ASADES

Avances en Energías Renovables y Medio Ambiente - AVERMA
Vol. 29, pp. 568-575 2025, ISSN 2796-8111

EVALUACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA LA CLASIFICACIÓN DE DÍAS DE CIELO CLARO

Pablo Cinco Reynaga[1], **Rubén Darío Ledesma**[1,2]
[1]Facultad de Ciencias Exactas, Universidad Nacional de Salta (UNSa)
[2]Instituto de Investigaciones en Energía No Convencional (INENCO, UNSa – CONICET)
E-mail: pablo.cinco.reynaga@gmail.com, rdledesma@exa.unsa.edu.ar

RESUMEN: El siguiente trabajo busca aplicar y comparar diferentes algoritmos de aprendizaje automático al problema de identificar días de cielo claro en los conjuntos de datos de radiación solar, con el objetivo de evaluar su desempeño. Se utilizará un conjunto de datos de la localidad de Cerrillos, Salta, Argentina para entrenar diferentes algoritmos de aprendizaje supervisado (Perceptrón multicapa y K-vecinos más cercanos), implementando técnicas de reducción de dimensión (Análisis de componen- tes principales), estimación de hiperparámetros (gridsearch) y, por último, se contrastarán los resultados mediante una matriz de confusión utilizando la precisión como métrica de desempeño.

Palabras clave: Irradiancia solar, GHI, Aprendizaje Automático, MLP, KNN

INTRODUCCIÓN

La medición de la irradiancia bajo condiciones de cielo despejado es fundamental para el análisis de los procesos radiativos en la atmósfera. Este tipo de datos permite eliminar la influencia de las nubes en el balance radiativo y evaluar con mayor precisión el impacto de gases y aerosoles. Analizar estos procesos en ausencia de nubosidad resulta clave para comprender su papel en el sistema climático. Sin embargo, esta tarea sigue siendo un desafío: no suele existir un sistema que registre de manera separada las observaciones en condiciones de cielo despejado, ni un método estandarizado para derivar dichas irradiancias a partir de mediciones realizadas bajo cielo total (Correa et al., 2022).

Uno de los principales obstáculos es la propia definición de lo que constituye un “día de cielo claro” o “día soleado”. En la comunidad científica dedicada a la radiación solar, no hay consenso: cada autor puede proponer criterios diferentes. Según (Gueymard et al., 2019), las metodologías para detectar cielo claro se dividen, en términos generales, en dos enfoques: la detección de días completamente libres de nubes y la detección de momentos con el sol despejado. El presente trabajo se enmarca en el primer enfoque, centrado en identificar días sin nubosidad.

En este contexto, se plantea evaluar el desempeño de algunos algoritmos de aprendizaje automático para clasificar si los datos correspondientes a un día determinado representan efectivamente un día de cielo claro. Los modelos seleccionados para esta tarea son el Perceptrón Multicapa (MLP, por sus siglas en inglés) y el algoritmo de k vecinos más cercanos (KNN).

Aunque una validación robusta requeriría datos provenientes de múltiples estaciones, en este estudio se utiliza la información de una sola. Esto es suficiente para obtener una primera aproximación al comportamiento de estos algoritmos frente a datos de irradiancia en condiciones de cielo claro.

PRE-PROCESAMIENTO DE DATOS

Descripción de los datos

Los datos utilizados corresponden a mediciones de radiación global en el plano horizontal (GHI, por sus siglas en inglés), expresadas en W/m², registradas a intervalos de un minuto. Las observaciones provienen de una estación meteorológica del Instituto Nacional de Tecnología Agropecuaria (INTA) enmarcadas en el proyecto ENARSOL (Moltoni et al., 2016), ubicada en la localidad de Cerrillos, provincia de Salta, Argentina, con coordenadas geográficas 24.89° S de latitud, 65.82° O de longitud, a una altitud de 1.235 m sobre el nivel del mar y se encuentra dentro de la clasificación climática Cwb (Beck et al., 2018).

