clasificación o una predicción. Generalmente el algoritmo se usa como un modelo de clasificación. Para problemas de clasificación se asigna una etiqueta de clase en base al número que se presenta con mayor frecuencia alrededor del punto dado. Para este caso se utilizo la metrica de minkowski con  $p = 2$  y considerando los 3 vecinos más cercanos.

**Árbol de decisión** Un árbol de decisión es un modelo basado en el aprendizaje supervisado. El modelo divide el espacio de predictores aplicando una serie de reglas o decisiones en la que contenga la mayor proporción posible de individuos de una de las categorías dadas. En este trabajo se utilizo la función gini para entrenar el modelo.

**Bosques aleatorios** El algoritmo de bosque aleatorio es un algoritmo de aprendizaje supervisado, el cual esta basado en un conjunto de árboles de decisión combinados con un método de votación. En este trabajo se utilizo la función gini para entrenar el modelo y 1000 estimadores.

**Naive Bayes Gaussiano** El algoritmo Naive Bayes Gaussiano es un algoritmo de aprendizaje supervisado. El algoritmo esta basado el el teorema de bayes. La asunción que toma el algoritmo es la independencia entre las categorías. E

## 5.2. Modelos basados en redes neuronales

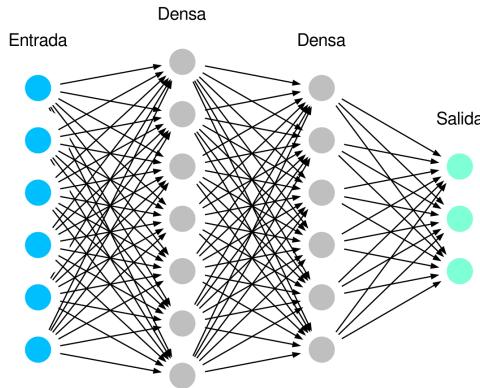
A partir de 1986,<sup>19</sup> los modelos conexionistas o redes neuronales han sido utilizados como herramientas de predicción y clasificación. Un modelo basado en redes neuronales es un sistema informático que mediante el uso de pesado auto-modificable.

**Perceptrón** El modelo neuronal perceptrón fue propuesto por Frank Rosenblatt.<sup>20</sup> El perceptrón usa una matriz para representar las redes neuronales, la cual es llamada matriz de pesos o solamente pesos. Los componentes de un modelo perceptrón son la capa de entrada, la capa oculta, una función de activación y la salida. En la capa oculta es donde se calcula la multiplicación con la matriz de pesos. En la tabla 4 se encuentran las funciones más usadas como funciones de activación.

Función	Definición
Lineal	$f(x) = x$
Sigmoide	$f(x) = \frac{1}{1+e^{-x}}$
ReLU	$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \leq 0 \end{cases}$
Tanh	$f(x) = \tanh(x)$

**Tabla 4:** Funciones de activación comúnmente usadas.

El perceptrón multicapa tiene una estructura similar a la de un modelo perceptrón. En este caso se incluyen capas ocultas donde todas se encuentran conectadas, a esto se le denomina como una capa densa. En cada iteración existe una actualización con propagación hacia atrás en las capas densas para entrenar el modelo. En nuestro caso se implemento un perceptrón multicapa con tres capas de 256, 128 y 3 capas densas con la función de activación sigmoide. En la figura 7 se representa de manera visual el modelo perceptrón.



**Figura 7:** Representación del modelo perceptrón multicapa.<sup>21</sup>

En la tabla 5 se muestra el número de parámetros y la dimensión de los vectores salida en capa del modelo creado.

