Artículos

Metodología para la evaluación de desempeño de plantas solares fotovoltaicas a través del uso de la ciencia de datos

Methodology for evaluating the performance of photovoltaic solar plants through the use of data science

Yajure-Ramírez, César A.; Rojas-Aranguren, Jairo J.

César A. Yajure-Ramírez
cyajure@gmail.com
Universidad Central de Venezuela, Venezuela
Jairo J. Rojas-Aranguren
jairo.rojas@rcenergia.com
R&C Ingeniería y Servicios SpA, Chile

Revista Tecnológica ESPOL - RTE
Escuela Superior Politécnica del Litoral, Ecuador
ISSN: 0257-1749
ISSN-e: 1390-3659
Periodicidad: Semestral
vol. 35, núm. 1, 2023
rte@espol.edu.ec

Recepción: 17 Febrero 2023
Aprobación: 11 Mayo 2023

URL: http://portal.amelica.org/ameli/journal/844/8444932007/

DOI: https://doi.org/10.37815/rte.v35n1.1011

Resumen: La evaluación continua de las plantas solares fotovoltaicas es fundamental para su operación, puesto que, se debe hacer seguimiento a sus variables principales, y así verificar que se entrega la energía eléctrica en óptimas condiciones de operación y de eficiencia. En esta investigación se presentó una metodología basada en la ciencia de datos con el fin de evaluar plantas solares fotovoltaicas. Se aplicó al conjunto de datos de una planta solar del Laboratorio Nacional de Energías Renovables de EEUU, haciendo un análisis de los datos para obtener curvas temporales de irradiancia y energía, y también de los principales indicadores de desempeño. Así también, se empleó el algoritmo K-Means para generar clústers dentro del conjunto de datos, y el algoritmo K-NN para crear modelos de predicción de clases de la energía y del indicador PR. Se obtuvieron clústers que agrupan los valores de potencia generada, y los valores del PR. El modelo de clasificación de las clases de energía tuvo una exactitud del 91,67%, mientras que el modelo de clasificación de las clases del indicador PR tuvo una exactitud del 83,33%. Dado que la tasa de ensuciamiento promedio en las escalas mensual y anual estuvo por encima del 90%, mientras que las del PR estuvieron alrededor del 70%, se recomienda hacer un estudio para determinar el origen de las pérdidas en la planta. Asimismo, se sugiere realizar un modelo para determinar el impacto de la temperatura ambiente, la temperatura del módulo fotovoltaico, y de la velocidad del viento en la producción de energía eléctrica.

Palabras clave: Energía eléctrica, K-Means, K-NN, tasa de desempeño, tasa de suciedad.

Abstract: The continuous evaluation of solar photovoltaic plants is essential for their operation. Their main variables must be monitored to verify that the electrical energy is delivered under optimal operating and efficiency conditions. This research presents a methodology based on data science to evaluate solar photovoltaic plants. This methodology was applied to the data set of a solar plant of the US National Renewable Energy Laboratory, analyzing the data to obtain temporal curves of irradiance and energy, as well as the leading performance indicators. Also, this study used the K-Means algorithm to generate clusters within the data set and the K-NN algorithm to create class prediction models of the energy and PR indicator. Clusters grouping the generated power values and the PR values

were obtained. The energy class classification model had an accuracy of 91.67%, while the PR indicator class classification model had an accuracy of 83.33%. Since the average fouling rate in the monthly and annual scales was above 90%, while those of the PR were around 70%, a study is recommended to determine the origin of the losses in the plant. It is also suggested that a model be developed to determine the impact of ambient temperature, PV module temperature, and wind speed on of electric power production.

Keywords: Electrical energy, K-Means, K-NN, performance ratio, soiling rate.

Introducción

En la sociedad actual, el uso de la energía eléctrica es tan común que cuesta imaginarse un mundo sin energía eléctrica. Desde finales del siglo XIX hasta el presente se han desarrollado distintas tecnologías para la producción de electricidad y, poco a poco, la tendencia ha sido utilizar fuentes de energía cada vez menos contaminantes y que minimicen el impacto ambiental debido a su uso. En ese sentido han surgido distintas tecnologías para captar energía desde fuentes renovables, entre ellas se encuentran aquellas provenientes del sol Usualmente se utilizan dos formas para captar esta energía solar, una es el método indirecto en el que esta energía se aprovecha para calentar un fluido, convertirlo a vapor, y hacer que mueva una turbina conectada a través de su eje con un generador eléctrico. El otro método es el directo, llamado sistema solar fotovoltaico y, según Yahyaoui (2018), este sistema "convierte la radiación solar en electricidad de manera directa", mediante el uso de celdas solares.

De acuerdo con ABB (2019), los sistemas de generación eléctrica del tipo solar fotovoltaico tienen como elemento principal el arreglo de paneles solares, a través del cual se genera energía eléctrica en corriente continua. Debido a que la mayoría de las cargas eléctricas consumen corriente alterna, o porque el sistema solar fotovoltaico se vaya a conectar a la red eléctrica externa en corriente alterna, la salida del arreglo de paneles debe conectarse a un inversor para transformar la energía de corriente continua a corriente alterna. Adicionalmente, podría poseer otros componentes tales como: reguladores de carga, contadores, transformadores, entre otros.

Ahora bien, tal como cualquier sistema de producción de energía, este tipo de sistemas requiere ser monitoreado, tomando las mediciones necesarias para determinar las pérdidas del sistema (Asolmex, 2018), y aplicando distintos indicadores, para de esta manera determinar y evaluar su desempeño. En ese sentido, el objetivo de esta investigación consiste en presentar y aplicar una metodología para evaluar plantas solares fotovoltaicas a través del uso de la ciencia de datos y los indicadores claves propuestos en la normativa vigente. Específicamente, en la etapa de modelación del proceso de ciencia de datos, se trabaja con el algoritmo de agrupamiento K-Means para generar clústers de registros de datos con características similares asociadas a la tasa de desempeño de la planta (PR: Performance Ratio), y a la energía eléctrica producida por la planta. Además, se aplica el algoritmo de clasificación K-NN para predecir el rango de valores de la tasa de desempeño y de la energía eléctrica producida.

En la revisión del estado del arte se encontraron varias investigaciones relacionadas con el tema de este trabajo, la mayoría con plantas solares fotovoltaicas en el orden de los megavatios. Por ejemplo, Verma et al. (2021) desarrollan la evaluación de desempeño de tres plantas solares fotovoltaicas ubicadas en regiones semiáridas con climas secos y cálidos, y ofrecen sugerencias para mejorar su eficiencia. Compararon los valores de desempeño obtenidos con valores simulados y observaron un mejor desempeño durante los meses de

marzo, abril, y mayo. Sugieren mantener un proceso de limpieza de polvo frecuente mientras esté operativa la planta.

Así también, en la investigación de Tackie & Cemal (2022) se realiza la evaluación de desempeño y estudios de viabilidad de una planta solar fotovoltaica ubicada en el norte de Chipre. Además del indicador tasa de desempeño PR, utilizan la eficiencia específica, el factor de planta FC, e indicadores de inversión. Como resultado obtienen una tasa de desempeño del 85,77%, un factor de planta de 17,71%, y que la producción de energía aumentaría un 27,88% sí se instalara en la planta un sistema de seguimiento.

En su investigación, Veerendra et al. (2022) realizan la evaluación de desempeño de una planta solar de 1,1 MW, ubicada en Louisiana, con distintos tipos de tecnologías de paneles solares, utilizando los indicadores PR, FC, y eficiencia del sistema. Concluyen que el sector de la planta con paneles de seleniuro de cobre, indio y galio (CIGS) tiene una mejor PR de 79% en comparación con el sector con paneles de silicio monocristalino y el silicio policristalino, que tienen PR de 77% y 73%, respectivamente.