La serie temporal está conformada por 1.875.636 puntos, desde 2016 hasta 2020. 945.791 son datos diurnos (50,4 %), de los cuales 209.023 (11,1 %) son datos faltantes. Se puede apreciar la distribución de los datos en la Figura 1.

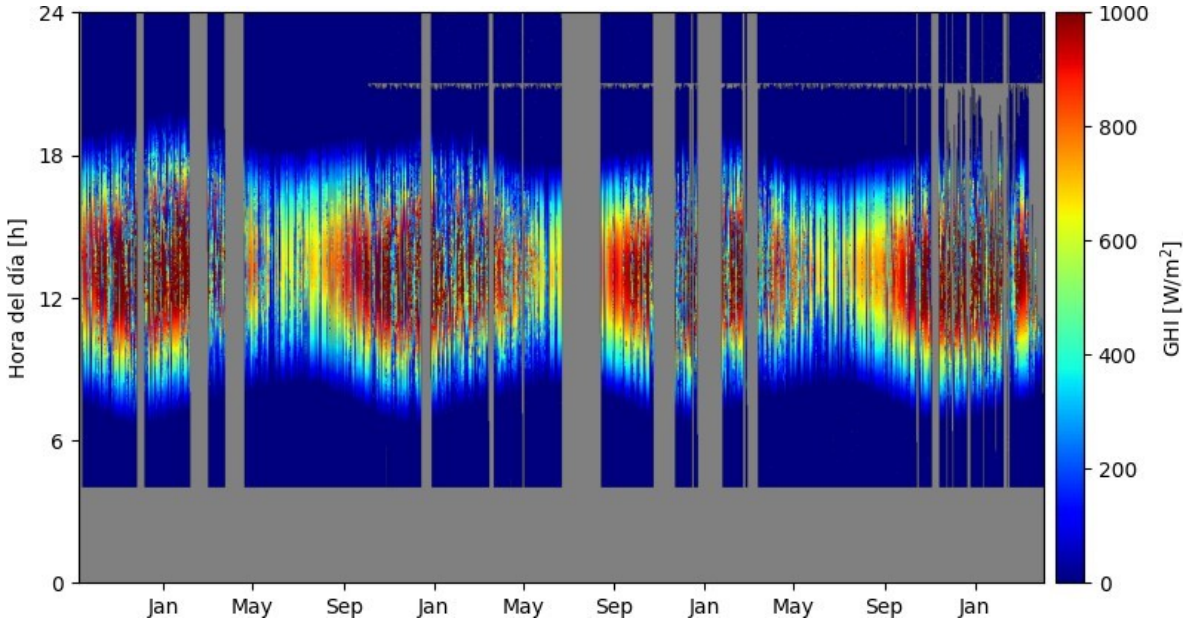

Figura 1: Datos de GHI. Cerrillos, Salta, Argentina. Periodo 2016-2020.

Evaluación de la calidad de los datos

Los controles de calidad permiten tener una noción de la confiabilidad de las bases de datos. Se utilizaron los filtros recomendados por la Baseline Surface Radiation Network (BSRN), según como lo detallan en el trabajo de (Long y Dutton, 2010). Particularmente se utilizarán los filtros "físicamente posibles" y "extremadamente raros". Como resultado 0 datos y 67 (<1 %) fueron filtrados por el primer y segundo filtro respectivamente.

Conjunto de entrenamiento

Para el procesamiento de la información, cada día se considera como una instancia o registro, y cada uno de los 1.440 minutos que lo componen se emplea como un atributo, tal como se presenta en la Tabla 1. Se eligió mantener todos los datos del día, incluyendo los nocturnos, a los aspectos de mantener fija la arquitectura de la red neuronal. Adicionalmente se tiene la hipótesis que sea factible encontrar desplazamientos horarios en la serie temporal, de esta manera se puede identificar si la serie de medidas está desplazada.