Capa	Salida	Número de parámetros
Flatten	24	0
Densa 1	256	6400
Densa 2	128	32896
Densa 3	3	387

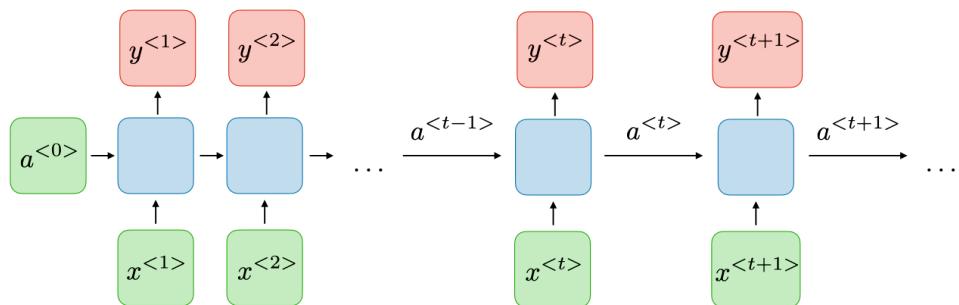
**Tabla 5:** Estructura del modelo perceptrón implementado.

**Red Recurrente** Las redes neuronales recurrentes (RNN) son una clase de redes neuronales que permiten conocer la salida anterior y utilizarla conociendo sus pesos. Para cada tiempo  $t$ , la función de activación ( $a_t$ ) y la salida ( $y_t$ ) pueden ser calculadas con las ecuaciones 8 y 9.

$$a_t = g_1(W_{aa}a_{t-1} + W_{ax}x_t + b_a) \quad (8)$$

$$y_t = g_2(W_{ya}a_t + b_y) \quad (9)$$

En la figura 8 se visualiza a arquitectura de una RNN.



**Figura 8:** Ilustración de una red neuronal recurrente.<sup>22</sup>

En la tabla 6 se muestra la estructura de la RNN creada para este proyecto. La capa RNN tiene como función de activación a la función ReLU y la capa densa una función sigmoidal.

Capa	Salida	Número de parámetros
RNN	64	4224
Densa	3	195

**Tabla 6:** Estructura del modelo RNN implementado.

**Red Convolucional** A diferencia de un modelo perceptrón, los modelos de redes neuronales convolucionales (CNN) consta de dos partes, una que se encarga del proceso de la convolución y otra del proceso de la predicción o la clasificación. En el proceso de convolución el objetivo es extraer la información mas relevante para la tarea. La operación convolución es comunmente usada para realizar transformaciones y obtener información de ella.<sup>23-25</sup> En nuestro caso, estamos tratando con un vector, por ende la convolución que se aplica sera una restringida en una dimensión. Después del proceso de las capas de convolución se obtienen una serie de vectores de una mayor dimensión, para reducir las mismas pasan por una capa llamada fully connected. La capa más común para realizar este proceso es llamada Max-polling, la cual dentro de una ventana en la serie de vectores obtiene únicamente el máximo. En la figura se muestra un ejemplo de esta capa en una matriz de 4x4.

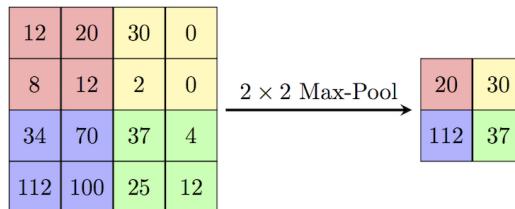


Figura 9: Ejemplo de la capa Max-polling sobre una matriz de 4x4.<sup>26</sup>

En nuestro caso, en vez de usar una capa de Max-polling se implemento un capa de Global-Average-Polling,<sup>27</sup> la cual realiza un promedio de los vectores obtenidos a partir de todas las convoluciones obtenidas. En la figura 10 se muestra un ejemplo de una arquitectura de una CNN.