Asimismo, Romero et al. (2019) realizan un análisis de desempeño de plantas con distintas tecnologías de paneles solares. Una planta de 3,3 kWp con paneles de silicio monocristalino ubicada en Arequipa, otra de 3,3 kWp también con paneles de silicio monocristalino situada en Tacna, una de 3 kWp con paneles de silicio policristalino localizada en Lima, y una última planta de 3,5 kWp con paneles con hetero unión de silicio amorfo/silicio cristalino ubicada también en Lima. Los rendimientos finales anuales obtenidos se encuentran entre 1770 y 1992 kWh/kW, entre 1505 y 1540 kWh/kW y entre 736 y 833 kWh/kW para Arequipa, Tacna y Lima, respectivamente, mientras que el rendimiento energético anual del conjunto fotovoltaico alcanzado por la planta con hetero unión es 1338 kWh/kW. El PR anual se mantiene en torno a 0,83 para las plantas en Arequipa y Tacna mientras que este parámetro oscila entre 70% y 77% para la planta de silicio monocristalino en Lima, y un valor de 97% para la segunda planta ubicada en Lima.

De igual manera, Nugroho y Sudiarto (2020) desarrollaron un estudio para evaluar una planta solar fotovoltaica en Indonesia empleando los datos del sistema de adquisición de datos desde marzo del 2016 hasta diciembre del 2019 para calcular el indicador PR diario, para el que obtuvieron valores entre 70% y 90%. Por otra parte, Vasisht et al. (2016) analizan el desempeño de una planta solar ubicada en el techo de una universidad en la India y determinan los efectos estacionales en la producción de energía eléctrica. Utilizan los indicadores de PR y FC para la evaluación de la planta y obtienen un FC del 16,5% y un PR de alrededor del 85%, pero adicionalmente determinan que el PR de la planta está correlacionado con la temperatura en las distintas épocas del año.

León-Ospina et al. (2023) desarrollan la evaluación del desempeño de proyectos fotovoltaicos ubicados en Latinoamérica, pero utilizando indicadores económicos-financieros. Por último, Ahire et al. (2018) realizaron el análisis de desempeño de una planta solar fotovoltaica de 10 kWp, usando un software de simulación de sistemas fotovoltaicos. Utilizan como principal indicador el PR, para el cual se obtuvieron valores mensuales alrededor del 78%.

El resto del artículo se distribuye de la siguiente manera. La sección 2 corresponde a la presentación de la metodología utilizada, en la sección 3 se presentan y discuten los resultados obtenidos y, finalmente, se presentan las conclusiones y recomendaciones que se derivan de la presente investigación.

Materiales y Métodos

Para la evaluación de desempeño de plantas solares, se han propuesto una serie de indicadores clave típicos; tales como el rendimiento de referencia, el rendimiento específico y la relación de desempeño. Estos indicadores clave de desempeño fueron propuestos por la Comisión Electrotécnica Internacional (IEC por sus siglas en inglés) en su documento de especificaciones técnicas para el desempeño de sistemas fotovoltaicos (IEC, 2016). En cuanto a la relación de desempeño, este representó el nivel de calidad de una planta solar fotovoltaica para un período de tiempo determinado, y se obtiene de la división entre el rendimiento

específico y el rendimiento de referencia, obtenidos para dicho período. Adicionalmente, se consideraron importantes los valores de energía eléctrica AC a la salida de la planta, los valores de irradiancia solar, la tasa de ensuciamiento de la planta, así como las variables climáticas de temperatura ambiente y velocidad del viento.

La metodología propuesta en esta investigación consistió en tomar los datos del sistema de adquisición de datos de la planta solar fotovoltaica para evaluar la planta de acuerdo con los indicadores mencionados en el párrafo anterior, aplicando los pasos o etapas de un proceso de ciencia de datos presentados por Cielen et. al. (2016). En ese sentido, la Ciencia de Datos no es más que el uso de técnicas especializadas para el análisis de grandes cantidades de datos con el fin de extraer conocimientos significativos de ellos. Entonces, en primer lugar, se fijaron los objetivos de la investigación, acción que requiere un conocimiento adecuado del negocio, en este caso, del funcionamiento y operación de las plantas solares fotovoltaicas. Seguidamente, se obtuvieron los datos necesarios para llevar a cabo la investigación. Los datos requeridos para obtener los indicadores de desempeño, así como la energía eléctrica de salida, y las variables climáticas provienen del sistema de medición que normalmente se instala en la planta solar fotovoltaica, lo que permitió hacer el seguimiento correspondiente. Como tercer paso estuvo la preparación de los datos, el cual incluyó la corrección de datos faltantes, datos duplicados, y/o datos atípicos. Además, se requirió de la transformación de datos, o la combinación de estos. Por ejemplo, con las mediciones básicas de irradiancia y potencia eléctrica ac, se podrían obtener las variables necesarias para calcular los datos de energía eléctrica y la tasa de desempeño de la planta.

El siguiente paso consistió en desarrollar un análisis exploratorio de los datos, utilizando técnicas analíticas estadísticas, así como también, técnicas gráficas. Fueron de interés, por ejemplo, las curvas horarias, diarias, y mensuales de generación de energía, así como las curvas mensuales y anuales de la tasa de desempeño, y las curvas de la tasa de ensuciamiento. Entonces, hasta este punto de la metodología se pudo tener suficiente conocimiento de la instalación fotovoltaica y sus datos, como para seleccionar y aplicar los algoritmos de aprendizaje automático adecuados, y así generar los modelos que permitan el alcance de los objetivos. En esta etapa de modelación, se ejecutaron, evaluaron y compararon los distintos modelos, y los resultados obtenidos se utilizaron para la etapa final de toma de decisiones.

Por lo general, los pasos del proceso de ciencia de datos no se aplican de manera lineal. En algunas aplicaciones será necesario sólo aplicar hasta el análisis exploratorio de los datos para lograr el conocimiento deseado, y no requerirse la etapa de modelación. Asimismo, en algunas ocasiones, pudiera haber necesidad de volver a etapas anteriores desde cualquier punto para obtener mejores datos, si los resultados obtenidos hasta ese momento así lo ameritan, o desde la modelación hacia el análisis exploratorio, entre otras posibilidades. En esta investigación se aplican todas las etapas del proceso desde el establecimiento del objetivo de investigación hasta la modelación de los datos para la toma de decisiones. Es así como, en esta sección se presenta la etapa de obtención de los datos, y la etapa de preparación de los datos, mientras que las restantes etapas se presentan en la siguiente sección de la investigación.

Obtención de los datos

Los datos utilizados en esta investigación provinieron del Laboratorio Nacional de Energías Renovables de los Estados Unidos (NREL por sus siglas en inglés). Específicamente, se tomaron de la página web del conjunto de datos públicos del sistema de adquisición de datos del NREL (PVDAQ, 2023). La información corresponde al sistema de adquisición de datos de las instalaciones fotovoltaicas del NREL ubicadas en Colorado, Estados Unidos. La planta está compuesta de cinco paneles solares de mono silicio, de 200 vatios pico cada uno (SolarDesignTool, 2023), en un montaje fijo, con 40° de inclinación, y ángulo azimut de 180°. Los datos corresponden a mediciones minutales de potencia de salida de la planta (“ac_power”), en vatios, temperatura ambiente (“ambient_temp”) en grados Celsius, irradiancia (“poa_irradiance”) en vatios por metro cuadrado, velocidad del viento en metros por segundo (“wind_speed”), y tasa de ensuciamiento

("soiling"). Las mediciones minutales iniciaron el 25 de febrero del 2010 y culminaron el 13 de diciembre del 2016, para un total de 1.558.875 filas (registros o instancias).