Etiquetas

Para generar las etiquetas de la base de datos, se determinó qué días correspondían a condiciones de cielo claro. Para ello, se solicitó a tres expertos en solarimetría que clasificaran cada día en una de tres categorías: "claro", "no claro" o "dudoso", mediante la inspección visual. Se le proporcionó gráficas de cada día con datos medidos de GHI y datos modelados de cielo claro de GHI como se puede ver los ejemplos de las Figuras 2 y 3, así como lineamientos para la inspección visual recomendados por (Abal et al., 2020). Los datos modelados fueron generados por el modelo ARGP (Ledesma et al., 2023).

Tabla 1: Ejemplo del conjunto de datos de entrenamiento. Cada fila contiene las mediciones de GHI de un día. Los valores están expresados en w/m^2.

	00:00	00:01	...	11:59	12:00	12:01	...	23:58	23:59
2016-09-08	nan	nan	...	nan	nan	nan	...	nan	nan
2016-09-09	nan	nan	...	786,02	802,51	780,42	...	nan	nan
2016-09-10	3,68	3,68	...	850,36	854,86	854,12	...	3,68	3,68
2016-09-11	3,68	3,68	...	875,99	877,09	879,81	...	3,68	3,68
...	...	...	...	...	...	...	...	...	...
2020-03-31	3,68	3,68	...	474,88	500,65	526,41	...	3,68	3,68
2020-04-01	3,68	3,68	...	356,55	nan	nan	...	nan	3,68
2020-04-02	nan	3,68	...	890,86	887,03	883,50	...	3,68	3,68
2020-04-03	nan	nan	...	nan	nan	nan	...	nan	nan

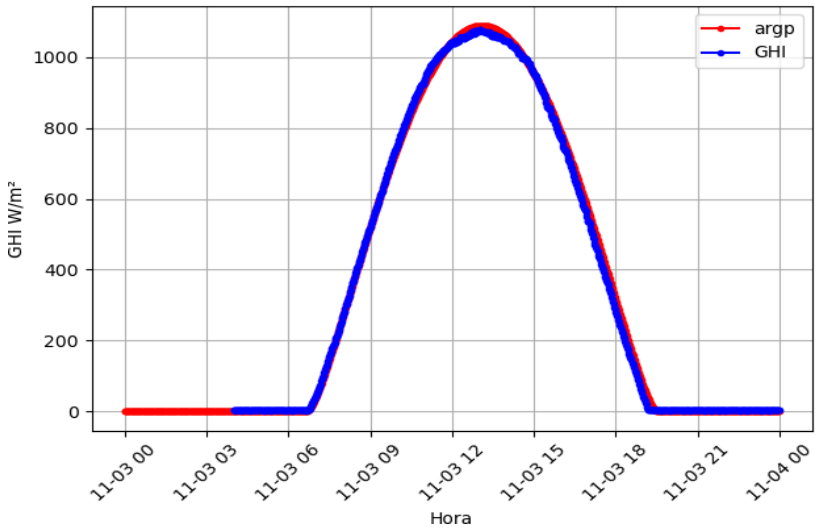

Figura 2: Ejemplo de gráfica de cielo claro usado para generar etiquetas. En azul se grafica las mediciones mientras que en rojo se grafica los datos modelados de cielo claro.