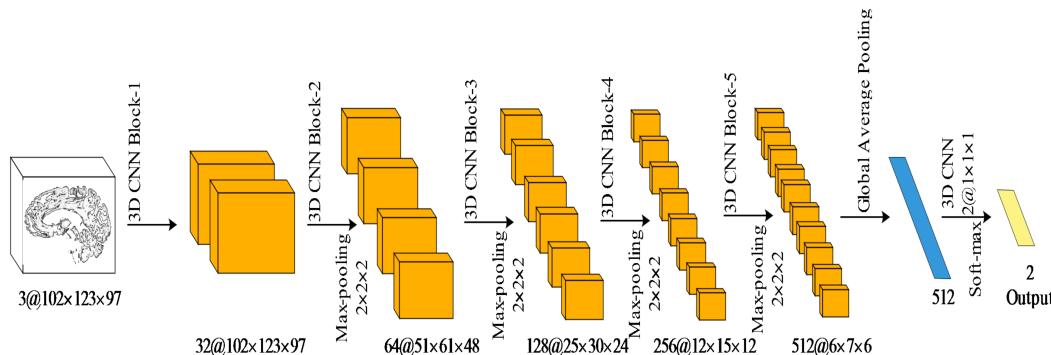


Figura 10: Arquitectura de una CNN.<sup>28</sup>

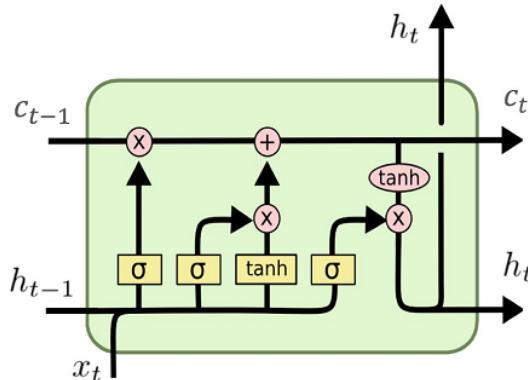
En la tabla 7 se muestra la estructura del modelo CNN implementado. Las capas convoluciones tienen como función de activación a la función ReLU y la capa densa la función sigmoide.

Capa	Salida	Número de parámetros
Conv1D 1	22,100	400
Conv1D 2	20,200	60200
Conv1D 3	18,200	120200
Global Average	200	0
Densa	3	603

Tabla 7: Estructura del modelo CNN implementado.

**Long short-term memory** El modelo long short-term memory (LSTM) es una red neuronal la cual contiene propagación hacia atras semejante a las RNN. La red LSTM es creada para modificar una característica de la red RNN. La característica que se modificada es que en el momento  $t$ , el estado oculto de los tiempos anteriores tienen una atribución menor conforme avanza el tiempo. Una red LSTM preserva la contribución de datos importantes independientemente de cuando aparezca.

Por lo tanto, puede tener una memoria de corto y largo plazo. En la figura se muestra una celda de la red LSTM. A diferencia de una celda de una red RNN, la LSTM contiene un elemento adicional llamado celda de estado ( $c_t$ ) . La celda de estado es la encargada de preservar la información relevante en cualquier tiempo.



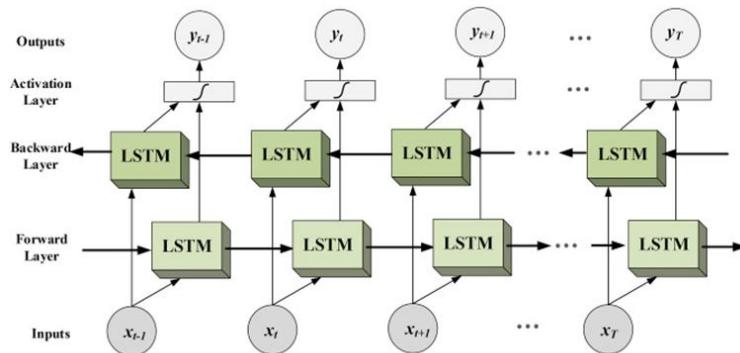
**Figura 11:** Celda de la red LSTM con sus elementos.<sup>29</sup>

En la tabla 8 se encuentra la estructura de la red LSTM implementada. La capa densa tiene como función de activación la función sigmoide.

Capa	Salida	Número de parámetros
LSTM 1	24,256	264192
LSTM 2	256	525312
Densa	3	771

**Tabla 8:** Estructura del modelo LSTM implementado.

**Bidireccional Long-short-term memory** El modelo Bidireccional Long short-term memory (Bi-LSTM) añade la característica de analizar los datos de entrada hacia delante y hacia atras. En la figura 12 se muestra la estructura interna de una red Bi-LSTM.



