ARTÍCULO DE INVESTIGACIÓN / RESEARCH ARTICLE
https://dx.doi.org/10.14482/inde.43.01.489.326

Evaluación de la calidad de la energía eléctrica suministrada por una planta solar fotovoltaica conectada a la red haciendo uso de algoritmos de aprendizaje automático y de minería de datos

Evaluation of the quality of the electrical energy supplied by a grid-connected solar photovoltaic plant using machine learning and data mining algorithms

César Aristóteles Yajure Ramírez*

* Universidad Central de Venezuela, Facultad de Ingeniería, Profesor del postgrado de Investigación de Operaciones, Magíster.
Orcid-ID: https://orcid.org/0000-0002-3813-7606. cyajure@gmail.com.

Correspondencia: César Yajure. Ciudad Universitaria, Los Chaguaramos, Caracas.
Teléfono: +582126053030. cyajure@gmail.com.

Resumen

La presencia de plantas solares fotovoltaicas para la producción de electricidad implica la disminución del uso de combustibles fósiles y la reducción de las emisiones contaminantes. La disponibilidad de la energía solar depende de las condiciones climáticas, por lo que los parámetros de la energía eléctrica a entregar podrían verse afectados. El objetivo de esta investigación es mostrar una metodología basada en la ciencia de datos para la evaluación de la calidad de la energía de plantas solares fotovoltaicas conectadas a la red, considerando los estándares vigentes. Se aplica a una planta de 260 kWp del Instituto Nacional de Estándares y Tecnología de los Estados Unidos. Los parámetros utilizados son la distorsión armónica total, fluctuaciones y desbalance de voltaje, frecuencia eléctrica, y factor de potencia. Casi el 100 % de los registros cumplen con los límites establecidos por los estándares para los parámetros, a excepción del factor de potencia, con un 63,56 %. Del modelo de clasificación del factor de potencia se obtuvo que las potencias aparente y activa y la frecuencia son las variables más importantes. Del algoritmo de descubrimiento de subgrupos se obtuvo que la irradiancia solar aparece en el 40 % de los subgrupos, y la frecuencia en el 50 %.

***Palabras* clave:** distorsión armónica, factor de potencia, frecuencia eléctrica, nivel de importancia, subgrupos.

Abstract

The presence of photovoltaic solar plants to produce electricity implies the reduction of the use of fossil fuels, and of polluting emissions. The availability of solar energy depends on weather conditions, so the parameters of the electrical energy to be delivered could be affected. The objective of this research is to present a methodology based on data science for the evaluation of the energy quality of photovoltaic solar plants connected to the grid, considering current standards. It is applied to a 260 kWp plant of the National Institute of Standards and Technology of the United States. The parameters used are total harmonic distortion, voltage fluctuations and unbalance, electrical frequency, and power factor. Almost 100% of the records comply with the limits established by the standards for the parameters, except for power factor, with 63.56%. From the power factor classification model, the knowledge that apparent and active power, and frequency, are the most important variables was gathered. From the subgroup discovery algorithm, that solar irradiance appears in 40% of the subgroups, and frequency in 50%.

Keywords: electrical frequency, harmonic distortion, importance level, power factor, subgrupos. subgroups.

Fecha de recepción: 8 de febrero de 2024
Fecha de aceptación: 19 de julio de 2024

INTRODUCCIÓN

El uso de la energía solar para la generación de energía eléctrica a través de plantas solares fotovoltaicas ha ido en constante aumento en los últimos años, por dos razones principales: fuente primaria gratis y bajas o nulas emisiones contaminantes hacia el medio ambiente. Es así como en su cuaderno técnico, la empresa ABB [1] estipula:

> Entre los diferentes sistemas que utilizan fuentes de energía renovables, la energía fotovoltaica es prometedora debido a las cualidades intrínsecas del propio sistema: tiene costos de servicio muy reducidos (el combustible es gratuito) y requisitos de mantenimiento limitados, es confiable, no hace ruido y es muy fácil de instalar.

Esto ha permitido la disminución del uso de fuentes primarias más contaminantes como los combustibles fósiles, los cuales, sin embargo, tienen una gran ventaja: la estabilidad de su suministro. Por el contrario, la energía solar, que permite la generación de energía eléctrica de las plantas solares fotovoltaicas, depende de las condiciones climáticas del área en que se encuentre la planta, por lo que la estabilidad de su disponibilidad no se garantiza de manera absoluta, por consiguiente, los parámetros eléctricos del suministro de energía de la planta se podrían ver afectados.

Por lo anterior, los organismos reguladores nacionales e internacionales han establecidos normativas que fijan las características que debe cumplir el suministro de energía eléctrica de las plantas solares fotovoltaicas cuando se conectan con la red eléctrica. Por ejemplo, se tiene el estándar IEC 61727 de la Comisión Internacional de Electrotecnia (IEC por sus siglas en inglés), específico para plantas solares fotovoltaicas de hasta 10 kVA; el estándar 1547 del Instituto de Ingenieros Eléctricos y Electrónicos (IEEE por sus siglas en inglés), el cual proporciona requerimientos y especificaciones técnicas y de prueba de interconexión e interoperabilidad para los recursos energéticos distribuidos; el estándar IEEE 519 para práctica recomendada y requisitos para el control de armónicos en sistemas eléctricos de potencia, entre otros.

Ahora bien, entre los índices de calidad de la energía eléctrica a suministrar por las plantas se tienen: distorsión armónica total (THD por sus siglas en inglés), desbalance y fluctuaciones de voltaje, variaciones de la frecuencia y variaciones del factor de potencia. A este respecto, el objetivo de este trabajo es mostrar una metodología para la evaluación de la calidad de la energía de plantas solares fotovoltaicas a través de la ciencia de datos y considerando la normativa internacional. En la etapa de modelación se utiliza el algoritmo de bosques aleatorios de aprendizaje automático de clasificación, y el algoritmo de minería de datos para el descubrimiento de subgru-

pos. A través de esta metodología se podrá determinar el nivel de cumplimiento de los estándares relacionados con la calidad de la energía eléctrica entregada por una planta solar fotovoltaica. Se ilustra la metodología evaluando la calidad de la energía de una planta solar fotovoltaica de 260 kWp del NIST (National Institute of Standards and Technology) de los Estados Unidos. El conjunto de datos de esta planta ya fue utilizado en [2] para la selección de variables de modelos de pronóstico de la energía eléctrica generada por dicha planta.