Preparación de los datos

En esta etapa, se aplicaron las técnicas sugeridas por McKinney (2018), las cuales incluyen revisión de los datos con el fin de detectar posibles datos faltantes o filas duplicadas, transformación de datos, combinación de columnas de datos, y verificación del formato adecuado para las distintas variables. Se puede decir que no se detectaron filas duplicadas. Sin embargo, se detectaron siete (7) datos faltantes en la variable de temperatura ambiente ("ambient_temp"), y 17.362 datos faltantes en la variable de velocidad del viento ("wind_speed"). Los datos faltantes de temperatura corresponden a menos del 0,01% de las filas totales, mientras que los de velocidad corresponden a aproximadamente el 1,11% de las filas totales y, no obstante, al ser porcentajes bajos, se optó por imputarlos con el valor medio de los tres datos más cercanos al dato faltante.

Adicionalmente, se crearon nuevas columnas de datos, correspondientes a nuevas variables. A partir de la columna de la fecha, se elaboraron columnas correspondientes al año, mes, día y hora de lectura de los datos. De igual manera, a partir de la columna de potencia, se elaboró la columna de energía, y a partir de la columna de irradiancia, se creó la de irradiación.

Se pudo detectar que para el año 2010, no existían registros para los meses de enero, septiembre y octubre, lo que podría perturbar los resultados del análisis exploratorio. En consecuencia, se eliminaron los datos del año 2010, y se hizo el análisis exploratorio con los datos de los restantes años, es decir, desde el año 2011 hasta el año 2016.

Resultados y Discusión

En esta sección se presentan los resultados obtenidos al desarrollar las etapas de análisis exploratorio de los datos, y de modelación de los datos.

Análisis exploratorio de los datos

Luego de la preparación de los datos quedó un conjunto de datos de lecturas minutales con 1.429.678 registros (filas), y 12 variables (columnas). Se hizo un análisis de correlación entre las variables, pero sin incluir las variables temporales ni las variables calculadas. Se consideró el método tradicional de Pearson, pero también el método de Spearman, y el de Kendall, ya que de acuerdo con Amat (2023), el coeficiente de Pearson aplica para datos que están normalmente distribuidos, mientras que los otros dos métodos, estadístico Rho de Spearman o el estadístico Tau de Kendall, son convenientes cuando los datos no siguen distribución alguna. En esta investigación los resultados fueron similares al comparar los tres métodos, por lo cual, en la Figura 1 se presentan para el caso tradicional de Pearson.

Los valores de correlación varían entre 0 y 1, mientras más cerca de 1 significa que el par de variables están altamente correlacionadas entre sí, pero si el valor se acerca a 0, representa que la correlación entre las variables es baja. Entonces, en la Figura 1 se puede ver que la irradiancia y la potencia eléctrica ac están altamente correlacionadas en sentido positivo, es decir, cuando una varía su valor, la otra variable también varía en el mismo sentido y casi con la misma magnitud. No se observa algún otro par de variables con valores de correlación significativos, puesto que, según lo que indica Ratner (2017), si los valores absolutos de correlación varían entre 0 y 0,3 hay una relación débil entre las variables, si está entre 0,3 y 0,7 la relación es moderada, y si se encuentra entre 0,7 y 1, la relación es fuerte.

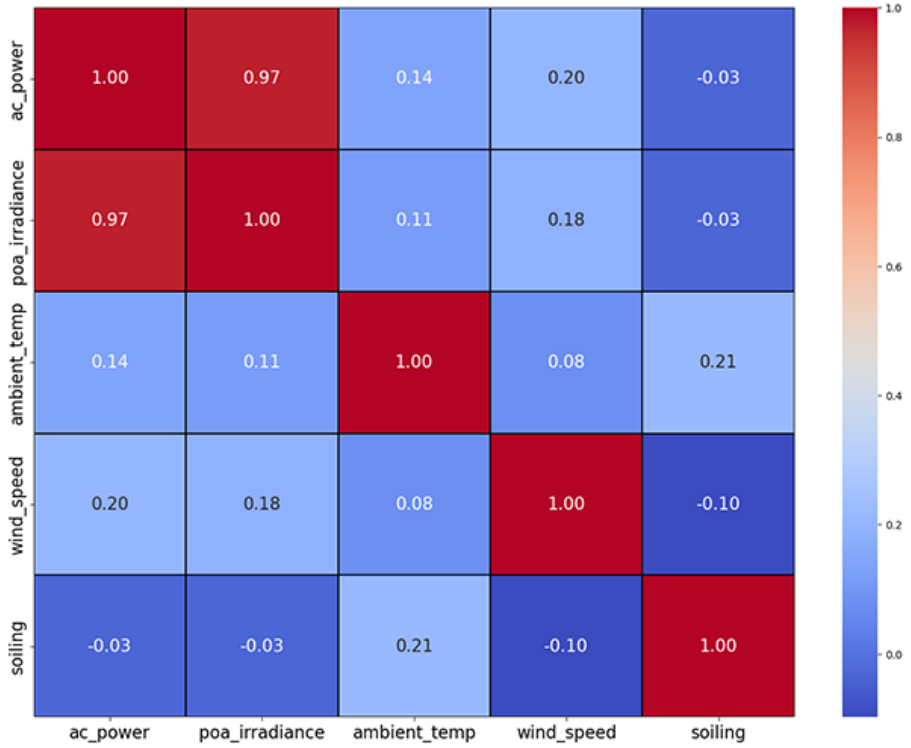

FIGURA 1
Matriz de correlación del conjunto de datos

Seguidamente se desarrolló un resumen estadístico descriptivo de los datos minutales, correspondientes a un total de 1.429.678 registros o filas, sin considerar las variables temporales. Esta información se presenta en la Tabla 1, en la que se puede observar que a excepción de la tasa de ensuciamiento ("soiling"), todas las demás variables tuvieron una alta variabilidad con respecto a su valor medio. También se puede ver, que todas las variables presentan un valor medio relativamente cercano a su mediana.

TABLA 1
Resumen descriptivo del conjunto de datos

Parámetro	ac_power	poa_irradiance	ambient_temp	wind_speed	soiling	ac_energy	irradiation
Media	374,88	483,55	15,52	1,77	0,96	6,25	8,06
DesvStd	301,26	368,68	9,79	1,20	0,04	5,02	6,14
Mínimo	-1,68	-8,36	-25,31	-0,11	0,76	-0,03	-0,14
1er. Cuartil	82,70	136,97	8,51	0,94	0,94	1,38	2,28
Mediana	314,67	406,62	16,28	1,51	0,98	5,24	6,78
3er. Cuartil	666,12	832,32	23,23	2,30	0,99	11,10	13,87
Máximo	1210,80	1646,61	68,57	14,65	1,00	20,18	27,44

Posteriormente, se desarrollaron curvas temporales con los valores promedios de la irradiancia en la planta, así como de la potencia AC generada. En la Figura 2 se presenta la curva horaria de irradiancia y de potencia AC, y de la misma se puede ver que tanto la irradiancia como la potencia tienen sus valores máximos alrededor de las 11 am, y la generación de potencia eléctrica ocurre principalmente entre las 6 am y las 5 pm. Lo anterior implica que la energía eléctrica entregada estará disponible durante ese mismo período del día. En cada punto

de la curva se presenta el intervalo de confianza respectivo, siendo la línea continua el valor promedio, y la parte sombreada son las bandas del intervalo de confianza.