**Figura 12:** Estructura interna de la red Bi-LSTM.<sup>30</sup>

En la tabla 9 se muestra la estructura interna del modelo Bi-LSTM implementado.

Capa	Salida	Número de parámetros
Bi-LSTM 1	24,512	528384
Bi-LSTM 2	512	1574912
Densa	3	1539

**Tabla 9:** Estructura del modelo Bi-LSTM implementado.

**Red convolucional con atención** La atención es una técnica que toma la idea de la atención cognitiva de los humanos. La idea principal es que el modelo se enfoque en cierta parte de la entrada e ignore la otra, esto por medio de un sistema de peso. El uso de múltiples mecanismos de atención afronta la desventaja que tiene la convolución.<sup>31,32</sup> Es por ello, que se implementó esta capa de atención en la estructura de la red CNN descrita en la tabla 7. En la tabla 10 se encuentra la estructura de la red CNN con atención implementada.

Capa	Salida	Número de parámetros
Conv1D 1	20,100	600
Conv1D 2	18,200	60200
Conv1D 3	16,200	120200
Atención	32	52800
Densa	3	99

**Tabla 10:** Estructura del modelo CNN con atención implementado.

**Esquema de votación** En base a los modelos neuronales implementados, se desarrolló un esquema de votación en el cual se seleccionan a los tres mejores modelos en base a su precisión para cada estación y se realiza una media aritmética de su resultado final.

### 5.3. Metricas

Para medir el rendimiento de cada modelo se utilizará como base la matriz de confusión. En la tabla 11 se presenta un ejemplo de una matriz de confusión. Cada columna representa el número de predicciones de cada clase, mientras que las filas representan a las clases reales.

Clase real	Clase predicha		
	Nublado	Parcialmente nublado	Despejado
Nublado	100	11	1
Parcialmente nublado	13	42	10
Despejado	0	4	51

**Tabla 11:** Ejemplo de una matriz de confusión con las clases de las condiciones de cielo.

Las metricas precision (ecuación 10) y recall (ecuación 11) que se pueden obtener a partir de la matriz de confusión. Estas metricas dependen de cada clase.

$$\text{Precision} = \frac{tp}{tp + fp} \quad (10)$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad (11)$$

donde tp son el número de predicciones correctamente etiquetadas, fn es el número de predicciones que predicen la inexistencia de cierta característica pero en realidad sí la presentan y fp es el número de predicciones que producen la presencia de cierta característica pero en realidad no la presentan.

La metrica F-Score combina los valores de las metricas precision y recall. La metrica F-Score está definida en la ecuación 12.

$$F = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (12)$$

Para resumir la información de las metricas antes mencionadas se utiliza la metrica accuracy. La metrica accuracy esta definida en la ecuación 13.

$$\text{Accuracy} = \frac{\text{Número correcto de predicciones}}{\text{Número total de predicciones}} \quad (13)$$

Para cada modelo se genera un reporte de metricas como el mostrado en la tabla 12.

	Precision	Recall	F-Score
<b>Nublado</b>	0.88	0.89	0.89
<b>Parcialmente nublado</b>	0.74	0.65	0.69
<b>Despejado</b>	0.82	0.93	0.87
<b>Accuracy</b>			0.83

**Tabla 12:** Ejemplo del reporte de metricas por cada modelo de clasificación.

## 6. Resultados

Se entrenaron los modelos clásicos y basados en redes neuronales con los vectores de comparación diarios. El vector diario obtenido a partir de la ecuación 5 es denominado diff y el vector obtenido a partir de la ecuación 6 se denomina ratio. Para los modelos clásicos se obtuvieron mejores metricas cuando se utiliza como vector de entrada los valores que se encuentran entre las 7 y 20 horas. En cambio, para los modelos basados en redes neuronales se obtienen mejores resultados utilizando los 24 valores diarios.

En las tablas 13a, 13b, 13c y 13d se muestran las metricas de accuracy para todos los modelos de clasificación implementados.

