

Irradiance Acquisition in Real Time with Long Term Data Logger and Postprocessing using Data Mining Methods

Aracely Zapién-Castillo , José Angel Zumaya-García , Pedro Martín García-Vite , *Member, IEEE*, Luciano Aguilera-Vázquez , and Francisca Hernández-Angel 

Abstract—Nowadays, there has been an increase in the usage of renewable energies, due to the environmental situation in our planet. Among the existing renewable energies, the photovoltaic energy is an attractive alternative. However, it has an intermittent production due to the environmental factors involved in its process. Particularly partial shading can make photovoltaic arrays consume energy instead of producing it. Also, it is useful to register the environmental factors, in order to implement control measurements. For this reason, in order to measure and register real-time temperature and irradiance precisely, a low-cost Data Acquisition System (DAS) is proposed. The implemented DAS has a wireless communication system, which allows the user to have a database through an interface. Subsequently, the available power is calculated through a simplified mathematical model, and used along with the irradiance and temperature as attributes. Finally, the data is classified through Data Mining techniques, such as k-means and ID3, and latter used to predict whether the bypass diode remains connected or not.

Index Terms—Solar Energy, Data Acquisition, Data Mining, Photovoltaic Panels, Irradiance, Temperature.

I. INTRODUCCIÓN

En las últimas décadas, debido al uso excesivo de las fuentes de generación de energía tradicionales, ha habido un incremento de los gases del efecto invernadero. Con el fin de contrarrestar este efecto negativo, han surgido nuevas propuestas que buscan reducir el uso de fuentes de generación contaminantes. Estas nuevas alternativas, además de ayudar en la lucha contra el cambio climático, son de bajo costo y asequible a la población. Dichas propuestas son las conocidas energías renovables y su uso ha venido aumentando a lo largo de los años. Una de las fuentes de energía renovable que más auge ha tenido es la solar Fotovoltaica (FV), debido principalmente a su reciente abaratamiento, bajo mantenimiento, fácil instalación, aumento en la eficiencia y vida útil de hasta 40 años [1].

Como parte de las propuestas para maximizar el aprovechamiento de la energía solar FV, se han creado sistemas de medición para los factores ambientales que afectan en

mayor medida la generación energética, en especial la temperatura e irradiancia como el reportado en [2]. De igual manera, en [3] se presenta el diseño de un sistema capaz de monitorear paneles FV conectados en cadena gracias a un circuito de conmutación que los desconecta y los vuelve a conectar, permitiendo así diagnosticar ya sea un panel dañado o condiciones de sombreado parcial. Sin embargo, para limitar la desconexión de los paneles, se han desarrollado trabajos que hacen predicciones con los datos de monitoreos previos como en [4] donde se pronostica la irradiancia con un error del 31.8 %. Los datos también se utilizan para implementar modelos matemáticos que permiten alcanzar permanentemente el punto de máxima potencia o incluso para determinar si se desconectan o no los paneles solares de un sistema FV [3]. Para recopilar estos datos, se han reportado trabajos como en [5] y [6], en donde se construyen sistemas de adquisición de datos (SAD) ambientales para un grupo de paneles FV; y algunos otros como [7] y [8] en donde se aplican diversas técnicas de aprendizaje automático para analizar la información de bases de datos ambientales.

A diferencia de los trabajos anteriores donde manejan SAD y análisis de la información recolectada de manera independiente, con el fin de contar con un sistema completo, en este trabajo se presentan dos contribuciones importantes: (i) la implementación de un Sistema de Adquisición de Datos (SAD) con conexión inalámbrica de bajo costo, empleada para crear una base de datos de las variables ambientales en tiempo real, y (ii) un análisis posterior de la información recolectada mediante técnicas de minería de datos, empleando en particular las técnicas k-means e ID3. Este análisis se basa en los datos de irradiancia y temperatura recopilados durante 26 días por medio del SAD propuesto y de la potencia calculada con un modelo matemático desarrollado en Matlab que utiliza la irradiancia y temperatura como datos de entrada. Esto a fin de predecir el momento en que los paneles no producirán energía debido a condiciones de sombreado parcial.

El resto del artículo está organizado como sigue: en la Sección II se presenta una revisión bibliográfica. Luego, en la Sección III, se explica el SAD propuesto, y en la Sección IV la interfaz codificada en Python. Posteriormente, se analizan diversas técnicas de minería de datos en la Sección V. El análisis y descripción del SAD, se realiza en la Sección VI y prosigue en la Sección VII con la aplicación de las técnicas de minería de datos a la información recopilada por el SAD. Por último, se muestran los resultados obtenidos en la Sección

Aracely Zapién-Castillo is with Instituto Tecnológico de Ciudad Madero e-mail:g20073012@cdmadero.tecnm.mx

José Angel Zumaya-García is with Instituto Tecnológico de Ciudad Madero e-mail:g20073007@cdmadero.tecnm.mx.