Se hizo una revisión de los artículos ya publicados relacionados con el tema, y ninguno de los revisados utiliza formalmente la ciencia de datos para la investigación respectiva. Por ejemplo, Alhussainy y Alquthami [3] utilizan la simulación para evaluar la calidad de la energía entregada por plantas entre los 250 kW y los 3 MW, incluyendo los circuitos de control correspondientes en el dominio del tiempo. No observan impacto directo entre el tamaño de la planta y los armónicos de voltaje. La distorsión armónica de corriente oscila menos del 5 % de su límite, y el THD viola el nivel aceptable de armónicos. En [4], los autores realizan un estudio de calidad de la energía y perfiles de voltaje de una planta solar fotovoltaica de 1,1 MW situada en Miami, Florida. Hacen uso de los datos históricos con resolución de un minuto, para crear curvas y determinar patrones en estos datos. Observan que el THD de corriente desencadena problemas cuando la generación es intermitente, mientras que no se detecta relación entre el THD de voltaje y la salida de la planta. Sus resultados también muestran que pudieran presentarse desviaciones de voltaje y pérdidas de alimentadores con penetración de la planta de al menos 60 % en días de baja carga. En su trabajo, Ibrik [5] investiga el desempeño de una planta solar fotovoltaica de 72,8 kW, ubicada en Palestina, desde el punto de vista de la calidad de la energía que entrega. Analiza el flujo de energía total en el sistema, la variación de voltaje, la desviación de frecuencia, los armónicos de corriente y voltaje, y el THDi y THDv en el punto de acoplamiento común (PCC), y en la salida del sistema fotovoltaico. De las mediciones realizadas del THD de voltaje y corriente, *flickers*, frecuencia y factor de potencia concluye que sus valores están dentro de los límites de los estándares vigentes.

En la investigación [6], los autores presentan una metodología basada en la medición de parámetros de interés para evaluar la calidad de la energía de una planta solar fotovoltaica ubicada en Shanghai, China. Luego de realizar el estudio obtuvieron que la máxima desviación de frecuencia en el punto de acoplamiento común es de ±0,2 Hz, el máximo desbalance de voltaje no supera el 2 % y el THD de voltaje máximo obtenido fue menor al 2,5 %, es decir, todos los parámetros cumplen con la normativa vigente. En [7], los autores analizan la calidad de la energía eléctrica producida por cinco sistemas fotovoltaicos conectados a la red eléctrica (I, II, III, IV, y V), y ubicados en la ciudad de Cuenca, Ecuador. Previamente, realizan una revisión de las regulaciones ecuatoriana, norteamericana y europea, relacionada con la calidad de la

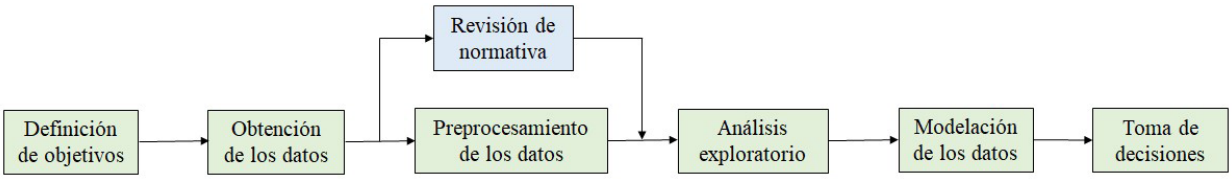
energía para determinar los parámetros regulatorios y sus límites admisibles. Para recabar los datos utilizados, instalaron medidores en las distintas plantas durante un período de una semana. El estudio determinó que la mayoría de los parámetros de los casos I, II, III, y V cumplen con los límites establecidos por las regulaciones analizadas, a excepción del desbalance de corriente. Asimismo, el caso IV no cumple con las regulaciones en cuanto a fluctuaciones y desbalance de voltaje, y parpadeos (*flickers*). En [8], los autores evalúan la calidad de la energía de una planta solar fotovoltaica de 5,5 kW conectada a la red eléctrica, ubicada en Egipto. Revisan el THD y estudian detalladamente cuatro técnicas para mitigar los armónicos. Modelan la planta y las cuatro técnicas haciendo uso del *software* Matlab/Simulink. La mejor solución correspondió a un filtro híbrido activo pasivo, el cual redujo el THD de la planta a 1,5 %, y la siguiente mejor solución fue un filtro shunt activo pasivo, que logró reducir el THD hasta el 4,5 %.

En [9], los autores realizan el monitoreo y análisis de la calidad de la energía de una planta solar fotovoltaica, compuesta por un sistema A de 14 kWp y un sistema B de 3,84 kWp, conectada a la red de bajo voltaje, ubicada en Rumania, y considerando el estándar EN 50160. Las mediciones de los parámetros se hicieron entre el 20 de julio de 2016 y el 22 de agosto del mismo año. Encontraron que la forma de onda entregada por el inversor no es perjudicada en condiciones de baja irradiancia solar, y con respecto a los armónicos, las tres fases cumplen con los límites establecidos en el estándar EN 50160. En cuanto a los *flickers* de corta y larga duración, los valores encontrados no caen dentro de los límites establecidos por el estándar. Finalmente, en [10], los autores realizan el análisis del desempeño de una planta solar fotovoltaica de 100 kW, conectada a la red eléctrica y ubicada en Tamil Nadu, India. El análisis lo desarrollan en cuatro fases: evaluación de factibilidad, evaluación del rendimiento energético, evaluación del ciclo de vida y evaluación de la calidad de la energía. Para realizar esta última evaluación, miden los parámetros durante catorce días seguidos, y obtuvieron que en promedio la frecuencia se mantiene con muy pocas variaciones alrededor de su valor nominal de 50 Hz, y el THD de voltaje siempre es menor al 3 % para cada una de las tres fases.

El resto del artículo se distribuye de la siguiente forma. En la sección 2 se presenta la metodología utilizada y los datos utilizados en esta investigación. Luego, en la sección 3, se analizan y discuten los resultados obtenidos. A continuación, en la sección 4, se presentan las conclusiones que se derivan del trabajo realizado. Finalmente, se presenta un listado con las referencias bibliográficas utilizadas.