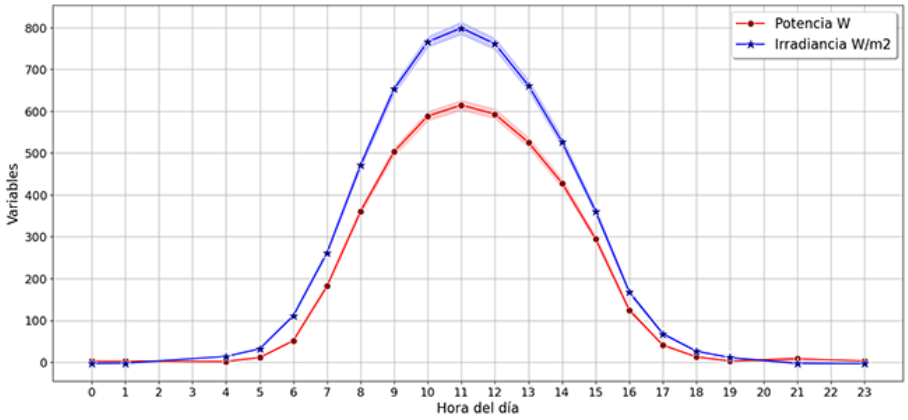

FIGURA 2
Curva horaria de irradiancia y potencia generada AC

En la Figura 3 se presenta la curva diaria de los valores promedios de irradiancia y potencia AC, con sus respectivos intervalos de confianza. Se puede ver que el valor promedio diario de la potencia se encuentra entre un valor superior a 300 W, pero menor a 375 W, siendo su valor mínimo el del día 15, y su valor máximo el del día 21. También se observa que la forma de curva de irradiancia es aproximadamente igual a la de la potencia, lo que confirma los resultados obtenidos en el análisis de correlación.

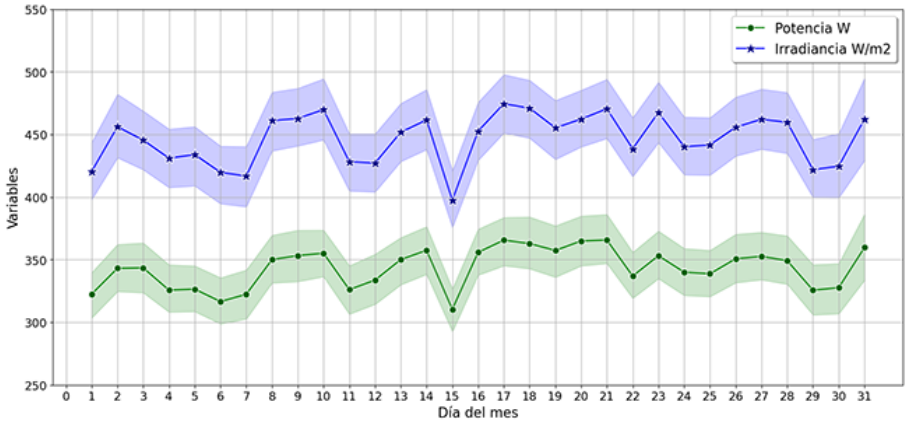

FIGURA 3
Curva diaria de irradiancia y potencia generada AC

El comportamiento de los valores promedios de estas mismas variables, pero ahora en una escala temporal mensual se presenta en la Figura 4. Se puede notar que los valores promedio máximos de irradiancia y potencia ocurrieron entre los meses de octubre y marzo, luego disminuyeron para obtener los valores mínimos entre los meses de mayo a julio. Los restantes meses fueron de transición.

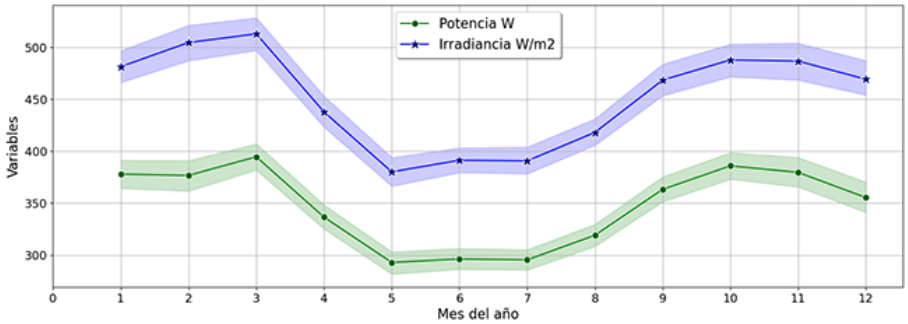

FIGURA 4
Curva mensual de irradiancia y potencia generada AC

Ahora bien, utilizando los datos de la columna de energía eléctrica, se calculan los datos para crear la columna del rendimiento específico (RendEsp), mientras que, con los datos de la columna de radiación, se calculan los datos para obtener la de rendimiento de referencia (RendRef). Posteriormente, con los datos de las columnas de rendimiento específico y rendimiento de referencia, se obtienen los datos para crear la columna de tasa de desempeño PR, el cual es uno de los indicadores utilizados para evaluar las plantas solares fotovoltaicas. Todos estos indicadores fueron obtenidos aplicando las ecuaciones presentadas en IEC (2016).

En la Figura 5 se presentan las curvas mensuales de los indicadores clave de desempeño de la planta solar. Se puede ver que la tasa de ensuciamiento (*soiling rate*) es alta y siempre mayor al 90% lo cual indica que la planta prácticamente no tiene pérdidas, para cada uno de los meses del año, pues según lo que se indica en Cordero et al. (2018), la tasa de ensuciamiento es igual a la relación entre la potencia de salida de los módulos cuando están sucios y la potencia de salida de los módulos cuando están limpios. Sin embargo, la tasa de desempeño es siempre mayor al 60% pero menor al 75%, lo cual evidencia un nivel de pérdidas en la planta que es relativamente alto, pero cuyo origen es diferente al nivel de ensuciamiento. Distintas razones, diferentes al soiling pueden causar disminución de PR, entre las esperadas se encuentran: la degradación del módulo solar, derrateo de potencia en inversores, entre otras. Según el NREL, en un estudio desarrollado en el año 2013, los paneles solares de silicio sufren una tasa promedio de degradación del 0,8% anual, con una mediana del 0,5% anual (Jordan & Kurtz, 2012).

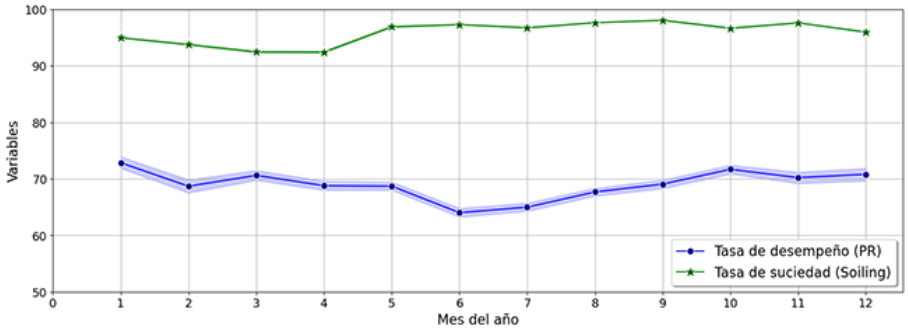

FIGURA 5
Curva mensual de PR y Soiling de la planta solar

El comportamiento de estos indicadores también se grafica en una escala anual, tal como se presenta en la Figura 6. Se puede ver que los valores promedio de la tasa anual de ensuciamiento siguen siendo altos, por lo que, las pérdidas razón serían notablemente bajas. En cuanto a la tasa de desempeño, esta inició con alrededor de 70% en el año 2011, se mantuvo en el año 2012, pero luego cayó, y quedó por debajo del 70% por el resto del período de estudio. Esta caída de la tasa de desempeño, con valores altos de tasa de ensuciamiento (bajo

nivel de suciedad), podría explicarse con la posible degradación de los distintos elementos que conforman la planta solar, entre otras razones posibles.