Pedro Martín García-Vite is with Instituto Tecnológico de Ciudad Madero e-mail:pedro.gv@cdmadero.tecnm.mx

Luciano Aguilera-Vázquez is with Instituto Tecnológico de Ciudad Madero e-mail:luciano.av@cdmadero.tecnm.mx

Francisca Hernández-Angel is with Universidad Autónoma de Tamaulipas and Universidad Politécnica de Altamira e-mail:franyhangel@gmail.com

VIII para finalizar con las conclusiones en la Sección IX.

II. ANTECEDENTES DEL PROBLEMA

A pesar de ser excelentes alternativas, las energías renovables son fuentes intermitentes y discontinuas. En los sistemas FV la potencia eléctrica está estrechamente relacionada con el nivel de factores ambientales incidentes. Estos factores varían a lo largo de 1 día y le impiden al sistema permanecer en el punto de máxima potencia, en especial la temperatura, la irradiancia y el sombreado. Este último es provocado por la sombra de algún cuerpo o el movimiento de las nubes, es decir, es consecuencia de condiciones de irradiancia no uniformes [9]. En tales casos, las celdas fotovoltaicas sombreadas se polarizan inversamente y actúan como una carga para las celdas completamente iluminadas consumiendo energía en lugar de producirla [10]. Debido a que esta energía se disipa en forma de calor causa daños irreversibles en los paneles. Por lo tanto para protegerlos, se instalan diodos bypass o de derivación en la caja de conexiones, los cuales se conectan en paralelo a cada conjunto de celdas o a un panel completo. De esta manera, una vez que su voltaje se vuelve negativo, el diodo se activa, creando un cortocircuito para que circule la corriente a través de él. Regularmente, se conectan varios paneles en serie o paralelo para alcanzar la potencia deseada, en estos casos si un diodo de derivación está activo, cambiará la cantidad de paneles FV activos dentro del sistema reduciendo su potencia de salida drásticamente [11]. Debido a esto, la comunidad científica se ha empeñado en encontrar soluciones que aumenten la eficiencia de los sistemas FV, usando tanto circuitos electrónicos como técnicas de minería de datos. Una revisión exhaustiva desde ambos enfoques es presentada a continuación.

A. Sistema de Adquisición de Datos

Existen diversos estudios en la literatura donde se han diseñado SAD para monitorear el estado de sistemas FV. En [12] se diseñó un registrador de datos de bajo costo para sistemas FV autónomos empleando como microcontrolador un Arduino UNO®. Este sistema mide parámetros eléctricos y meteorológicos, los muestra en un display y los almacena en una memoria SD (Secure Digital).

En [13] se describe una estación de investigación para condiciones de alta irradiancia durante ocho meses. Dicha estación mide irradiancia y temperatura cada diez segundos, posteriormente cada minuto calcula y registra un promedio. Además, genera la curva I-V de los cuatro tipos de paneles solares utilizados. En [14] se mide la irradiancia de 6 a.m. a 7 p.m. usando un seguimiento de doble eje, sensores y un algoritmo para rastrear la posición del sol. Además, se utiliza un SAD comercial el cual recibe señales de control desde un programa de LabView y envía los datos a una interfaz web.

Estos proyectos tienen aportaciones significativas, debido a que, al tener conexiones cableadas, permiten gran velocidad de transferencia de datos, con muy baja interferencia y sin riesgo de comprometer la confidencialidad de la información. Sin embargo, el costo de implementación suele ser elevado y la expansión del sistema compleja. Debido a lo anterior,

TABLA I
COMPARACIÓN DE DIVERSOS PROTOCOLOS DE
COMUNICACIÓN INALÁMBRICA.

Parámetro	ZigBee	Wi-Fi	Bluetooth
Rango	<100m	<100m	<10m
Ancho de banda	<2.4GHz	2.4GHz - 5GHz	<2.4GHz
Fiabilidad	Alta	Alta	Media
Latencia	<20ms	<10ms	<10ms
Uso de potencia	Bajo	Medio	Bajo
Cobertura	Baja	Baja	Baja
Costo por módulo	\$1 - \$5	\$1 - \$10	\$2 - \$5

se han desarrollado sistemas inalámbricos que se instalan rápidamente y permiten agregar nodos en condiciones extremas o ubicaciones geográficas mucho más amplias, incluso en altamar. Un buen ejemplo es el mostrado en [15], donde utilizan un SAD inalámbrico, que mide variables eléctricas y ambientales en una estación de bombeo FV. Además, mediante una placa Arduino y un módulo GSM (Global System for Mobil Communications) controlan y monitorean el sistema a través de mensajes SMS (Short Message Service), sin embargo, no permite un registro completo de datos para postprocesamiento.

Por otro lado, ciertamente existen desventajas en las redes inalámbricas, sin embargo, si se selecciona la tecnología adecuada para el sistema propuesto se pueden minimizar los inconvenientes. Actualmente, las tecnologías inalámbricas más destacadas son Bluetooth, WiFi y ZigBee, las cuales son adecuadas para transmitir información en sistemas de adquisición de datos [16]. Cada una de ellas cuenta con características específicas, las cuales se comparan en la Tabla I.