Modelo	diff GHI	diff RS	ratio GHI	ratio RS
SVM	0.71	0.72	0.72	0.73
KNN	0.81	0.79	0.71	0.70
Bosques aleatorios	0.77	<b>0.82</b>	0.77	0.80
Naive Gaussiano	0.66	0.68	0.68	0.67
Árbol de decisión	0.73	0.75	0.72	0.77
Perceptron	0.62	0.74	0.74	0.71
CNN	<b>0.82</b>	0.81	<b>0.85</b>	<b>0.85</b>
LSTM	0.73	0.79	0.77	0.79
RNN	0.78	0.69	0.76	0.76
Bi LSTM	0.74	0.81	0.80	0.84
Attention CNN	0.43	0.42	0.82	0.82
Votación	0.74	0.79	0.82	<b>0.85</b>

Modelo	diff GHI	diff RS	ratio GHI	ratio RS
SVM	0.80	0.80	0.80	0.81
KNN	0.78	0.79	0.78	0.78
Bosques aleatorios	<b>0.84</b>	<b>0.85</b>	0.83	<b>0.88</b>
Naive Gaussiano	0.69	0.70	0.77	0.78
Árbol de decisión	0.78	0.75	0.79	0.82
Perceptron	0.66	0.70	0.81	0.84
CNN	<b>0.84</b>	0.81	0.82	0.84
LSTM	0.81	<b>0.85</b>	0.76	0.82
RNN	0.73	0.79	0.81	0.81
Bi LSTM	0.77	0.80	<b>0.84</b>	0.83
Attention CNN	0.48	0.48	0.81	0.83
Votación	0.77	0.80	0.81	0.83

(a) Estacion Noroeste.

(b) Estacion Noreste.

Modelo	diff GHI	diff RS	ratio GHI	ratio RS
SVM	0.74	0.75	0.75	0.75
KNN	0.75	0.76	0.70	0.69
Bosques aleatorios	0.81	0.83	<b>0.83</b>	0.83
Naive Gaussiano	0.69	0.67	0.71	0.70
Árbol de decisión	0.76	0.75	0.77	0.75
Perceptron	0.65	0.70	0.74	0.73
CNN	<b>0.85</b>	<b>0.84</b>	0.82	0.83
LSTM	0.74	0.76	0.80	0.83
RNN	0.74	0.75	0.81	<b>0.84</b>
Bi LSTM	0.72	0.81	0.79	<b>0.84</b>
Attention CNN	0.42	0.42	0.82	0.82
Votación	0.74	0.76	0.82	0.83

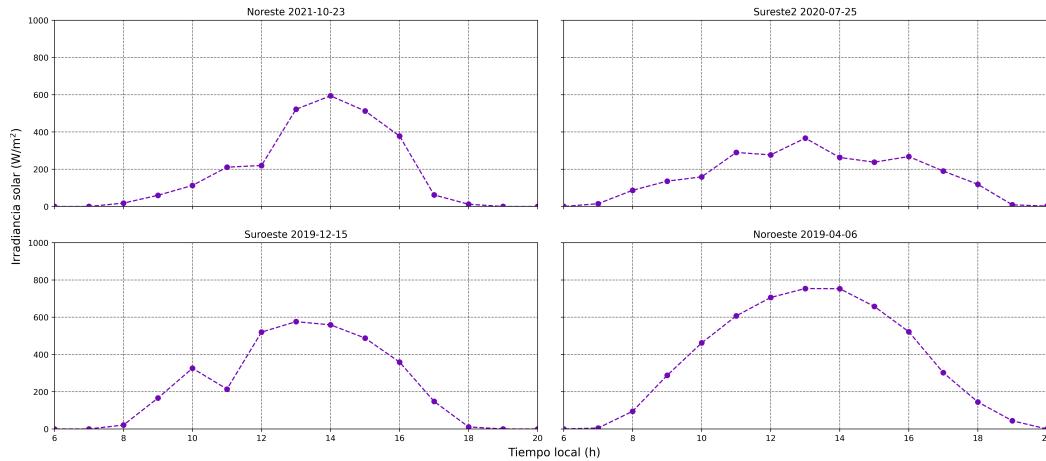
Modelo	diff GHI	diff RS	ratio GHI	ratio RS
SVM	0.73	0.76	0.77	0.81
KNN	0.76	0.81	0.70	0.67
Bosques aleatorios	0.82	<b>0.84</b>	0.82	0.83
Naive Gaussiano	0.67	0.71	0.71	0.70
Árbol de decisión	0.71	0.80	0.75	0.79
Perceptron	0.50	0.64	0.76	0.80
CNN	<b>0.86</b>	0.80	<b>0.84</b>	<b>0.86</b>
LSTM	0.71	0.79	0.78	0.83
RNN	0.80	0.79	0.80	0.82
Bi LSTM	0.72	0.80	0.80	0.84
Attention CNN	0.47	0.47	0.82	0.83
Votación	0.72	0.79	0.80	0.83