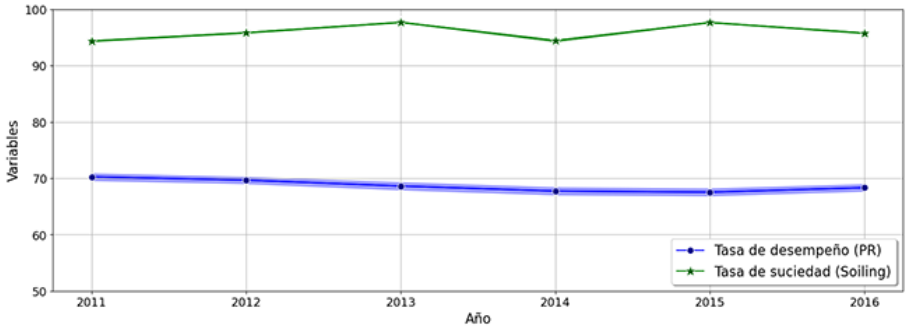

FIGURA 6
Curva anual de PR y Soiling de la planta solar

Modelación de los datos

En esta sección se discuten los resultados obtenidos luego de usar los algoritmos de aprendizaje automático a los datos, y obtener los respectivos modelos. Se agruparon los datos en clústers al aplicar el algoritmo de agrupamiento K-Means, adicionalmente, se generaron modelos de predicción de clases al emplear el algoritmo de K vecinos más cercanos K-NN.

Aplicación del algoritmo K-Means

El algoritmo para agrupamiento o clustering de datos K-Means es un algoritmo de aprendizaje automático de tipo no supervisado, a través del cual, se definen grupos o clústers de manera tal que cada elemento en un grupo específico presente una desviación mínima con respecto a los restantes elementos del grupo. Según lo planteado en Igual et al. (2017), el agrupamiento por K-Means consiste en juntar elementos que sean parecidos entre sí. Cuando exista más de un grupo para un conjunto de datos, los elementos de un mismo grupo o clúster deben ser parecidos entre sí, y los elementos de grupos diferentes deben tener características diversas entre sí.

Para este algoritmo, el hiperparámetro corresponde al número de clústers, el cual lo debe definir el analista. Se han diseñado distintas técnicas para encontrar su valor óptimo. Umargono et al. (2019) plantean que el "método del codo" es adecuado para definir el número de clústers K, este es un método gráfico que consiste en determinar en una gráfica de inercia vs. número de clústers , cuál es el número de clústers para el cual se observa la reducción más drástica de la inercia. Por su parte, Russano et al. (2020) consideran que la inercia es una métrica que se utiliza regularmente para obtener el valor óptimo de K, y explican que esta métrica no es más que el cuadrado de la distancia euclidiana entre cada punto del clúster y su centroide.

En la práctica, se ha demostrado que esta técnica funciona bien, pero en otros casos no es así, por lo que, se utilizan otros métodos. Un método muy popular consiste en seleccionar el número de clústers que maximiza la métrica silhouette. En Yuan y Yang (2019) indican que la métrica silhouette combina los factores de cohesión y separación. El factor de cohesión representa la similitud del elemento y los otros elementos de su clúster, mientras que el factor de separación revela que tan diferente es el elemento cuando se compara con los de otros clústers.

En esta investigación, para la aplicación del algoritmo K-Means se utilizan los datos horarios (26.172 registros) y se aplican las dos técnicas para la determinación del número óptimo de clústers. Previamente,

se crean columnas adicionales en el conjunto de datos correspondientes al rango de valores de la potencia generada (cuatro clases), y al rango de valores de la tasa de desempeño (tres clases).

Al aplicar las dos técnicas para obtener el valor óptimo de K, y considerando el rango de valores de la potencia AC, el resultado obtenido fue igual a cuatro para ambas técnicas. En la Figura 7 se presenta la gráfica de número de clústers K vs. Inercia, en la cual puede observar claramente que la inercia sufre una variación drástica justo cuando el número de clústers es igual a cuatro.

Entonces, aplicando el algoritmo de K-Means, se obtienen entonces cuatro grupos o clústers. El primer clúster tiene 12.169 registros, el segundo 5.459 registros, el tercero 5.341 registros, y el cuarto 3.203 registros. Los elementos de estos clústers se diferencian entre sí sólo por el rango de potencia AC al que pertenecen, y también por la hora del día al que pertenece el registro respectivo. Por ejemplo, el primer clúster tiene registros asociados a 20 de las 24 horas del día, el segundo tiene registros con horas que van desde las 8 am hasta las 3 pm, el tercer clúster presenta registros desde 7 am hasta las 4 pm y, por último, el cuarto tiene sólo registros con horas entre las 9 am y las 2 pm.

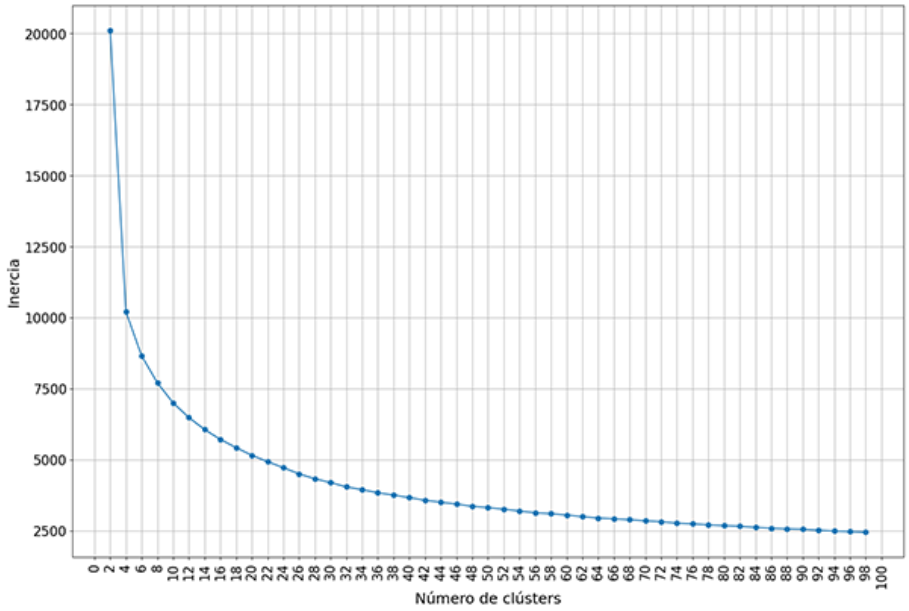

FIGURA 7
Determinación del K óptimo – Rango de potencia AC

En cuanto al rango de potencia AC, en la Figura 8 se presenta como fue su distribución con respecto a los clústers. Se puede ver que en el primer clúster solo hay registros cuya potencia AC es menor o igual a los 250 W, en el segundo se tienen registros cuya potencia AC está entre 500 W y 750 W, en el tercero clúster se evidencian registros cuya potencia AC está entre 250 W y 500 W, mientras que en el cuarto se muestran registros cuya potencia AC es mayor a 750 W.