ZigBee es similar a la tecnología Bluetooth, pero destaca debido a que su distancia de transmisión de datos es de hasta 100 metros, es decir, 10 veces mayor que la distancia de transmisión de Bluetooth. Además Zigbee tiene una alta fiabilidad de transferencia y un buen precio, por lo que es ampliamente utilizado como protocolo de comunicación en SAD [17]. En [18] se muestra un buen ejemplo donde la comunicación entre paneles FV y la computadora principal se realiza mediante este protocolo de manera exitosa.

B. Análisis de Datos

Una vez construido el SAD, usualmente se continúa con el análisis de los valores obtenidos. El análisis y procesamiento de datos en sistemas de energía solar FV, generalmente se realiza mediante técnicas estadísticas o de minería de datos, según la literatura.

En relación a las técnicas estadísticas, [19] propone una metodología simple pero eficaz para elegir una señal de irradiancia diaria representativa de un conjunto fijo de 294 días, donde toman en cuenta la variabilidad de la irradiancia y la energía total diaria producida. La variabilidad la definen como la velocidad a la cual la irradiancia cambia a lo largo del día. El criterio de elección del valor representativo de irradiancia fue el que más se acercara al promedio de la variabilidad y de la energía total. Asimismo, la selección de datos fue también útil para considerar días con valores muy altos de irradiancia. Similarmente, en [20] se reportan

valores de temperatura, humedad relativa e irradiancia para utilizarse en la planificación de energía FV. Los sensores se conectan a un sistema Arduino y los datos se almacenan en una memoria SD. Posteriormente, los datos recopilados se procesan con la técnica de Correlación de Pearson y el método de Mínimos Cuadrados Ordinarios en Matlab para obtener valores predictivos de irradiancia, además calculan el Error Porcentual Absoluto Medio de los valores reales de irradiancia contra los valores predictivos calculados.

Respecto a los trabajos que incluyen minería de datos, se encuentra [21] y [22] donde en ambos se predice la potencia eléctrica que generarán los sistemas FV con la técnica de máquinas de soporte vectorial (MSV). En [21] la potencia se predice con una hora de anticipación gracias al uso de redes neuronales, mientras que en [22] mejoran el desempeño de las MSV utilizando el algoritmo de colonia de abejas artificiales. En otro trabajo reportado [23], se extrae un conjunto de valores de irradiancia horizontal global, directa y difusa, temperatura de bulbo seco y humedad relativa de una base de datos de acceso público procedente de la localidad “Nevada Power Clark Station” [24]. Dichos datos se tomaron de un día y fueron guardados en intervalos de un minuto. Se emplearon dos técnicas de minería de datos (k -NN y Naive Bayes) para modelar la irradiancia horizontal global, reportando la técnica de k -NN superior con base al porcentaje de Error Medio Absoluto. En [25] emplean un método basado en un análisis de agrupamiento múltiple junto con el modelo de distribución de probabilidad de los datos para modelar la potencia generada de un sistema FV. Además, se analizan dos grupos de datos: unos en escala de horas y otros en escala de minutos, y para ambas muestras de datos se calculan sus correspondientes centroides.

En este trabajo, se aborda desde la construcción del SAD para obtener una propia base de datos y se continúa con el posterior análisis de dichos datos, que incluye un análisis estadístico y un procesamiento con minería de datos.

III. DISEÑO SAD PROPUESTO

Diversos factores modifican el punto de máxima potencia de la curva característica de un panel FV. Los principales son la irradiancia y la temperatura, por tanto, el SAD propuesto adquiere los datos de estas variables, las cuales inciden en un sistema FV de 24 V del tipo aislado. Tal como se muestra en la Fig. 1, el procesador del SAD es el microcontrolador Atmega328P. En él se encuentra un programa escrito en lenguaje C, que se encarga de la lectura de los sensores, el procesamiento de las señales y el despliegado de información en unidades correspondientes, en un display de cristal líquido (LCD) 20x4, donde además se exhiben la hora y la fecha. Pese a que el microcontrolador es de 8 bits, tomando un voltaje de referencia de 5 V, la conversión analógico-digital la realiza sobre una base de hasta 10 bits, permitiendo una resolución de 5/1024 V/bit.

Para la temperatura se emplea el módulo DHT11. Este contiene un sensor analógico con un intervalo de medición de 0° a 50°C y un voltaje de operación de 3.3–5 V. Dentro del módulo se hace la conversión analógico-digital y se envía una trama al microcontrolador con los datos. Para la

irradiancia se utiliza un sensor de radiación solar RK200-04, el cual tiene un intervalo de operación de 0–1500 W/m², se alimenta con 12 V y salida de 4–20 mA. Antes de conectarse al microcontrolador pasa por una etapa de acondicionamiento de señal para convertir la salida de corriente de 4–20 mA a una señal de 1–5 V capaz de ser leída y convertida a señal digital (0–1023) por el Atmega328P. Al utilizar este intervalo como señal de salida aun cuando la irradiancia detectada por sea nula, llegará un voltaje que confirma el funcionamiento del sensor, lo cual no se sabría si el offset fuera cero.