METODOLOGÍA

La metodología consiste en aplicar las etapas de un proceso de ciencia de datos para evaluar la calidad de la energía que entrega una planta solar fotovoltaica. De acuerdo con lo planteado en [10], un proceso de este tipo consta por lo general de seis etapas; en la primera de ellas se fija el o los objetivos del proyecto, seguidamente se obtienen los datos a utilizar para el estudio, luego se preprocesan estos datos. Posteriormente, está el análisis exploratorio de los datos, la modelación de los datos y, por último, la toma de decisiones con base en los resultados obtenidos en las etapas previas. En la figura 1 se presenta un esquema gráfico de la metodología planteada.



Fuente: elaboración propia con base en [10].

Figura 1. Esquema de metodología planteada

Como se observa en la figura 1, al inicio se fija el objetivo o los objetivos del proyecto, lo cual en este caso sería la evaluación de la calidad de la energía que suministra una planta solar fotovoltaica. Seguidamente, se obtienen los datos a utilizar para hacer el estudio, los cuales podrán provenir tanto de fuentes internas como externas. Como tercera etapa está el preprocesamiento de los datos, en la que se aplican las técnicas propuestas por Mukhiya y Ahmed [11], vale decir: detección e imputación de datos faltantes, manejo de datos duplicados, detección y filtrado de datos atípicos, combinación de atributos, transformación de atributos, entre otros. Para esta investigación se agrega en paralelo al preprocesamiento de los datos, la revisión de la normativa vigente sobre la calidad de la energía que suministran las plantas solares fotovoltaicas, de manera tal de generar los índices requeridos para hacer la evaluación, tales como: desbalance y fluctuación de voltaje, THD, variación del factor de potencia, entre otros. A continuación, se desarrolla el análisis exploratorio de los datos, en el que se utilizan la estadística descriptiva, así como gráficos simples o combinados. En esta etapa ya se podría conseguir conocimiento útil (*insights*) que podría coadyuvar en la toma de decisiones o servir de base para la siguiente etapa. Posteriormente, se tiene la modelación de los datos, en la que se aplican algoritmos de aprendizaje automático supervisados para obtener modelos de regresión o clasificación, o no supervisados para detectar patrones de interés en el conjunto de datos. Finalmente, con todo el conocimiento adquirido en las etapas previas se procede a la toma de decisiones. Es importante destacar que estas etapas podrían aplicarse de manera secuencial, como

se sugiere en la figura 1, pero eventualmente podría haber realimentaciones hacia las etapas iniciales, dependiendo del caso particular que se esté tratando.

En esta sección se desarrollan las etapas de obtención de los datos, preprocesamiento de los datos y revisión de la normativa vigente. Mientras que en la sección siguiente se desarrollan las etapas de análisis exploratorio de los datos y modelación de los datos.

Indicadores de calidad de la energía eléctrica

La definición de calidad de la energía eléctrica o calidad del suministro es establecida en el estándar IEC 61000-4-30 como "Características de la electricidad en un punto dado de una red de energía eléctrica, evaluadas con relación a un conjunto de parámetros técnicos de referencia" [12]. Entonces, existe un número determinado de indicadores para medir la calidad de la energía eléctrica que entrega una planta de generación. En esta investigación se presentan aquellos que son adecuados para hacer la evaluación de una planta solar fotovoltaica, y que además se pueden calcular con las variables presentes en el conjunto de datos original. El primer indicador es la distorsión armónica total THD, la cual según [13] "es el valor efectivo de las componentes armónicas de una forma de onda distorsionada", y se calcula utilizando (1). Para baja tensión (menos de 1000 V), el estándar IEEE 519-2014 indica que el THD máximo debería ser del 8 % [14], y en [15] indican que el estándar IEEE 1547 sugiere un máximo de 5 %, el cual es el valor considerado en esta investigación.

$$THD(\%) = \frac{\sqrt{\sum_{h>1}^{h_{max}} M_h^2}}{M_1} \tag{1}$$

Donde:

M_h: Es el valor rms de la *h-ésima* componente de la cantidad M, que pudiera ser voltaje o corriente.

Por otra parte, se tiene el desequilibrio o desbalance de voltaje, definido como la máxima desviación del promedio de las tensiones trifásicas, dividido por el promedio de las tensiones trifásicas, expresadas en porcentaje, tal como se establece en [16] y se muestra en (2). En [13] se plantea que otra forma de calcular el desbalance de voltaje es utilizando las componentes simétricas, más específicamente, como la relación entre la componente de secuencia negativa (o cero) y la componente de secuencia positiva.

$$desb\_volt = \frac{\max(V_a, V_b, V_c)}{promedio(V_a, V_b, V_c)} \times 100 \tag{2}$$

El estándar IEEE 1547-2018 indica que se tolera un desbalance del 5 % siempre y cuando no exceda una duración de 60 segundos, o un valor del 3 % siempre y cuando su duración no exceda de 300 segundos [17]. En esta investigación se trabaja con el límite del 3 %.

Asimismo, se tienen las fluctuaciones de voltaje hacia abajo o hacia arriba del valor nominal. El estándar IEC 61727-2004 establece que el rango normal de operación para el voltaje, en el punto de conexión con la red, se establece entre el 85 % y el 110 % del valor nominal [18]. El mismo estándar establece que los valores de operación normal para la frecuencia deben estar dentro del rango de 59 a 61 Hz, mientras que el estándar IEEE 1547-2018 indica que el rango normal de operación va de 58,8 a 61,2 Hz.

En cuanto al factor de potencia (fp), el estándar IEC 61727-2004 indica que cuando la potencia activa alcanza el 50 % del valor nominal del inversor, el fp debe ser en atraso y con un valor superior a 0,9. En [15] presentan una tabla con los valores mínimos permitidos en distintos países, variando entre 0,85 y 0,95. Este último valor es el utilizado en esta investigación.

Obtención de los datos

El conjunto de datos proviene de la página web del NIST [19]. Incluye variables asociadas al clima, y también variables eléctricas, las que corresponden a las mediciones minutales realizadas del período 2015 - 2018, de las estaciones de medición de una planta solar fotovoltaica, ubicada en Maryland, Estados Unidos. Contiene 1.152 paneles solares de silicio monocristalino marca Sharp, con 235 Wp por panel [20], y consta de un inversor de 260 kW nominales de potencia AC marca PVPowered, voltaje nominal de 480 VAC, conexión en Y, 4 hilos [21].