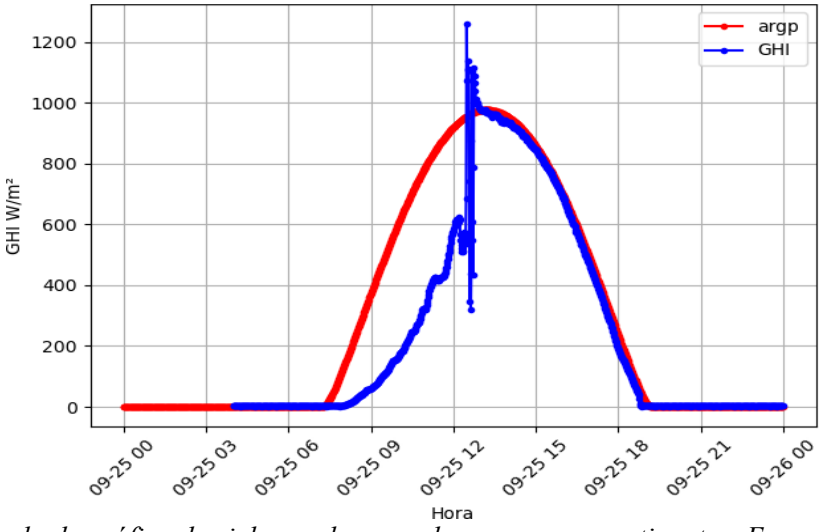

Figura 3: Ejemplo de gráfica de cielo no claro usado para generar etiquetas. En azul se grafica las mediciones mientras que en rojo se grafica los datos modelados de cielo claro

Posteriormente, los conjuntos de etiquetas generados por los expertos se combinaron para formar un único conjunto de etiquetas siguiendo las siguientes reglas:

- El día se clasifica como “no claro” si al menos uno de los expertos lo calificó así.
- El día se clasifica como “claro” si todos los expertos indicaron “dudoso” excepto uno, que lo calificó

como “claro”.
- Si todos los expertos indicaron “dudoso”, el día se clasifica como “no claro”.

Aplicando este procedimiento, se generaron 1.304 etiquetas (una por cada día del periodo analizado), transformando el problema en una tarea de clasificación binaria.

Índice de claridad k_t

A partir del conjunto de datos de GHI, se generará otro conjunto de Índice de Claridad (k_t en adelante), el cual se calcula dividiendo GHI sobre la radiación extraterrestre en el plano horizontal I_0 (Duffie y Beckman, 2013), tal como se ve en la Ec (1).

$$k_t = GHI/I_0 \quad (1)$$

Se determinará el rendimiento de los algoritmos con este conjunto también.

Estandarización de datos

Ya que todos los atributos de los conjuntos de datos representan medidas de un mismo instrumento, se considera que todas están dentro de un mismo rango de variación. Así que se considerará que no hace falta estandarizar los datos.

Métodos de reducción de dimensión

Se aplicó el análisis de componentes principales (en adelante PCA) al conjunto de datos para reducir la cantidad de variables. Así, se redujeron de 1440 columnas a 53 para los datos de GHI y 56 para los datos de k_t. Estas columnas mantienen el 95 % de la variabilidad de los datos originales. Se puede ver una comparación entre los resultados con y sin el análisis en la Tabla 2.

Se puede apreciar que la mayoría de las métricas mejoran con la reducción de dimensiones.

Tabla 2: Comparación de métricas con y sin PCA

	GHI		Kt	
	MLP	KNN	MLP	KNN
Sin PCA	Accuracy = 0.89 Recall = 0.52 Precisión = 0.50 F-Score = 0.51	Accuracy = 0.79 Recall = 0.97 Precisión = 0.34 F-Score = 0.50	Accuracy = 0.90 Recall = 0.86 Precisión = 0.53 F-Score = 0.66	Accuracy = 0.84 Recall = 0.93 Precisión = 0.40 F-Score = 0.56
Con PCA	Accuracy = 0.92 Recall = 0.86 Precisión = 0.60 F-Score = 0.70	Accuracy = 0.87 Recall = 0.93 Precisión = 0.46 F-Score = 0.61	Accuracy = 0.93 Recall = 0.76 Precisión = 0.68 F-Score = 0.73	Accuracy = 0.89 Recall = 0.83 Precisión = 0.49 F-Score = 0.62

Conjunto de entrenamiento y validación

Se separaron los datos en un conjunto de entrenamiento y en otro de validación con una relación del 80 % y 20 % del total de los datos respectivamente. La selección de los datos para cada conjunto fue aleatoria; si bien pueden existir relaciones en la estacionalidad de los días, se desestimó a los aspectos prácticos de simplificar el problema.