Simultáneamente, tal como se muestra en el diagrama de flujo de la Fig. 2, cada minuto el sistema almacena el valor de irradiancia y temperatura en la memoria EEPROM externa AT24C256. Este dispositivo dispone de 256 kB de memoria organizada en 32,768 palabras por 8 bits y se puede acceder a ella por medio del bus circuito ínter-integrado o también llamado I2C, que es un estándar de comunicación entre microcontroladores. Adicionalmente, cuenta con un módulo DS1307 el cual es un reloj de tiempo real (RTC). Éste trabaja a 32.7 kHz y posee una batería de litio que permite tener el registro de la hora exacta sin importar que el SAD no esté conectado a una computadora. La información almacenada cada minuto ocupa un espacio de 8 bytes de memoria, 4 corresponden a la hora y al minuto de muestreo, 2 a la irradiancia y 2 a la temperatura. Debido a que el SAD trabaja 15 horas, de las 6:00 a.m. a las 9:00 p.m., y hace el muestreo por minuto, se ocupan 7.2 kB de memoria por día. Lo que significa que se pueden almacenar los datos de hasta 4.5 días.

Al cabo de este periodo, o en el momento que el usuario desee, se puede exportar esta información a una computadora de forma inalámbrica. Para tal propósito se utiliza el protocolo ZigBee mediante módulos de radiofrecuencia llamados Xbee, conectados al microcontrolador (Fig. 1).

El flujo de información es bidireccional; enviando y recibiendo la información almacenada cuando es solicitada por el usuario. El Atmega328P procesa los datos sensados y forma una sucesión de bits organizados llamada trama que se envía cuando el Xbee coordinador la requiera. Esta comunicación se lleva a cabo en Modo API (Application Programming Interface), el cual permitirá a futuro que varios SAD envíen su información al mismo Xbee Coordinador mediante una red de estrella.

En el mercado existen diversos SAD comerciales, sin embargo el propuesto tiene características diferenciales. Por ejemplo, el propuesto cuenta con RTC, display que permite ver en tiempo real las variables sensadas e interfaz en código abierto, a diferencia del Sistema de Adquisición Powerlab 4/26 de AD Instruments que requiere un software especializado.

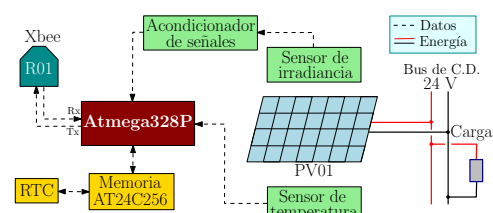


Fig. 1. Diagrama de bloques del SAD propuesto.

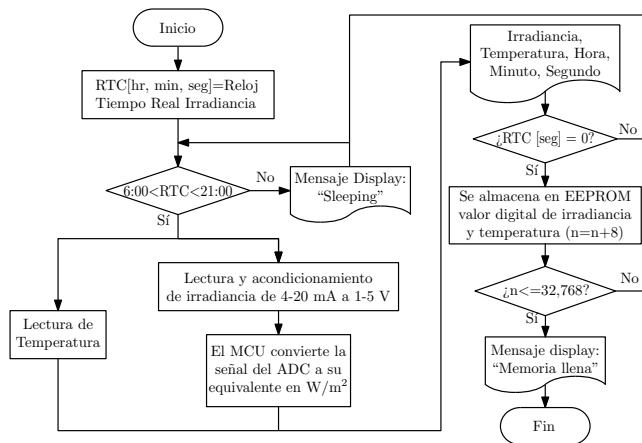


Fig. 2. Diagrama de flujo del proceso de adquisición de datos.

Asimismo, el prototipo tiene una alimentación de 12 V para conectar sensores de irradiancia, contrario a la NI USB-6009 de National Instrument que solo permite conectar sensores de 5 V. Además, la portabilidad del SAD e interfaz permiten extraer inalámbricamente la información almacenada, a diferencia del Sistema de Monitoreo Comet MS6R de Comet Instruments, el cual trabaja alámbricamente. Pero principalmente destaca su costo asequible de 125 USD, muy inferior al de las alternativas mencionadas que van desde 550 USD hasta 12,000 USD.

IV. INTERFAZ

Por medio de la interfaz gráfica mostrada en la Fig. 3 y desarrollada en Python, el cliente es capaz de comunicarse con el SAD para obtener la información almacenada en la memoria EEPROM externa. Este proceso se lleva a cabo inalámbricamente desde una computadora, ya que conectado a su puerto USB se encuentra un módulo Xbee coordinador, tal como lo presenta la Fig. 4.