(c) Estacion Suroeste.

(d) Estacion Sureste 2.

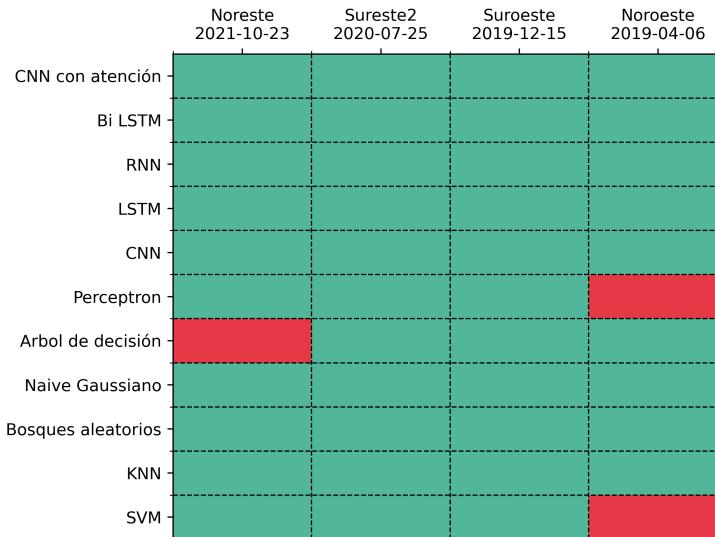
**Tabla 13:** Accuracy obtenido por los diferentes modelos de clasificación para las estaciones seleccionadas con base a los diferentes tipos de comparación.

Se seleccionaron aleatoriamente una fecha por estación para probar los modelos. En la figura 13 se muestran las fechas seleccionados.



**Figura 13:** Fechas de la base de datos elegidos aleatoriamente.

El modelo de arboles de decisión fallo con la medición de la estación Noreste, este lo clasificó como parcialmente nublado. Los modelos de Perceptron y SVM fallaron con la medición de la estación Noreste, clasificando este día como parcialmente nublado. En las demás fechas, los modelos clasificaron de forma correcta a las mediciones tomadas aleatoriamente. En la figura 14 se muestran los resultados de las clasificaciones de los modelos implementados para las fechas mostradas en la figura 13



**Figura 14:** Resultados de la clasificación usando los modelos implementados para las fechas seleccionadas. El color verde y rojo indica si el modelo clasificó de forma correcta ó incorrecta la condición de cielo del día.

## 7. Discusión y conclusiones

Con base en las tablas 13a, 13b, 13c y 13d se pueden obtener las siguientes conclusiones:

- Los modelos entrenados con base en los vectores diarios ratio con el modelo RS exhiben un mejor accuracy en comparación a los modelos entrenados con los otros vectores diarios.
- Los modelos de bosques aleatorios y el CNN muestran los mejores resultados de accuracy independientemente de las estaciones en las cuales fueron entrenados. Sin embargo, el modelo CNN mostró una mayor estabilidad en el momento de pruebas de parámetros. Esto debido a que los valores de accuracy eran cercanos a los mostrados en este trabajo. El modelo de bosques aleatorios era muy sensible a los parámetros, por lo que si se eligen parámetros aleatoriamente puede que se obtengan peores resultados a los antes mostrados.