MODELOS

Para este problema de clasificación, se seleccionaron los algoritmos de aprendizaje supervisado: Multi-layer Perceptron y K-Nearest Neighbors.

Un perceptrón multicapa (MLP) es una red neuronal artificial compuesta por varias capas de neuronas: una capa de entrada, donde se ingresan los datos, una o más capas ocultas y una capa de salida, de donde se obtiene la predicción. En este modelo, cada neurona de una capa está conectada con todas las neuronas de la capa siguiente. Los MLP se emplean en distintas aplicaciones de clasificación y regresión, y son perfectas para representar relaciones no lineales complejas entre los datos (Piccioni Costa et al., 2023).

El algoritmo K-Nearest Neighbors (KNN) es un método de aprendizaje supervisado utilizado tanto para clasificación como para regresión. Su funcionamiento se basa en la comparación de un punto de datos nuevo con los K ejemplos más cercanos en el conjunto de entrenamiento, según una métrica de distancia, comúnmente la distancia euclidiana. La predicción se realiza a partir de la clase mayoritaria (en clasificación) o del promedio de valores (en regresión) (Cunningham y Delany, 2007).

Métrica de evaluación

Ya que la salida de los algoritmos es binaria; "claro" o "no claro", se utilizará una matriz de confusión para determinar el desempeño contrastando la salida de estos con datos reales. Dicha matriz clasifica los resultados en cuatro clases: verdadero positivo, verdadero negativo, falso positivo y falso negativo. Verdadero positivo significa que el algoritmo identificó correctamente un día de cielo "claro", según los datos reales, mientras que un falso positivo quiere decir que identificó un cielo "claro" cuando en realidad era "no claro". De la misma manera, verdadero negativo y falso negativo es cuando el algoritmo identificó un día de cielo "no claro" cuando los datos reales dictaban que era un día de cielo "no claro" y "claro" respectivamente.

Una posible aplicación para el modelo es la selección de días de cielo claro para usarse en procesos de validación, aquí es necesario tener conjuntos de datos compuestos únicamente por días despejados. Por lo tanto, nos interesa maximizar la cantidad de verdaderos positivos a la vez que se minimizan los falsos positivos. Así, la métrica de evaluación será la precisión, Eq 2, tal como se describe en (Olivas Soria, 2022).

$$Presición = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos} \tag{2}$$

Esta métrica varía entre 0 y 1, donde un valor cercano a 1 nos dice que la mayor parte de los resultados arrojados por el algoritmo fueron verdaderos positivos.

Ajuste de hierparámetros

Se utilizó una técnica llamada Búsqueda en Rejilla (o Grid Search) para buscar la mejor combinación de hiperparámetros. La misma consiste en explorar sistemáticamente las diferentes combinaciones de valores para los hiperparámetros. Luego, cada combinación es evaluada respecto a nuestra métrica de interés, en este caso, la precisión, utilizando el método de la validación cruzada. La misma consiste en, para cada combinación de hiperparámetros, dividir el conjunto de datos en 5 subconjuntos, donde 4 se usaban para entrenar el algoritmo y 1 para validar, el proceso se itera 5 veces cambiando el conjunto de validación. La aplicación de este método arrojó los resultados que se pueden ver en las Tablas 3, 4, 5 y 6.

Tabla 3: Resultados de gridsearch. Modelo: KNN. Datos: GHI

Parámetros	split0_ score	split1_ score	split2_ score	split3_ score	split4_ score	mean_ score	std_ score	rank
metric: euclidean n_neighbors: 100 weights: distance	0.684	0.545	0.458	0.478	0.846	0.602	0.145	1
metric: minkowski n_neighbors: 100 weights: distance	0.684	0.545	0.458	0.478	0.846	0.602	0.145	2

metric: minkowski n_neighbors: 50 weights: distance	0.5	0.432	0.325	0.307	0.592	0.431	0.106	3