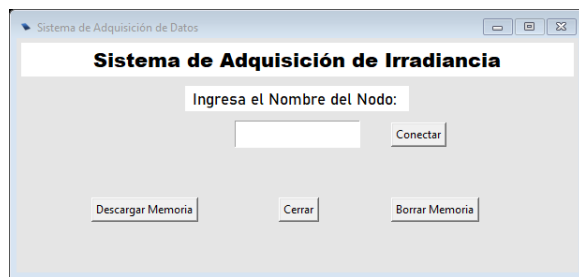


Fig. 3. Interfaz gráfica para extraer la información del SAD.

Aunque este proyecto está integrado por una red de un solo nodo, en casos donde se cuente con una red más amplia el cliente puede conectarse al nodo que desee a través de la interfaz gráfica. Para conocer la información almacenada primero se debe entablar la conexión entre el Xbee coordinador y el Xbee Router. Después de enlazarlos exitosamente se procede a descargar los datos de la memoria. Al hacer clic en el botón Descargar Memoria, el programa hace la solicitud por radiofrecuencia, la cual es recibida por el Xbee Router que se encuentra conectado al microcontrolador del SAD. En ese momento, el Atmega328P contesta organizando la información

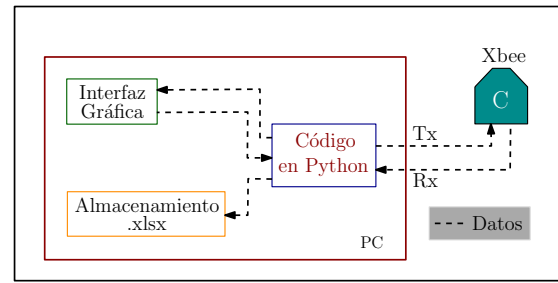


Fig. 4. Proceso interno en computadora personal.

de la memoria en tramas, las cuales se transfieren a gran velocidad a través de los Xbees. Iniciando desde la localidad de memoria 0, hasta la que contiene el último dato guardado. Estas secuencias de bits son recibidas por el Xbee coordinador y desempaquetadas por medio del programa codificado en Python. La interfaz gráfica muestra la recepción de los datos a gran velocidad. Estos datos son valores de 0 a 1023, los cuales no significan nada en términos de irradiancia por lo cual dentro del programa se hace una conversión. Debido a que el sensor de irradiancia RK200-04 tiene una señal de salida 4–20mA, que corresponde a un intervalo de 0–1500 W/m² de irradiancia (Δ_G), el rango (0-1023) se reduce y el valor de 204 representa la irradiancia 0 (G_{off}), mientras que 1023 el máximo de 1500 W/m². Para esto se considera (1), en donde ADC_n es el valor enviado por el microcontrolador y G_{med} es el valor medido de irradiancia en W/m². De esta manera, la irradiancia tiene una resolución de 819/1024 (W/m²)/b.

$$G_{med} = \frac{\Delta_G(ADC_n - G_{off})}{(2^{10} - 1 - G_{off})} \quad (1a)$$

$$G_{med} = \frac{1500(ADC_n - 204)}{819} \quad (1b)$$

Cuando se termina este proceso, todos los datos recibidos se almacenan de forma organizada y secuencial en un archivo .xlsx para su posterior análisis. A dicho archivo se le nombra según la hora y la fecha de descarga de la información. Posteriormente, el proceso de muestreo se reinicia de forma automática, a la espera de que el cliente vuelva a solicitar la información acumulada. Este proceso se representa en el diagrama de flujo de la Fig. 5.

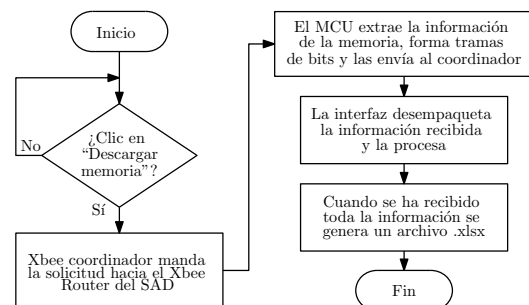


Fig. 5. Diagrama de flujo para la comunicación inalámbrica.

V. MINERÍA DE DATOS

Similar a [25], donde emplean minería de datos para agrupar los valores, en este trabajo es necesario preprocesar los datos, ya que con el SAD solo se obtienen los valores sin estar previamente clasificados, es decir una cadena de valores secuenciales. Por ello, para clasificar los datos se emplea la técnica de agrupamiento de k -means, destacando esta técnica entre las 10 más utilizadas [26].