Del subconjunto de variables eléctricas se tienen: potencia activa AC en kilovatios (“‘PwrMtrP_kW_Avg”), potencia reactiva en kilovoltio amperios reactivos (“PwrMtrP_kVAR_Avg”), potencia aparente en kilovoltio amperios aparentes (“PwrMtrP_kVA_Avg”), frecuencia eléctrica en Hertz (“PwrMtrFreq_Avg”), factor de potencia (“PwrMtrPF_Avg”), voltajes por fase en voltios (‘PwrMtrVa_Avg’, ‘PwrMtrVb_Avg’, ‘PwrMtrVc_Avg’), corrientes por fase en amperios (‘PwrMtrIa_Avg’, ‘PwrMtrIb_Avg’, ‘PwrMtrIc_Avg’), entre otras.

El subconjunto de variables climáticas incluye: irradiancia solar en vatios por metro cuadrado (“SEWSPOAIrrad_Wm2_Avg”), temperatura ambiente en grados Celsius (“SEWSAmbientTemp_C_Avg”), temperatura promedio en los paneles solares en grados Celsius (“SEWSModuleTemp_C_Avg”), velocidad promedio del viento en metros por segundo (“WindSpeedAve_ms”), entre otras.

El grupo de datos está constituido por 2.103.810 registros de las mediciones minutales de un total de noventa y nueve variables, separados en 1.461 archivos con datos diarios, propios de cada uno de los días entre 2015 y 2018.

Preprocesamiento de los datos

Inicialmente, se combinaron los 1.461 archivos de datos diarios, para crear un solo archivo, y así alcanzar la totalidad de registros mencionados con anterioridad. Posteriormente, se realizó un análisis de datos faltantes, implicando que de las noventa y nueve columnas (variables), la fecha fue la única que no tuvo datos faltantes, siendo 9.587 el valor mínimo y 128.369 el valor máximo de datos faltantes por columna. Para esta investigación no eran de utilidad las variables asociadas al inversor, puesto que el conjunto de datos original incluye columnas con los registros requeridos para el cálculo de los indicadores utilizados en este trabajo, y además eran las variables con mayor cantidad de datos faltantes, por lo que fueron eliminadas. Luego, se eliminaron los registros con al menos un dato faltante, para quedar un total de 1.997.418 registros sin datos faltantes. No se detectaron datos duplicados.

Seguidamente, se estableció que los valores significativos de la irradiancia solar están presentes entre las seis de la mañana y las seis de la tarde, por lo que se optó por eliminar los registros fuera de este intervalo de tiempo. Asimismo, se detectaron 528 registros ilógicos o sin sentido para la mayoría de las variables, por lo que fueron eliminados. Luego de estos dos ajustes quedaron en total 1.030.901 filas o registros en el conjunto de datos. A continuación, se hizo un chequeo de la columna de frecuencia eléctrica, observando valores con error de medición, procediendo a eliminar las filas correspondientes. Asimismo, se detectaron valores negativos para la irradiancia solar, cuyas filas también fueron eliminadas. Finalmente, quedaron 1.029.561 filas o registros para ser utilizados en los análisis posteriores.

Subsiguientemente, utilizando las columnas de los voltajes por fase, se obtuvo una columna que indica si todos los voltajes de fase están dentro de los límites establecidos por la normativa (“estatus_limV”), y otra columna que muestra el desbalance de voltaje (“desb_volt”). A partir de esta última se generó una columna adicional que indica si el desbalance de voltaje cumple o no con la normativa (“estatus_desbV”). De igual manera, a partir de la columna de la frecuencia, se creó una columna que indica si los valores de esta variable están dentro de los límites establecidos por la normativa (“estatus_freq”). También se creó una columna para verificar si el THD cumple con la normativa (“estatus_THD”), y otra para verificar si el factor de potencia está dentro de los límites establecidos por la normativa (“estatus_FP1”).

DISCUSIÓN DE RESULTADOS

A continuación, se analizan y discuten los resultados generados en las etapas de análisis exploratorio de los datos y su modelación, considerando los indicadores de calidad de la energía obtenidos en la etapa de preprocesamiento.

Análisis exploratorio de los datos

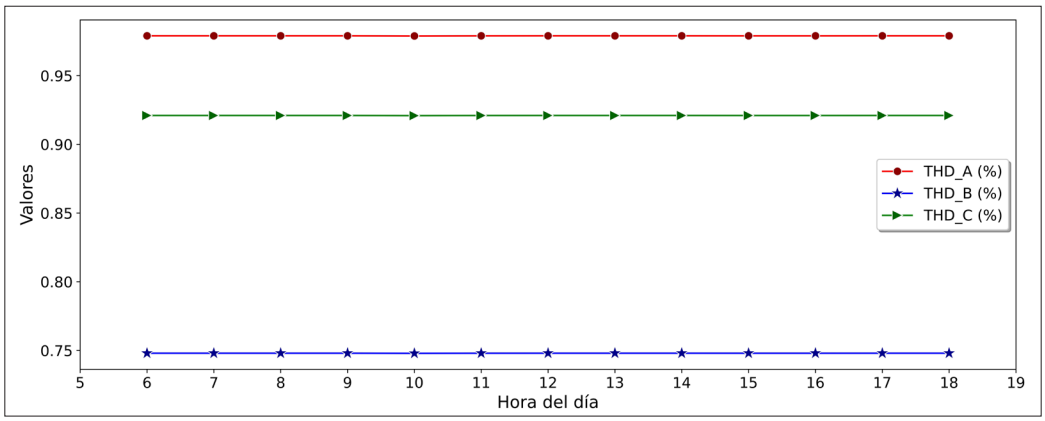
En primer lugar, se realizó un análisis descriptivo de las principales variables eléctricas, lo cual se presenta en la tabla 1. Se puede ver que, para las variables de potencia y factor de potencia, el valor medio está alejado de la mediana de los datos. Asimismo, los valores de THD presentan una variación nula con respecto a la media, y además siempre es menor al 1 %. El valor máximo del desbalance de voltaje es igual a 3,66 %.