Tabla 4: Resultados de gridsearch. Modelo: KNN. Datos: Kt

Parámetros	split0_ score	split1_ score	split2_ score	split3_ score	split4_ score	mean_ score	std_ score	rank
metric: euclidean n_neighbors: 3 weights: distance	0.993	0.976	0.976	0.964	1.0	0.982	0.012	1
metric: minkowski n_neighbors: 3 weights: distance	0.993	0.976	0.976	0.964	1.0	0.982	0.012	2
metric: minkowski n_neighbors: 5 weights: distance	0.987	0.976	0.981	0.965	0.993	0.981	0.009	3

Tabla 5: Resultados. Modelo: MLP. Datos: GHI

Parámetros	split0_ score	split1_ score	split2_ score	split3_ score	split4_ score	mean_ score	std_ score	rank
activation: logistic alpha: 0.1 hidden_layer_sizes: (50,) learning_rate: constant solver: adam	0.75	0.818	1.0	0.7	1.0	0.853	0.125	1
activation: logistic alpha: 0.01 hidden_layer_sizes: (50,) learning_rate: invscaling solver: adam	0.6	1.0	0.666	0.777	1.0	0.808	0.166	2
activation: logistic alpha: 0.0001 hidden_layer_sizes: (100, 50) learning_rate: invscaling solver: lbfgs	0.666	0.722	0.625	0.769	1.0	0.756	0.131	3

Tabla 6: Resultados. Modelo: MLP. Datos:Kt

Parámetros	split0_ score	split1_ score	split2_ score	split3_ score	split4_ score	mean_ score	std_ score	rank
activation: logistic alpha: 0.1 hidden_layer_sizes: (50,) learning_rate: adaptive 'solver: lbfgs	0.988	0.972	0.988	0.994	0.994	0.987	0.007	1
activation: relu alpha: 0.1 hidden_layer_sizes: (100,) learning_rate: constant solver: lbfgs	0.988	0.983	0.983	0.988	0.988	0.986	0.002	2

activation: logistic alpha: 0.001 hidden_layer_sizes: (100,) learning_rate: adaptive solver: lbfgs	1.0	0.978	0.983	0.983	0.983	0.985	0.007	3

RESULTADOS

Utilizando los mejores parámetros determinados en la sección anterior, se utilizó el conjunto de prueba para determinar las métricas finales que se pueden ver en el cuadro 7.

Se denota una mejora sustancial con el uso de datos de k_t para modelos de KNN, logrando una precisión del $98,1\%$. Con el uso de datos de GHI, esta mejora no es tan marcada para los modelos de MLP; sin embargo, sus resultados tienen una variabilidad considerablemente menor.

Se determina que el mejor algoritmo para trabajar es el K-Vecinos más Cercanos, entrenado con datos de Índice de Claridad (k_t) con los siguientes hiperparámetros.

- métrica de distancia: euclidiana
- n de vecinos = 3
- pesos: en relación a la distancia

Hay que considerar que, debido a que el entrenamiento se realizó con un conjunto de datos perteneciente a una localización en particular, el algoritmo está adaptado a ese sitio en particular.

Por otro lado, hay que tener en cuenta que solo se usaron datos de GHI para el etiquetado de los datos. En algunos casos, las mediciones de GHI se pueden parecer mucho a las que habría en condiciones de cielo despejado si hay una cubierta de nubes que aumente la radiación difusa lo suficiente.

Tabla 7: Resultados de los modelos con el conjunto de prueba

	GHI	Kt
MLP	0.720	0.972
KNN	0.458	0.981

CONCLUSIONES

Tras aplicar los mejores hiperparámetros obtenidos mediante la búsqueda en rejilla, se evaluó el desempeño final de los modelos utilizando el conjunto de prueba. Los resultados muestran que el uso de Kt como variable de entrada mejora significativamente el rendimiento, especialmente en el algoritmo KNN, que alcanzó una precisión del 98,1 %. En comparación, con datos de GHI, la mejora en el MLP no fue tan pronunciada, aunque el modelo presentó una menor variabilidad en sus resultados.