El algoritmo de k -means es un método iterativo de aprendizaje no supervisado, donde los datos no se encuentran clasificados inicialmente. Por lo que el algoritmo se encargará de agrupar en k clases al conjunto de datos de entrada a través de los siguientes pasos:

1. *Asignar k* : k es el número de grupos en los que se desea clasificar el conjunto de datos, y se asigna de manera arbitraria según el contexto de los datos.
2. *Definir los centroides iniciales*: serán k centroides iniciales definidos, y se pueden elegir de manera aleatoria del mismo conjunto de datos.
3. *Asignar los datos al centroide más cercano*: a cada valor del conjunto de datos se le asigna el centroide más cercano, resultando así la separación de los datos en k grupos. La distancia de los datos a los centroides se calcula generalmente con la distancia Euclideana [25].
4. *Reubicación de centroides*: cada centroide representativo se reubica al centro (la media) de todos los datos asignados a dicho centroide. El algoritmo concluye cuando la ubicación de los centroides no cambia.

Una vez clasificados los datos obtenidos con el algoritmo k -means, es necesario continuar con una técnica de minería de datos que aprenda a clasificar nuevos datos, empleando el conjunto de datos ya clasificado como conjunto de entrenamiento. Para la segunda parte del análisis se optó por utilizar la técnica de árboles de decisión o ID3 (por sus siglas en inglés, Induction Decision Trees), ya que al igual que k -means, es una de las 10 técnicas de minería más empleadas y es sencilla de implementar [26].

La técnica de árboles de decisión es un algoritmo de aprendizaje supervisado que utiliza un conjunto de datos de entrenamiento para posteriormente ser capaz de clasificar nuevos valores. Partiendo del conjunto de entrenamiento que posee diversos atributos y una clase asignada a cada punto, la técnica ID3 desarrolla reglas de decisión que clasifican cada punto (o caso) del conjunto de entrenamiento, y posteriormente, será capaz de predecir la clase de nuevos datos.

Para saber cuál es el atributo más significativo en cada caso, se utilizan dos conceptos: entropía y ganancia. La entropía se puede definir como la incertidumbre que se tiene sobre la clase a la que pertenecerá un ejemplo escogido al azar del conjunto de datos [27]. Con base en el concepto de entropía, la ganancia se define como la disminución de entropía que se produce al dividir el conjunto de datos de entrenamiento según los distintos valores v_i de un atributo A [27]. Para llevar a cabo el algoritmo ID3, considerando los conceptos anteriores se siguen los siguientes pasos:

1. *Construir tablas de frecuencia*: con base en el conjunto de entrenamiento, se construyen n tablas de frecuencias,

siendo n el número de atributos. Asimismo, cada tabla de frecuencia debe contener los valores de A y sus correspondientes clasificaciones.

2. *Calcular la entropía E del conjunto S* : definida por (2), donde S es el conjunto de datos, C el conjunto de clases y p_i la probabilidad de que un valor dado pertenezca a la clase i ,

$$E(S) = \sum_{i \in C} -p_i \log_2(p_i) \quad (2)$$

3. *Calcular la ganancia por atributo*: en (3) se describe la ganancia por atributo, donde V es el conjunto de valores que puede tomar el atributo A mientras que S_{v_i} es el subconjunto de datos de S cuyo atributo A tiene el valor v_i , es decir, el segundo término de la diferencia en la ganancia es la suma de las entropías de los conjuntos resultantes, ponderadas por el tamaño de dichos conjuntos con respecto al inicial [27],

$$G(S, A) = E(S) - \sum_{v_i \in V} \frac{|S_{v_i}|}{|S|} E(S_{v_i}). \quad (3)$$

4. *Construir el árbol de decisión*: para realizar este paso primero se elige la ganancia más alta, la cual definirá el atributo o la nueva hoja del árbol de decisión. Con esto, las tablas de frecuencia se harán más pequeñas por hoja, y los pasos del 1 al 3 deberán repetirse hasta que exista homogeneidad en las tablas resultantes.
5. *Simplificar el árbol de decisión*: este paso se realizará en cuanto las tablas de frecuencia sean homogéneas, ya que solo hay que representar las hojas del árbol para facilitar la lectura del mismo.
6. *Construir las reglas de decisión*: hay que recorrer el árbol final de izquierda a derecha y de arriba hacia abajo, estableciendo todas las combinaciones posibles.

Teniendo el árbol de decisión construido, será posible clasificar nuevos valores, tomando en cuenta que los valores seguirán alguna de las reglas de decisión del árbol.

VI. IMPLEMENTACIÓN SAD

El SAD se puso en funcionamiento en el municipio de Tampico, Tamaulipas, México, ubicado en las coordenadas 22°13'06.6"N 97°50'42.7"W entre los días 28 de junio y 10 de agosto del 2021. El sensor de irradiancia se colocó en la azotea del edificio, tal como lo muestra la Fig. 6, a fin de evitar obstáculos. Mientras que el sensor de temperatura se colocó en un lugar estratégico en la terraza del edificio, permitiendo capturar la temperatura propia de la celda. La información se registró de 6 a.m. a 9 p.m., realizando 900 muestreos de irradiancia y temperatura por día.