Tabla 1. Estadística descriptiva de variables eléctricas

Estadístico	Freq	Va	Vb	Vc	P_kW	Q_kVAR	S_kVA	fp	VaTHD	VbTHD	VcTHD	desb_volt
Media	60,1	271,6	271,2	272,7	73,9	5,9	75,2	0,41	0,98	0,75	0,92	0,33
DesvStd	5,4	2,2	2,1	2,1	73,6	11,2	73,4	0,83	0,00	0,00	0,00	0,10
min	57,3	246,0	258,0	260,6	-0,8	-9,2	0,0	-1,00	0,00	0,04	0,00	0,00
Q1	60,0	270,0	270,0	271,0	7,9	-2,6	9,6	-0,31	0,98	0,75	0,92	0,25
Mediana	60,0	272,0	271,9	273,0	47,9	1,2	47,9	0,99	0,98	0,75	0,92	0,32
Q3	60,0	273,0	273,0	274,0	131,9	14,3	132,7	1,00	0,98	0,75	0,92	0,38
max	493,8	287,3	287,4	288,4	258,4	37,8	260,6	1,00	0,98	0,75	0,92	3,66

Fuente: elaboración propia.

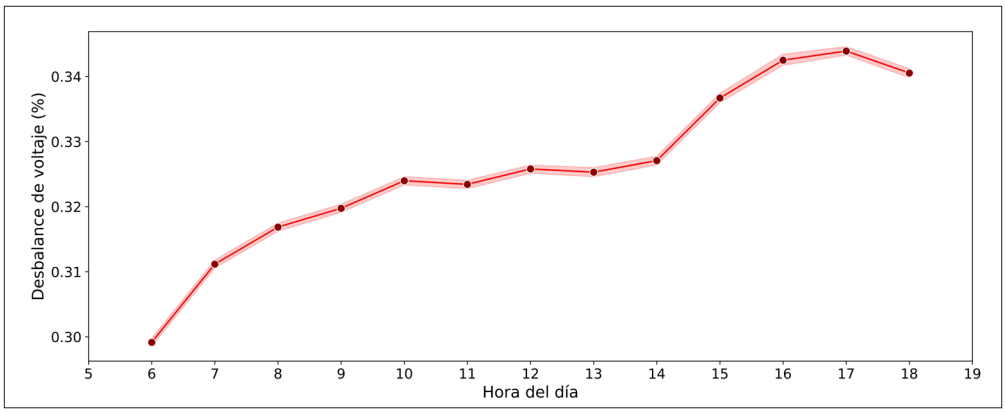
Con el fin de verificar el comportamiento del THD, se generó una gráfica con el valor promedio horario para cada una de las fases, la cual se presenta en la figura 2. Cada valor horario mostrado se corresponde con el intervalo de confianza al 95 % de los sesenta valores minutales de la hora respectiva. Se confirma que los valores del THD son menores al 1 %, y que, además, no hay variabilidad en estos datos.



Fuente: elaboración propia.

Figura 2. Curva horaria del THD

De igual forma, se obtuvo la curva horaria del desbalance de voltaje, la cual se muestra en la figura 3. En este caso se nota la diferencia entre la curva más oscura (media) y la franja más clara (bandas del intervalo de confianza), lo cual indica que hay algo de variabilidad en el desbalance.



Fuente: Elaboración propia.

Figura 3. Curva horaria del desbalance de voltaje

Como se dijo previamente, para el THD se fijó un límite máximo del 5 %, para el desbalance un valor máximo del 3 %, para la frecuencia los valores deberían estar entre 59 Hz y 61 Hz, y para el factor de potencia, el límite considerado fue de 0,95 (en atraso o en adelanto). Se contabilizó la cantidad de registros que cumplieron con estos límites, y los resultados se presentan en la tabla 2.

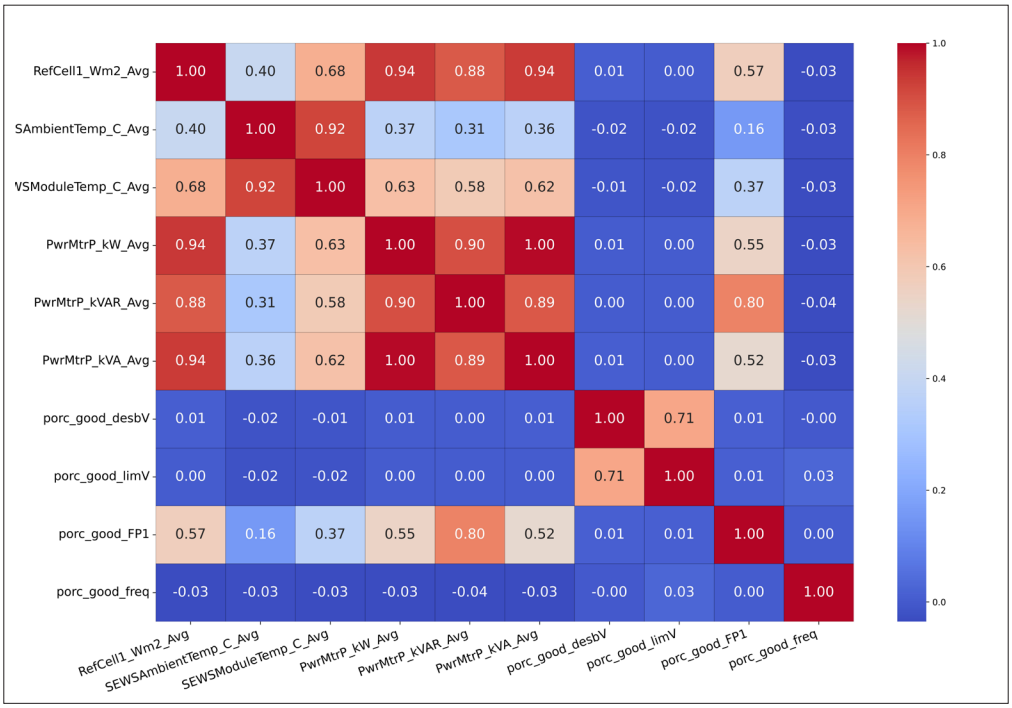
Tabla 2. Cumplimiento de normativa

1: SI O: NO	estatus_desbV	estatus_THD	estatus_limV	estatus_freq	estatus_FP1
1	1.029.508	1.029.561	1.029.503	1.028.859	654.343
0	53	0	58	702	375.218
% Cumplimiento	99,99%	100,00%	99,99%	99,93%	63,56%

Fuente: elaboración propia.