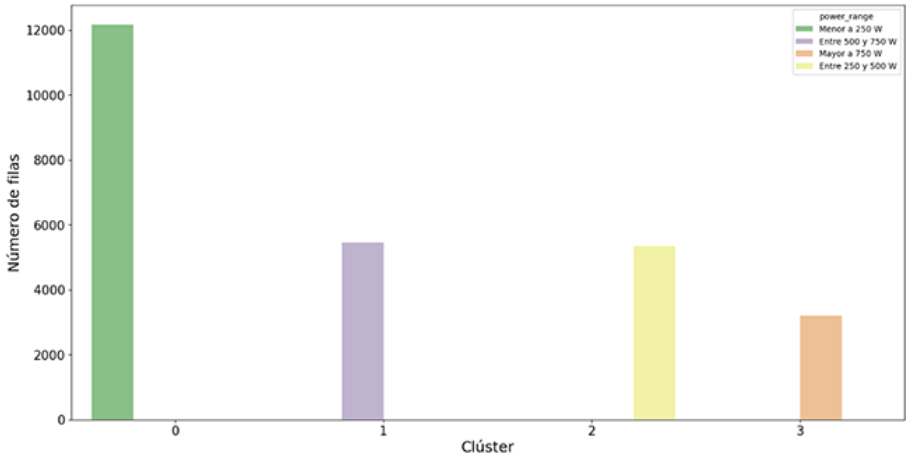

FIGURA 8
Clústers vs. Rango de potencia AC

Es así que se deduce que la planta solar genera potencia AC mayor a 750 W, sólo entre las 9 am y las 2 pm del día, mientras que entre las 8 am y las 3 pm la potencia generada varía entre 500 y 750 W. Asimismo, se obtuvo que para el segundo clúster se obtuvieron valores de la tasa de desempeño mayores o iguales al 70%, al igual que para el cuarto clúster.

Posteriormente, se repite el procedimiento, pero considerando para la aplicación del algoritmo el rango de valores de la tasa de desempeño (tres clases). Al aplicar las dos técnicas para obtener el valor óptimo de K, el resultado fue igual a cuatro para la técnica que usa la métrica inercia, y fue igual a tres para la técnica que usa la métrica silhouette. En la Figura 9 se presenta el resultado con la métrica silhouette, aquí el valor máximo ocurre cuando el número de clústers es igual a tres.

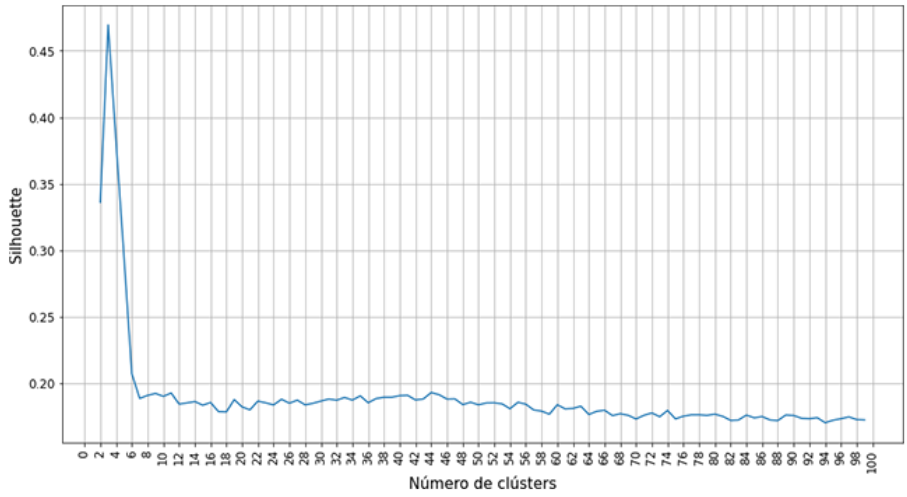

FIGURA 9
Determinación del K óptimo – Rango de tasa de desempeño

Por lo tanto, se aplica nuevamente el algoritmo K-Means, considerando tres grupos o clústers y al rango de valores de la tasa de desempeño. En esta oportunidad, se tienen 8.204 registros en el primer clúster, 9.801 registros en el segundo, y 8.167 en el tercero o. Los elementos de distintos clústers se diferencian entre sí sólo por la hora del día del registro, y por la clase del rango de la tasa de desempeño a la que pertenece el registro correspondiente. El primer clúster tiene registros con horas que van desde las 4 am hasta las 7 pm, además con una tasa de desempeño menor o igual a 70%. De igual forma, el segundo clúster evidencia registros con

horas especialmente desde las 7 am hasta las 3 pm y valores de tasa de desempeño mayores a 80%. Finalmente, en el tercer clúster se presentan registros con horas entre las 4 am y las 7 pm, y tasas de desempeño entre 70% y 80%. Adicionalmente, en el primer clúster hay valores, principalmente, de potencia AC menores a 250 W, mientras que para los otros dos clústers hay amplia variedad en los valores de potencia AC.

Aplicación del algoritmo K-NN

El algoritmo K-NN, o de los K vecinos más cercanos, es del tipo supervisado para clasificación. Uno de sus principales usos es predecir la clase o categoría de un set de datos a partir de un grupo de variables predictoras. Algunos autores como Lee (2019), consideran que este es uno de los algoritmos más simples al compararse con los otros algoritmos de aprendizaje supervisado para clasificación. Su principio de funcionamiento consiste en contrastar la distancia entre cada registro de referencia y las otras instancias del set de entrenamiento, seleccionando los K vecinos más cercanos a ellos.

Para desarrollar el modelo, en primer lugar, se debe definir el número de vecinos K, el cual es el hiperparámetro para este algoritmo. El procedimiento para obtenerlo consiste en seleccionar el valor de K que maximiza alguna métrica de desempeño. La métrica más utilizada para este fin es la exactitud (accuracy), puesto que, para un modelo de clasificación, Fenner (2020) indica que nos da "el porcentaje de respuestas que coinciden con la variable objetivo".

En este trabajo se aplicó el algoritmo en dos partes. Se consideran los datos mensuales desde marzo del 2010 (sin considerar agosto y septiembre de ese año) hasta noviembre del 2016 (para un total de 79 registros) y así generar un modelo que permitió predecir si la energía eléctrica producida estará por debajo o por encima del valor promedio de energía de los datos del período de estudio. De igual manera, se utilizaron estos datos para generar un modelo que posibilitó predecir si los valores de la tasa de desempeño de la planta estaban por debajo o por encima de su valor promedio.

Para la primera parte se tuvo como variables predictoras a la temperatura ambiente, la irradiancia, la tasa de ensuciamiento, y la velocidad del viento. Como variable objetivo se consideró a la clase de la energía eléctrica generada, por encima de 125 kWh (valor medio) o por debajo de 125 kWh. El valor óptimo del número de vecinos más cercanos K resultó para este caso igual a 15, el cual se utilizó para aplicar el algoritmo. Seguidamente, los 79 registros del conjunto de datos mensuales se dividieron en dos partes: el 70% (55 registros) para el set de entrenamiento que permitió la generación del modelo, y el 30% (24 registros) para el set de prueba que facilitó la evaluación del modelo obtenido.

Este modelo tuvo una exactitud del 91,67%, la cual es bastante alta, pero otra herramienta que usualmente se utiliza para mostrar el resultado de la evaluación del modelo de clasificación es la matriz de confusión, que es una matriz cuadrada cuya longitud de sus lados es igual al número de clases o categorías de la variable objetivo. Las celdas de la matriz tienen la siguiente información: los verdaderos negativos y los verdaderos positivos en la diagonal principal, y los falsos negativos y los falsos positivos en las otras celdas. Para este caso, ya que se tuvieron sólo dos categorías, se obtuvo una matriz de 2x2. La matriz de confusión obtenida, para los 24 registros del set de prueba, se presenta en la Figura 10.