El análisis permite concluir que el modelo con mejor desempeño para esta tarea es el KNN entrenado con datos de Kt, configurado con métrica de distancia euclidiana, k=3 vecinos y pesos inversamente proporcionales a la distancia. Sin embargo, debe tenerse en cuenta que el entrenamiento se realizó con datos de una única ubicación, lo que hace que el modelo esté adaptado específicamente a las condiciones locales de Cerrillos, Salta.

Asimismo, se debe considerar que el etiquetado de los datos se basó únicamente en mediciones de GHI. En ciertos casos, condiciones de nubosidad pueden generar valores de GHI similares a los de cielo despejado, debido al aumento de la radiación difusa, lo que podría inducir a errores en la clasificación.

En síntesis, los resultados evidencian el potencial de los modelos de aprendizaje supervisado para la identificación de días de cielo claro, destacando la importancia de la selección de variables y el ajuste de hiperparámetros en la optimización del rendimiento.

REFERENCIAS

Abal, G., Alonso-Suárez, R., y Laguarda, A. (2020). *Radiación Solar: Notas del curso Fundamentos del Recurso Solar*. Laboratorio de Energía Solar, Uruguay, versión 4.0 edición.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., y Wood, E. F. (2018). Present and future köppen-geiger climate classification maps at 1-km resolution. *Scientific Data*, 5(1):180214.

Correa, L., Folini, D., Chtirkova, B., y Wild, M. (2022). A method for clear sky identification and long term trends assessment using daily surface solar radiation records. *Earth and Space Science*, 9.

Cunningham, P. y Delany, S. (2007). k-nearest neighbour classifiers. *Mult Classif Syst*, 54.

Duffie, J. y Beckman, W. (2013). *Solar Engineering of Thermal Processes*. Ingenieria de la Energia. Wiley.

Gueymard, C. A., Bright, J. M., Lingfors, D., Habte, A., y Sengupta, M. (2019). A posteriori clear-sky identification methods in solar irradiance time series: Review and preliminary validation using sky imagers. *Renew. Sustain. Energy Rev.*, 109:412–427.

Ledesma, R. D., Salazar, G. A., y de Castro Vilela, O. (2023). Argp-v2 un modelo práctico para la estimación de irradiancia global horizontal en condiciones de cielo claro para sitios de altura. *Avances en Energías Renovables y Medio Ambiente - AVERMA*, 26:283–289.

Long, C. N. y Dutton, E. G. (2010). Bsrn global network recommended qc tests, v2. x.

Moltoni, A., Clemares, N., Gorandi, E., y Moltoni, L. (2016). Enarsol. red de medición de radiación solar interconectada. En *III Congreso Argentino de Ingeniería y IX Congreso Argentino de la Enseñanza en Ingeniería*, Resistencia, Argentina.

Olivas Soria, E. (2022). *Inteligencia artificial: Casos prácticos con aprendizaje profundo*. Ediciones de la U.

Piccioni Costa, L., Guerreiro, M., Puchta, E., Tadano, Y., Antonini Alves, T., Kaster, M., y Siqueira, H. (2023). *Multilayer Perceptron*, p. 105.

ASSESSMENT OF MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION OF CLEAR SKY DAYS

ABSTRACT: The following work seeks to apply and compare different machine learning algorithms to the problem of identifying clear-sky days in solar radiation datasets, with the aim of evaluating their performance. A dataset from the town of Cerrillos, Salta, Argentina will be used to train different supervised learning algorithms (Multilayer Perceptron and K-Nearest Neighbors), implementing dimension reduction techniques (Principal Component Analysis), hyperparameter estimation (GridSearch), and finally, the results will be contrasted using a confusion matrix using precision as a performance metric.

Keywords: Solar Irradiance, GHI, Machine Learning, MLP, KNN