Se puede notar que el 100 % de los valores del THD cumple con el límite fijado de acuerdo con la normativa. Por otro lado, los valores del desbalance de voltaje, las fluctuaciones de voltaje y la frecuencia cumplen casi en su totalidad con los límites fijados por la normativa. La excepción la representa el factor de potencia, ya que solo el 63,56 % de los registros cumple con el límite de 0,95.

Seguidamente, se obtuvo la matriz de correlación considerando las variables climáticas, la potencia generada, así como los porcentajes de cumplimiento de los indicadores con respecto a la normativa. Los resultados se presentan en la figura 4.



Fuente: elaboración propia.

Figura 4. Matriz de correlación con porcentaje de cumplimiento

De la figura 4 se puede observar que el porcentaje de cumplimiento del factor de potencia tiene relación directa con la irradiancia solar, la temperatura ambiente, la temperatura de los paneles solares y con las potencias activa, reactiva y aparente. Para analizar las magnitudes se recuerda que un valor absoluto de correlación entre 0 y 0,3 implica una débil relación entre el par de variables, entre 0,3 y 0,7 esa relación es moderada, mientras que valores entre 0,7 y 1 indican una fuerte relación entre las variables [22].

Entonces, del análisis de correlación se puede decir que el porcentaje de cumplimiento del factor de potencia tiene una fuerte relación (aunque en el límite) con la potencia reactiva, relación moderada con la irradiancia solar, la temperatura de los paneles solares, la potencia activa y la potencia aparente, y una relación débil con la temperatura ambiente.

Modelación de los datos

Para la modelación de los datos se trabajó con dos algoritmos: el de bosques aleatorios, perteneciente al grupo de algoritmos de aprendizaje automático supervisado, y el algoritmo de descubrimiento de subgrupos, que se ubica dentro de la minería de datos.

Algoritmo de bosques aleatorios

Este algoritmo forma parte de los métodos de conjunto (*ensemble methods*) que se basan en combinar los resultados de un conjunto de estimadores simples [23]. Para el caso de bosques aleatorios, los estimadores simples son árboles de decisión, a partir de los cuales se crea un modelo más robusto que tiene un mejor desempeño de generalización y es menos susceptible al sobreajuste [24].

De la sección previa se obtuvo que el 63,56 % de los registros cumplieron con el límite fijado para el factor de potencia con el fin de dar cumplimiento a la normativa. Entonces, se utiliza este algoritmo para crear un modelo de clasificación que permita predecir si las nuevas mediciones cumplirán con el valor de 0,95 del factor de potencia (uno) o no lo cumplen (cero).

La variable objetivo es el estatus del factor de potencia, y las variables explicativas son: irradiancia solar, temperatura de los paneles solares, temperatura del medio ambiente, velocidad promedio del viento, los tres voltajes de fase, la frecuencia, la potencia activa y la potencia aparente. Los datos se dividen en dos partes: 75 % para en entrenamiento del modelo y el restante 25 % para la prueba del modelo. Se considera un total de 100 estimadores simples, y para la evaluación del modelo se considera la matriz de confusión además de la métrica exactitud.

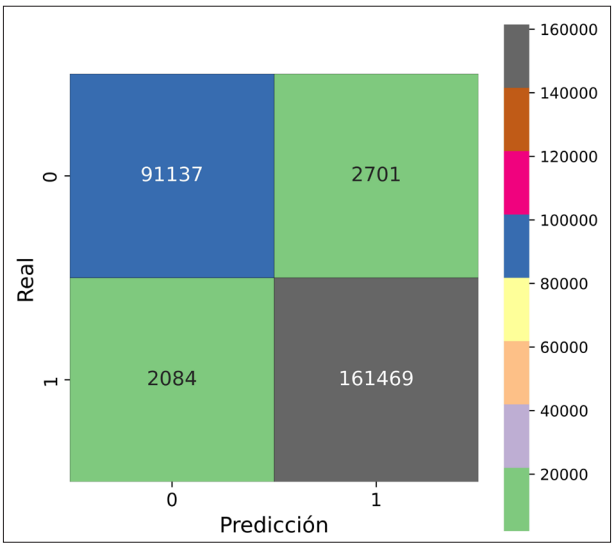
Este algoritmo entrega el nivel de importancia de cada variable explicativa. En la tabla 3 se presentan estos niveles de importancia ya jerarquizados. Se puede observar que, de acuerdo con este algoritmo, la potencia aparente, la potencia activa y la irradiancia solar están en los tres primeros lugares de importancia. La frecuencia eléctrica y la velocidad del viento resultaron con los menores niveles de importancia.

Tabla 3. Niveles de importancia

Variable	Importancia
PwrMtrP_kVA_Avg	0,378
RefCell1_Wm2_Avg	0,248
PwrMtrP_kW_Avg	0,237
SEWSModuleTemp_C_Avg	0,049
SEWSAmbientTemp_C_Avg	0,023
PwrMtrVc_Avg	0,018
PwrMtrVb_Avg	0,017
WindSpeedAve_ms	0,014
PwrMtrVa_Avg	0,014
PwrMtrFreq_Avg	0,002

Fuente: elaboración propia.

Por otra parte, en la figura 5 se presenta la matriz de confusión del modelo, de la cual se puede observar que el conjunto de prueba estuvo compuesto por 257.391 registros. De esos, 93.838 correspondían a factor de potencia menor a 0,95 y el modelo clasificó de forma correcta 91.137 de esos registros, es decir, 97,12 %. Los restantes 163.553 registros tenían factor de potencia mayor o igual a 0,95 y el modelo clasificó de manera correcta el 98,73 % de estos (161.469). Estos resultados dan una exactitud global del 98,14 %.



Fuente: elaboración propia.