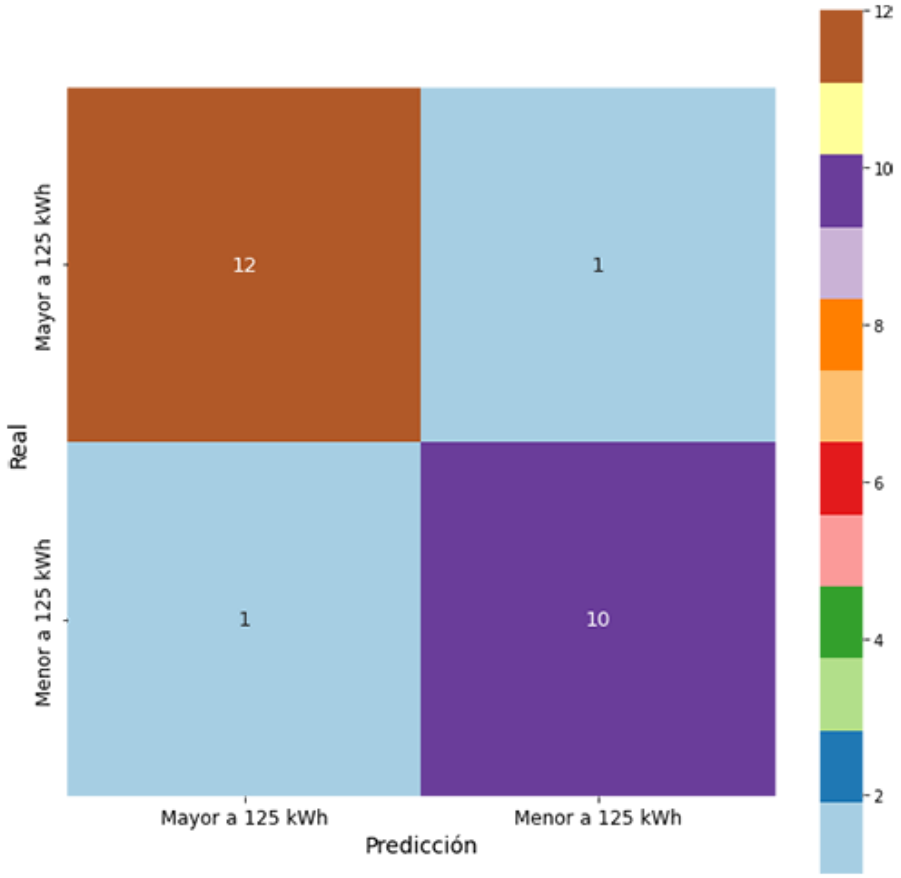

FIGURA 10
Matriz de confusión – Energía eléctrica

De la matriz de confusión se puede observar que del set de prueba se tuvo trece registros reales con energía eléctrica mensual mayor o igual a 125 kWh, y el modelo clasificó correctamente doce de ellos. Por otra parte, dicho set de prueba generó once registros reales con energía eléctrica mensual menor a 125 kWh, y el modelo clasificó correctamente diez de esos registros.

Posteriormente, para el segundo modelo se obtuvieron como variables predictoras a la potencia AC generada, la temperatura ambiente, la irradiancia, la tasa de ensuciamiento, y la velocidad del viento. Como variable objetivo se tuvo a la clase de la tasa de desempeño de la planta por encima del 69% (valor medio) o por debajo del 69%. El valor óptimo del número de vecinos más cercanos K resultó para este caso igual a 5, y este valor es el que se utiliza para aplicar el algoritmo. Al igual que en el caso anterior, los 79 registros del conjunto de datos mensuales se dividieron en dos partes: el 70% (55 registros) para el set de entrenamiento que permitió la generación del modelo, y el 30% (24 registros) para el set de prueba que facilitó la evaluación del modelo obtenido. En este caso, el modelo obtenido tuvo una exactitud del 83,33% en el conjunto de prueba.

La matriz de confusión de este modelo, para los 24 registros del set de prueba, se presenta en la Figura 11. Se puede notar que el set de prueba consta de doce registros reales con tasa de desempeño mayor a 69%, y el modelo clasificó diez de estos registros de manera correcta. De igual forma, el set de prueba constó también de doce registros reales, y el modelo organizóde manera correcta diez de los registros.

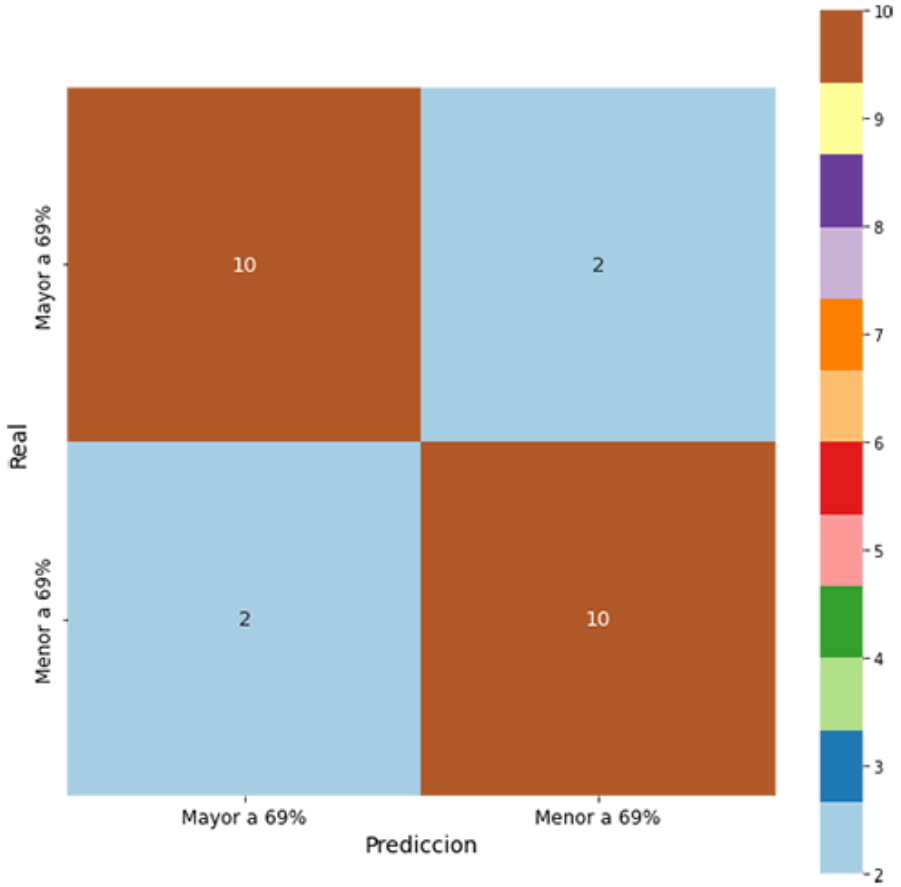

FIGURA 11
Matriz de confusión – Tasa de desempeño de la planta

Conclusiones

Del análisis de correlación, con los datos minutales, se pudo observar una relación lineal alta entre el nivel de irradiancia capturada por la instalación y la potencia AC generada. Este resultado fue confirmado con las gráficas temporales de irradiancia y potencia, en las escalas horaria, diaria y mensual, en las que se observó un comportamiento similar de ambas variables.