En general, cuando un día es clasificado con condiciones de cielo despejado o nublado, es poco probable que el modelo se haya equivocado. Teniendo más seguridad cuando el día es clasificado como nublado. Por otro lado, cuando un día es clasificado como día de cielo parcialmente nublado se recomienda que sea revisado manualmente, debido a que puede estar mal clasificado. En el caso cuando se tenga una topografía representativa, es recomendable entrenar un modelo por cada localidad o lugar de uso. Esto debido a que se mostraron rendimientos menores cuando la base de datos no fue discriminada por estaciones de monitoreo.

## 8. Referencias

- [1] Iqbal M. An Introduction To Solar Radiation; 1983.
- [2] Suehrcke H, McCormick PG. The frequency distribution of instantaneous insolation values. *Solar Energy*. 1988;40(5):413-22. Available from: <https://www.sciencedirect.com/science/article/pii/0038092X88900965>.
- [3] Skartveit A, Olseth JA. The probability density and autocorrelation of short-term global and beam irradiance. *Solar Energy*. 1992 dec;49(6):477-87. Available from: <https://doi.org/10.1016%2F0038-092x%2892%2990155-4>.
- [4] Jurado M, Caridad JM, Ruiz V. Statistical distribution of the clearness index with radiation data integrated over five minute intervals. *Solar Energy*. 1995 dec;55(6):469-73. Available from: <https://doi.org/10.1016%2F0038-092x%2895%2900067-2>.
- [5] Maafi A, Harrouni S. Preliminary results of the fractal classification of daily solar irradiances. *Solar Energy*. 2003 jul;75(1):53-61. Available from: <https://doi.org/10.1016%2Fs0038-092x%2803%2900192-0>.
- [6] Harrouni S, Guessoum A, Maafi A. Classification of daily solar irradiation by fractional analysis of 10-min-means of solar irradiance. *Theor Appl Climatol*. 2005 sep;80(1):27-36. Available from: <https://doi.org/10.1007%2Fs00704-004-0085-0>.
- [7] Gueymard CA. Clear-sky irradiance predictions for solar resource mapping and large-scale applications: Improved validation methodology and detailed performance analysis of 18 broadband radiative models. *Solar Energy*. 2012 aug;86(8):2145-69. Available from: <https://doi.org/10.1016%2Fj.solener.2011.11.011>.
- [8] Pérez-Burgos A, Díez-Mediavilla M, Alonso-Tristán C, Rodríguez-Amigo MC. Analysis of solar direct irradiance models under clear-skies: Evaluation of the improvements for locally adapted models. *Journal of Renewable and Sustainable Energy*. 2017 mar;9(2):023703. Available from: <https://doi.org/10.1063%2F1.4981798>.
- [9] Ineichen P. Validation of models that estimate the clear sky global and beam solar irradiance. *Solar Energy*. 2016 jul;132:332-44. Available from: <https://doi.org/10.1016%2Fj.solener.2016.03.017>.
- [10] Hasni A, Sehli A, Draoui B, Bassou A, Amieur B. Estimating Global Solar Radiation Using Artificial Neural Network and Climate Data in the South-western Region of Algeria. *Energy Procedia*. 2012;18:531-7. Available from: <https://doi.org/10.1016%2Fj.egypro.2012.05.064>.
- [11] Kumar S, Kaur T. Efficient solar radiation estimation using cohesive artificial neural network technique with optimal synaptic weights. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*. 2019 oct;234(6):862-73. Available from: <https://doi.org/10.1177%2F0957650919878318>.
- [12] Ozgoren M, Bilgili M, Sahin B. Estimation of global solar radiation using ANN over Turkey. *Expert Systems with Applications*. 2012 apr;39(5):5043-51. Available from: <https://doi.org/10.1016%2Fj.eswa.2011.11.036>.