Figura 5. Matriz de confusión del modelo de clasificación

Algoritmo de descubrimiento de subgrupos

Es una técnica de minería de datos que se utiliza para encontrar relaciones entre distintos atributos de un conjunto de datos, con respecto a una variable objetivo. En [25] plantean que un aspecto importante es la estrategia de búsqueda utilizada por el algoritmo. Adicionalmente, indican que otro aspecto importante consiste en evaluar la calidad del subgrupo respectivo, para lo cual existe un número determinado de métricas que usualmente llaman "medidas de calidad". En [26], los autores realizan un análisis de las distintas medidas utilizadas para evaluar la calidad de los subgrupos obtenidos. La variable objetivo podría ser del tipo nominal o del tipo numérico, y para ambos casos existen medidas de calidad específicas.

Entonces, se generaron rangos de valores para la mayoría de las variables utilizadas en el modelo de forma tal de manejarlas como variables categóricas. Estas variables son: la potencia aparente, la potencia activa, la temperatura de los paneles solares y la irradiancia solar. Como variable objetivo se considera al estatus del factor de potencia, específicamente cuando su valor es igual o mayor a 0,95. Es importante recordar que el número de registros es igual a 1.029.561, y de esos, 654.343 cumplen con el valor deseado para el factor de potencia.

El nivel de profundidad seleccionado para la búsqueda es tres, es decir, se encontrarán subgrupos de hasta tres atributos relacionados, mientras que la función de calidad para evaluar los subgrupos es WRAcc (*Weighted Relative Accuracy*), y la estra-

tegia de búsqueda de subgrupos se realiza utilizando el algoritmo de búsqueda de Haz (*Beam Search*). En [27], los autores utilizan la función WRAcc para casos en los cuales la variable objetivo es del tipo de clasificación, como en esta investigación, así como el mismo algoritmo de búsqueda. Luego de realizar la búsqueda de subgrupos, los resultados obtenidos se presentan en la tabla 4.

Tabla 4. Subgrupos encontrados

Sg	Calidad	Subgrupo	Tamaño	Positivos	Cobertura	Elevación
1	0,1182	Irrad==’Mayor a 400’ AND PwrMtrFreq_Avg: [60.0:60.01[	357.344	348.759	0,533	1,536
2	0,1181	Irrad==’Mayor a 400’	358.068	349.152	0,534	1,534
3	0,1153	PwrMtrFreq_Avg: [60.0:60.01[AND kVA==’Mayor a 100’	335.387	331.897	0,507	1,557
4	0,1153	kVA==’Mayor a 100’	336.068	332.278	0,508	1,556
5	0,1151	PwrMtrFreq_Avg: [60.0:60.01[AND kW==’Mayor a 100’	334.161	330.865	0,506	1,558
6	0,1151	PwrMtrFreq_Avg: [60.0:60.01[AND kVA==’Mayor a 100’ AND kW==’Mayor a 100’	334.161	330.865	0,506	1,558
7	0,1150	kVA==’Mayor a 100’ AND kW==’Mayor a 100’	334.839	331.243	0,506	1,557
8	0,1150	kW==’Mayor a 100’	334.839	331.243	0,506	1,557
9	0,1140	Irrad==’Mayor a 400’ AND PwrMtrFreq_Avg: [60.0:60.01[AND kVA==’Mayor a 100’	331.439	328.000	0,501	1,557
10	0,1139	Irrad==’Mayor a 400’ AND kVA==’Mayor a 100’	332.116	328.377	0,502	1,556

Fuente: elaboración propia.

De la tabla 4 se puede notar que los diez subgrupos encontrados están ya jerarquizados de acuerdo con la función de calidad mencionada previamente, pero se incluyen dos medidas de calidad adicionales, la cobertura (*coverage*), que es la relación de registros positivos en el subgrupo respectivo con respecto a los registros positivos del conjunto de datos total, y la elevación (*lift*), que es la proporción de registros positivos en el subgrupo con respecto a la proporción de registros positivos en el conjunto de datos total. En ambas medidas, valores altos implica mejores características del subgrupo respectivo.

El primer subgrupo corresponde a registros con irradiancia solar mayor a 400 W/m² y frecuencia eléctrica igual a 60 Hz, es decir, los registros que incluyen esos valores para esos dos atributos impactan positivamente el alcance del objetivo. El segundo subgrupo corresponde a registros solo con irradiancia solar mayor a 400 W/m², pero el tercer subgrupo corresponde a registros con una frecuencia eléctrica igual a 60 Hz y potencia aparente mayor a 100 kVA. Se puede ver que el quinto subgrupo incluye también a los registros con una frecuencia igual a 60 Hz, pero con los valores de potencia activa mayores a 100 kW. El resto de los subgrupos se interpreta de la misma forma.

Ahora, si se analizan los subgrupos desde el punto de vista de la elevación, se puede deducir que los subgrupos 5 y 6 serían los mejores, y ambos incluyen a la frecuencia igual a 60 Hz y a potencias activas mayores a 100 kW, acotando que el número seis también incluye a potencias aparentes mayores a 100 kVA.

CONCLUSIONES

Se presentó una metodología basada en la ciencia de datos para evaluar la calidad de la energía que entrega una planta solar fotovoltaica conectada a la red. Se utilizaron indicadores definidos en la normativa vigente sobre la calidad de la energía de este tipo de plantas. Esta metodología permite determinar el nivel de cumplimiento de los estándares relacionados con la calidad de la energía eléctrica producida por este tipo de plantas. El uso de la ciencia de datos es un enfoque original que permite el manejo de grandes cantidades de datos históricos para evaluar la calidad de la energía que ha entregado, y entrega, una planta solar fotovoltaica, identificando los factores que inciden en el cumplimiento de los estándares vigentes.

Los indicadores seleccionados fueron, el desbalance de voltaje, la distorsión armónica total, fluctuaciones de voltaje y factor de potencia. De los datos históricos se obtuvo que la mayoría cumple con los parámetros exigidos por la normativa en alrededor del 99 % de los registros, a excepción del factor de potencia, del que solo el 63,56 % de los registros cumple con el límite exigido por la normativa vigente.

Con el algoritmo de bosques aleatorios se obtuvo un modelo de clasificación para predecir si en los registros nuevos el valor del factor de potencia cumplirá o no con el límite del 0,95. Este modelo en su fase de prueba tuvo una exactitud del 98,14 %.

Considerando como meta un factor de potencia mayor o igual a 0,95, y utilizando un algoritmo de minería de datos, se generaron diez subgrupos dentro del conjunto de datos para determinar cuál o cuáles atributos influyen en la obtención de ese valor

límite del factor de potencia, siendo la irradiancia solar uno de los atributos más influyentes.