La tasa de ensuciamiento (*soiling rate*) promedio mensual es superior al 90% para todos los meses del año, lo cual demostró que las pérdidas por ensuciamiento en la planta son notablemente bajas y, por lo tanto, no tienen efecto alguno en la producción de energía eléctrica. Este resultado se repitió cuando se hizo el análisis en la escala anual. Por otra parte, la tasa de desempeño promedio mensual se movió alrededor del 70%, con valores mínimos del 65% para los meses de junio y julio. Mientras que, en la escala anual, la tasa de desempeño también se osciló desde un máximo de 70% para el primer año de estudio (2011), y luego una tendencia a la baja a partir del año 2012.

Cuando se consideró a la potencia AC generada dentro de los datos de entrenamiento, el algoritmo K-Means generó grupos donde los mayores niveles de potencia se suscitaron entre las 9 am y las 2 pm, junto con una tasa de desempeño superior al 70%. De igual forma, se obtuvo un clúster con registros entre las 8 am y las 3 pm, con niveles de potencia relativamente alto (entre 500 y 750 W), con tasa de desempeño superior al 70%. Los registros con menor potencia generada también tuvieron baja tasa de desempeño (menor al 70%).

Al considerar a la tasa de desempeño de la planta dentro del conjunto de datos de entrenamiento del modelo, con el algoritmo K-Means, se obtuvo un clúster con registros con horas desde las 7 am hasta las 3 pm,

valores de tasa de desempeño mayores a 80%, y valores de potencia generada mayoritariamente superiores a 500 W. En contraste, se tuvo un clúster con horas que van desde las 4 am hasta las 7 pm, con tasa de desempeño menores o igual a 70%, y potencia generada menor a los 250 W.

El modelo de predicción de la categoría de energía eléctrica generada tuvo una exactitud de casi el 92%, clasificando correctamente a 22 de los 24 registros del conjunto de prueba. Por otra parte, el modelo de predicción de la categoría de la tasa de desempeño de la planta tuvo una exactitud del 83%, categorizando correctamente a 20 de los 24 registros del conjunto de prueba.

Sería recomendable desarrollar un modelo de regresión lineal múltiple con los datos diarios de la energía eléctrica AC, y determinar cuál es el impacto de las variables climáticas de temperatura y velocidad del viento en la generación de energía eléctrica. Adicionalmente, se sugiere realizar un estudio sobre las causas de tener un PR relativamente bajo (alrededor del 70%), siendo que la tasa de ensuciamiento es relativamente alta (alrededor del 90%).

Referencias

Ahire, N., Agrawal, A., & Sharma, D. (2018). Performance Analysis of PV Solar Power System. *IOSR Journal of Electrical and Electronics Engineering*, 35-41. DOI: 10.9790/1676-1302013541.

Amat Rodrigo, J. (15 de 02 de 2023). *Ciencia de Datos, Estadística, Machine Learning y Programación*. (Joaquin Amat Rodrigo) Recuperado el 01 de Dicembre de 2022, de https://www.cienciadedatos.net/documentos/pystats05-correlacion-lineal-python.html

Asea Brown Boveri. (2019). *Technical Application Paper. Photovoltaic plants-Cutting edge technology. From sun to socket.*https://search.abb.com/library/Download.aspx?DocumentID=9AKK107492A3277&LanguageCode=en&DocumentPartId&Action=Launch.

Asociación Mexicana de Energía Solar. (2021). *Operación y Mantenimiento. Guía de Mejores Prácticas / Edición México.*https://asolmex.org/2021/04/29/operacion-y-mantenimiento/.

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science.* Shelter Island, NY: Manning Publications Co.

Cordero, R., Damiani, A., Laroze, D., MacDonell, S., Jorquera, J., Sepúlveda, E., . . . Torres, G. (2018). Effects of soiling on photovoltaic (PV) modules in the Atacama Desert. *Scientific Reports*, 1-14. DOI:10.1038/s41598-018-32291-8.

Fenner, M. E. (2020). *Machine Learning with Python for Everyone.* Boston: Pearson Education, Inc.

Igual, L., & Seguí, S. (2017). *Introduction to Data Science - A Python Approach to Concepts, Techniques and Applications.* Switzerland: Springer International Publishing.

International Electrotechnical Commission. (2016). *IEC TS 61 724-3 Photovoltaic system performance – Part 3: Energy evaluation method.* IEC.

Jordan, D., & Kurtz, S. (2012). Photovoltaic Degradation Rates — An Analytical Review. *National Renewable Energy Laboratory*.

Lee, W. M. (2019). *Python Machine Learning.* Indianapolis: John Wiley & Sons, Inc.

León-Ospina, C., Arias-Zarate, H., & Hernandez, C. (2023). Performance Evaluation of Photovoltaic Projects in Latin America. *International Journal of Advanced Computer Science and Applications*, 201-212. https://dx.doi.org/10.14569/IJACSA.2023.0140123.

McKinney, W. (2018). *Python for Data Analysis.* Sebastopol, CA: O'Reilly Media, Inc.

Nugroho, W., & Sudiarto, B. (2021). Performance evaluation of 5 MW Solar PV Power Plant in Kupang. *Materials Science and Engineering*. doi:10.1088/1757-899X/1098/4/042069.

PVDAQ NREL. (15 de 02 de 2023). *Duramat.* Obtenido de Duramat: https://datahub.duramat.org/dataset/pvdaq-time-series-with-soiling-signal

Ratner, B. (2017). *Statistical and Machine-Learning Data Mining - Techniques for Better Predictive Modeling and Analysis of Big Data.* Boca Raton, FL: CRC Press Taylor & Francis Group.

Romero-Fiances, I., Muñoz-Cerón, E., Espinoza-Paredes, R., Nofuentes, G., & de la Casa, J. (2019). Analysis of the Performance of Various PV Module Technologies in Peru. *Energies*. doi:10.3390/en12010186.

Russano, E., & Ferreira Avelino, E. (2020). *Fundamentals of Machine Learning Using Python.* Oakville, Canadá: Arcler Press.

SolarDesignTool. (15 de 02 de 2023). Obtenido de SolarDesignTool site: http://www.solardesigntool.com/components/module-panel-solar/Sanyo/2735/HIP200BA3/specification-data-sheet.html

Tackie, S., & Özerdem, Ö. (2022). Performance Evaluation and Viability Studies of Photovoltaic Power Plants in North Cyprus. *International Journal of Renewable Research*, 2237-2247. https://doi.org/10.20508/ijrer.v12i4.13670.g8583.

Umargono, E., Suseno, J. E., & Gunanwan S.K, V. (2019). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. *Advances in Social Science, Education and Humanities Research, 474*. https://doi.org/10.2991/assehr.k.201010.019.

Vasisht, M., Srinivasan, J., & Ramasesha, S. (2016). Performance of solar photovoltaic installations: Effect of seasonal variations. *Solar Energy*, 39-46. http://dx.doi.org/10.1016/j.solener.2016.02.013.

Veerendra Kumar, D., Deville, L., Ritter III, K., Raush, J. R., Ferdowsi, F., Gottumukkala, R., & Chambers, T. (2022). Performance Evaluation of 1.1 MW Grid-Connected Solar Photovoltaic Power Plant in Louisiana. *Energies*. https://doi.org/10.3390/en15093420

Verma, S., Yadav, D., & Sengar, N. (2021). Performance Evaluation of Solar Photovoltaic Power Plants of Semi-Arid Region and Suggestions for Efficiency Improvement. *International Journal of Renewable Energy Research*, 762-775. https://dorl.net/dor/20.1001.1.13090127.2021.11.2.25.4.

Yahyaoui, I. (2018). *Advances in Renewable Energies and Power Technologies - Volume 1: Solar and Wind Energies.* Cambridge: Elsevier Inc.

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *Multidisciplinary Scientific Journal*, 226-235. doi:10.3390/j2020016.