- [13] Sahan M, Yakut E. Estimation of monthly global solar radiation in the eastern Mediterranean region in Turkey by using artificial neural networks. EPJ Web of Conferences. 2016;128:06001. Available from: <https://doi.org/10.1051%2Fepjconf%2F201612806001>.
- [14] Ruiz-Arias JA, Gueymard CA. Worldwide inter-comparison of clear-sky solar radiation models: Consensus-based review of direct and global irradiance components simulated at the earth surface. Solar Energy. 2018 jul;168:10-29. Available from: <https://doi.org/10.1016%2Fj.solener.2018.02.008>.
- [15] Kwarikunda N, Chiguvare Z. Performance Analysis of Clear Sky Global Horizontal Irradiance Models: Simple Models Adapted for Local Conditions. Journal of Renewable Energy. 2021 sep;2021:1-12. Available from: <https://doi.org/10.1155%2F2021%2F4369959>.
- [16] Haldorai A, Ramu A. Canonical Correlation Analysis Based Hyper Basis Feedforward Neural Network Classification for Urban Sustainability. Neural Process Lett. 2020 aug;53(4):2385-401. Available from: <https://doi.org/10.1007%2Fs11063-020-10327-3>.
- [17] KangSeung-Ho, JeongIn-Seon, LimHyeong-Seok. A Method for the Classification of Water Pollutants using Machine Learning Model with Swimming Activities Videos of Caenorhabditis elegans. Journal of the Korea Institute of Information and Communication Engineering. 2021 7;25(7):903-9.
- [18] Verhaar HJM, Solbé J, Speksnijder J, van Leeuwen CJ, Hermens JLM. Classifying environmental pollutants: Part 3. External validation of the classification system. Chemosphere. 2000 apr;40(8):875-83. Available from: <https://doi.org/10.1016%2Fs0045-6535%2899%2900317-3>.
- [19] Rumelhart DE, McClelland JL. In: A General Framework for Parallel Distributed Processing; 1987. p. 45-76.
- [20] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review. 1958;65(6):386-408. Available from: <https://doi.org/10.1037%2Fh0042519>.
- [21] Isaksson M. Create a drawing of a feed-forward neural network.; 2021. Available from: <https://github.com/martisak/dotnets>.
- [22] Amidi S, Amidi A. Recurrent Neural Networks cheatsheet;. Available from: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks>.
- [23] Unser M, Thevenaz P, Yaroslavsky L. Convolution-based interpolation for fast, high-quality rotation of images. IEEE Transactions on Image Processing. 1995;4(10):1371-81.
- [24] Boellaard R, van Herk M, Mijnheer BJ. A convolution model to convert transmission dose images to exit dose distributions. Med Phys. 1997 feb;24(2):189-99. Available from: <https://doi.org/10.1118%2F1.598066>.
- [25] Gao H, Choi HF, Claus P, Boonen S, Jaecques S, van Lenthe GH, et al. A fast convolution-based methodology to simulate 2-D/3-D cardiac ultrasound images. IEEE Trans Ultrason, Ferroelectr, Freq Contr Transactions on Ultrasonics, Ferroelectrics and Frequency Control. 2009 feb;56(2):404-9. Available from: <https://doi.org/10.1109%2Ftuffc.2009.1051>.
- [26] Jauregui Fernández A. Qué son y cómo crear una red neuronal convolucional con Keras;. Available from: <https://anderfernandez.com/blog/que-es-una-red-neuronal-convolucional-y-como-crearlaen-keras/>.
- [27] Lin M, Chen Q, Yan S. Network In Network; 2013. Available from: <https://arxiv.org/abs/1312.4400v3>.
- [28] Qu L, Wu C, Zou L. 3D Dense Separated Convolution Module for Volumetric Medical Image Analysis. Applied Sciences. 2020 jan;10(2):485. Available from: <https://doi.org/10.3390%2Fapp10020485>.

- [29] Rahuljha. LSTM Gradients; 2020. Available from: <https://towardsdatascience.com/lstm-gradients-b3996e6a0296>.
- [30] Deep Dive into Bidirectional LSTM; 2019. Available from: <https://www.i2tutorials.com/deep-dive-into-bidirectional-lstm/>.
- [31] Chen Y, Kalantidis Y, Li J, Yan S, Feng J. *A<sup>2</sup>-Nets: Double Attention Networks*; 2018.
- [32] Chen Y, Fan H, Xu B, Yan Z, Kalantidis Y, Rohrbach M, et al.. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution; 2019.