El SAD propuesto se construyó en un circuito impreso, como se aprecia en la Fig. 7. Este se colocó dentro de las instalaciones del edificio a fin de protegerlo de las inclemencias del tiempo. El circuito tiene una fuente de 12 V la cual alimenta al sensor de irradiancia. Mientras que un regulador de voltaje LM7805, acompañado de dos capacitores de 10 μ F, generan 5 V para alimentar al resto de los componentes electrónicos. Entre estos se encuentra el microcontrolador Atmega328P con un cristal de 16 MHz que genera los pulsos de reloj, dos

capacitores de 22pF de desacoplo y su botón de reset. También cuenta con un LCD 20x4, un módulo Xbee Router, un módulo AT24C256 que contiene memoria EEPROM externa y finalmente el RTC en el módulo DS1307. Asimismo, al circuito se conectan los sensores DHT11 y RK200-04 para medir la temperatura e irradiancia, respectivamente. Este último, antes de conectarse al microcontrolador, pasa por una etapa de acondicionamiento de señal constituida por una resistencia de precisión de 250 Ω . Las unidades de las variables son W/m^2 para irradiancia y $^{\circ}C$ para la temperatura.

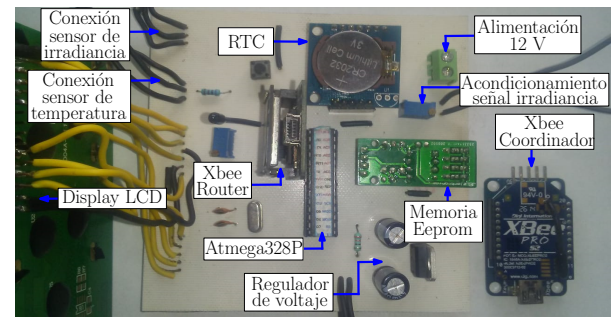


Fig. 7. Fotografía del circuito propuesto del SAD.

VII. ANÁLISIS E INTERPRETACIÓN DE DATOS

Una vez recopilados los datos del intervalo de fechas mencionadas, se introdujeron directamente los datos crudos al programa de Minitab® para elaborar un análisis estadístico rápido de los mismos. Con la opción de *Resumen Gráfico* se muestra el histograma de la distribución de los datos, así como la media, desviación estándar, curtosis, cuartiles y el valor p de los datos. Este último indica si se sigue o no una distribución normal, es decir, si $p < 0.05$, los datos no son normales. La anomalía de los datos generalmente se debe a alguna anomalía en el proceso o a que la simple naturaleza de los datos no sigue una distribución normal. Para el caso particular de la irradiancia y temperatura durante el día no siguen una distribución normal, ya que estas variables fluctúan hora a hora dependiendo de las condiciones climáticas. Por lo tanto, la anomalía de estos datos está justificada. Este análisis estadístico es útil para conocer los valores promedios de la irradiancia y de la temperatura diarios en la zona, y a su vez permite visualizar la distribución de los datos, sus medidas de tendencia central y dispersión, todo con la finalidad de conocer su comportamiento. En la Fig. 8 se puede visualizar un resumen gráfico obtenido del día 23 de julio de 2021, donde se evidencia la anomalía de los datos de irradiancia a lo largo del día, así como sus medidas de tendencia central y dispersión.

Adicional a la irradiancia y la temperatura, la potencia extraída del panel PV se toma en cuenta como tercer atributo. Para el cálculo de la potencia, se consideran los coeficientes de temperatura de voltaje y de corriente [28] α_{SC} y β_{OC} , respectivamente, proporcionados por el fabricante bajo condiciones estándares de prueba. El modelo matemático que describe el funcionamiento del circuito mostrado en la Fig. 9, es el siguiente:



Fig. 6. Sensor RK200-04 adquiriendo datos de irradiancia.

$$I_{fv} = I_{ph} - I_d \quad (4a)$$

$$I_{ph} = \frac{G}{G_n} [I_n + \alpha_{SC} \Delta T] \quad (4b)$$

$$I_d = I_{os} \left[\exp \left(\frac{q(V_{fv} + I_{fv} \cdot R_s)}{AKT_j} \right) - 1 \right] \quad (4c)$$

donde $\Delta T = T_j - T_n$, y los demás parámetros se muestran en el apéndice A.

Después se desarrolló en Python un código para aplicar la técnica de k -means, con ayuda de la librería `sklearn.cluster`. Los datos se representaron gráficamente en 3 dimensiones y se eligió el valor de $k = 2$. Este valor se basó en [29] donde a la intensidad de la sombra se le denomina factor de sombreado y su valor oscila entre 0 (sin sombra) y 1 (sombra completa). Sin sombra significa que toda la irradiancia disponible incide sobre el módulo y sombra completa considera que la sombra se proyecta sobre todas las celdas del módulo FV. Al saber si existe sombra o no, es posible conocer el momento en que los diodos bypass entrarán