REFERENCIAS

[1] Asea Brown Boveri, «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, 2019.

[2] C. A. Yajure-Ramírez, «Enfoque multicriterio para la selección óptima de variables explicativas para modelos de pronóstico de la energía eléctrica de plantas solares fotovoltaicas,» *Revista Tecnológica* ESPOL, vol. 35, nº 3, pp. 84-98, 2023. https://doi.org/10.37815/rte.v35n3.1045.

[3] A. A. Alhussainy y T. S. Alquthami, «Power quality analysis of a large grid-tied solar photovoltaic system,» *Advances in Mechanical Engineering*, pp. 1-14, 2020. DOI: 10.1177/1687814020944670.

[4] A. Anzalchi, A. Sundararajan, A. Moghadasi y A. Sarwat, «Power Quality and Voltage Profile Analyses of High Penetration Grid-tied Photovoltaics: A Case Study,» IEEE *Industry Applications Magazine*, pp. 1-8, 2019.

[5] I. Ibrik, «Power Quality and Performance of Grid-Connected Solar PV System in Palestine,» *International Journal of Engineering Research and Technology*, pp. 1500-1507, 2019.

[6] J. Cui, H. Siming, Z. Bingbing, Z. Hua, D. Chang, Z. Guofa y M. Bo, «Monitoring and Analysis of Power Quality in Photovoltaic Power Generation System,» de E3S W *eb of Conferences* ICPRE, 2018.

[7] M. Barreto, A. Guananga, A. Barragán, E. Zalamea y X. Serrano, «Power Quality Analysis of Photovoltaic Systems,» de 21*th International Conference on Renewable Energies and Power Quality* (ICREPQ'23), 2023.

[8] W. Salem, W. G. Ibrahim, A. M. Abdelsadek y A. A. Nafeh, «Grid connected photovoltaic system impression on power quality of low voltage distribution system,» *Cogent Engineering*, pp. 1-18, 2022. https://doi.org/10.1080/23311916.2022.2044576.

[9] L.-A. El-Leathey, A. Nedelcu y M. Dorian, «Power Quality Monitoring and Analysis of a Grid-Connected PV Power Plant,» ELECTROTEHNICA, ELECTRONICA, AUTOMATICA (EEA), 2017.

[10] A. Lavanya, J. Divya Navamani, A. Geetha, V. Ganesh, M. Jagabar Sathik, K. Vijayakumar y D. Ravichandran, «Smart energy monitoring and power quality performance

based evaluation of 100-kW grid tied PV system,» Heliyon, pp. 1-19, 2023. https://doi.org/10.1016/j.heliyon.2023.e17274.

[11] S. K. Mukhiya and U. Ahmed, Hands-On Exploratory Data Analysis with Python, Birmingham, UK: Packt Publishing Ltd., 2020.

[12] International Electrotechnical Commission, «IEC 61000-4-30 - Técnicas de ensayo y de medida. Métodos de medida de la calidad del suministro,» IEC, Geneva, Switzerland, 2003.

[13] R. C. Dugan, M. F. McGranaghan, S. Santoso y H. W. Beaty, Electrical Power Systems Quality, The McGraw-Hill Companies, 2004.

[14] The Institute of Electrical and Electronics Engineers, «IEEE Std 519 - Recommended Practice and Requirements for Harmonic Control in Electric Power Systems,» IEEE, New York, 2014.

[15] A. Q. Al-Shetwi, M. A. Hannan, K. P. Jern, A. A. Alkahtani y A. E. PG Abas, «Power Quality Assessment of Grid-Connected PV System in Compliance with the Recent Integration Requirements,» *Electronics* MDPI, pp. 1-22, 2020. doi:10.3390/electronics9020366.

[16] National Electrical Manufacturers Association, «Electric Power Systems and Equipment-Voltage Ratings (60 Hertz),» ANSI, Virginia, USA, 2005.

[17] The Institute of Electrical and Electronics Engineers, «IEEE Std 1547 - Standard for Interconnection and Interoperability of Distributed Energy Resources with Associated Electric Power Systems Interfaces,» IEEE, New York, 2018.

[18] International Electrotechnical Commission, «IEC 61727 - Photovoltaic (PV) systems. Characteristics of the utility interface,» IEC, Geneva, Switzerland, 2004.

[19] National Institute of Standards and Technology, «National Institute of Standards and Technology,» 04 Mayo 2023. [En línea]. Available: https://catalog.data.gov/dataset/nist-campus-photovoltaic-pv-arrays-and-weather-station-data-sets-05b4d.

[20] Datasheet, «Datasheet,» 04 Mayo 2023. [En línea]. Available: https://www.datasheets.com/en/part-details/nu-u235f2-sharp-46351940#datasheet.

[21] SolarDesignTool, «SolarDesignTool,» 04 Mayo 2023. [En línea]. Available: http://www.solardesigntool.com/components/inverter-grid-tie solar/PVPowered/137/PVP260kW/specification-data-sheet.html.

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

[23] J. VanderPlas, Python Data Science Handbook - Essential Tools for Working with Data, Sebastopol, CA: O'Reilly Media, Inc., 2017.

[24] S. Raschka y V. Mirjalili, Python Machine Learning - Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow, Birmingham: Packt Publishing Ltd., 2017.

[25] A. Lopez-Martinez, J. M. Juarez, M. Campos y B. Canovas-Segura, «VLSD—An Eficient Subgroup Discovery Algorithm Based on Equivalence Classes and Optimistic Estimate,» *Algorithms* - MDPI, vol. 16, nº 6, pp. 1-26, 2023. https://doi.org/10.3390/a16060274.

[26] L. W. Rizkallah y N. M. Darwish, «An Analysis of Subgroup Discovery Quality Measures,» JOURNAL OF ENGINEERING AND APPLIED SCIENCE, pp. 109-131, 2019.

[27] M. Meeng y A. Knobbe, «For real: a thorough look at numeric attributes in subgroup discovery,» *Data Mining and Knowledge Discovery*, vol. 35, p. 158–212, 2021. https://doi.org/10.1007/s10618-020-00703-x.

[28] D. Cielen, A. D. B. Meysman y M. Ali, Introducing Data Science, Shelter Island, NY: Manning Publications Co., 2016.