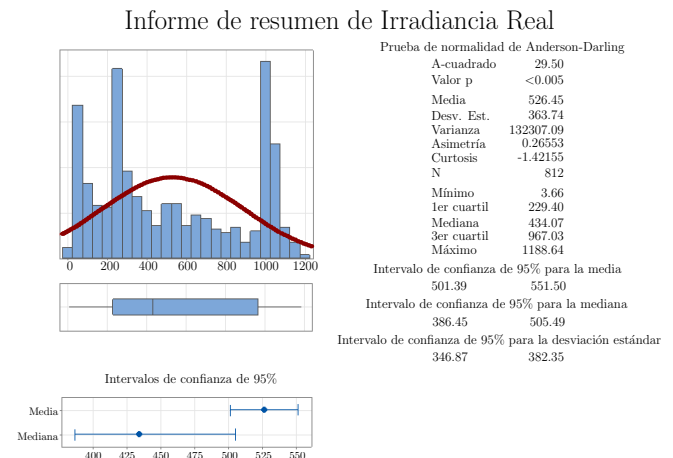


Fig. 8. Resumen gráfico de irradiancia del 23 de julio de 2021.

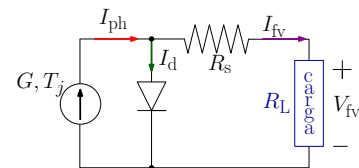


Fig. 9. Circuito eléctrico equivalente del panel fotovoltaico.

en funcionamiento y por lo tanto predecir cuándo bajará la potencia generada por el sistema FV. Por tanto, las 2 clases empleadas en el código son “diodos bypass desactivados” y “diodos bypass activados”. El código de Python generó la Fig. 10 basada en los datos obtenidos del SAD y en el algoritmo implementado, y arrojó dos centroides: C_1 :(40.71, 692.12, 0.27) y C_2 :(42.28, 747.43, 0.30), donde los ejes (x, y, z) están dados por temperatura, irradiancia y potencia, respectivamente. Además, se generó un archivo .xlsx con los valores de temperatura, irradiancia, y potencia ya clasificados.

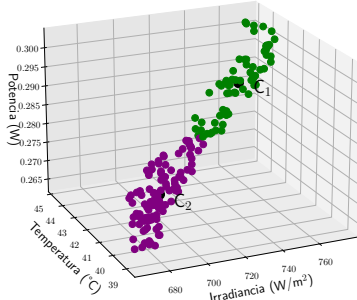


Fig. 10. Datos de irradiancia, temperatura y potencia clasificados con k -means.

El archivo .xlsx previamente generado se utilizó como conjunto de entrenamiento de la técnica ID3, implementada con la ayuda de la librería `DecisionTreeClassifier`. Esta es capaz de realizar una clasificación en un conjunto de datos con varias clases. Para esto toma como entrada dos matrices: una matriz X que contiene las muestras de entrenamiento, en este caso la irradiancia, temperatura y potencia, y una matriz Y que contiene las etiquetas de clase para las muestras de formación, es decir diodos bypass desactivados o activados.

Con estas matrices se entrenó el programa y se generó el árbol de decisiones de la Fig. 11. Este comienza con una condición donde se evalúa si la irradiancia es menor o igual a 719.82 y a su vez si la potencia es menor o igual a 0.283. Cuando es verdadera la condición toma la rama de la izquierda, caso contrario la de la derecha, es decir indica si pertenecerá a la clase diodos bypass desactivados o activados.

La variable G_{ini} representa el porcentaje de respuestas similares y utiliza (5), donde x es el número de respuestas positivas (“Sí”), y el número de respuestas negativas (“No”) y n es el número de muestras. Por otro lado, si se desea, el árbol de decisión también se puede exportar en formato textual con la función `export text` para que se pueda apreciar como un conjunto de reglas.

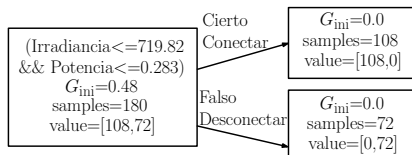


Fig. 11. Árbol de decisión generado con programa de Python.

$$G_{ini} = 1 - \left(\frac{x}{n}\right)^2 - \left(\frac{y}{n}\right)^2 \quad (5)$$

La técnica ID3 también se puede utilizar para predecir la clase a la que pertenece una muestra dada por el usuario. Es decir, se ingresan datos de irradiancia, temperatura y potencia al programa ya entrenado y mediante la instrucción `dtree.predict` Python predice la clase correspondiente. En este caso señala si los diodo bypass se encuentran desactivados o activados con base a las condiciones ambientales y eléctricas introducidas al programa. Este proceso se llevó a cabo con los paneles FV monocristalinos marca “EPCOM POWER LINE®” modelo PROSE-5012, cuyos parámetros a STC son: $P_{mp} = 50$ W, $V_{mp} = 17.5$ V, $I_{mp} = 2.86$ A, $V_{OC} = 21.5$ V, $I_{SC} = 3.01$ A.

VIII. RESULTADOS

El SAD de temperatura e irradiancia propuesto se integró dentro de un gabinete plástico para garantizar la fiabilidad de la información recopilada. Así como se muestra en la Fig. 12, en la parte delantera cuenta con conectores tipo banana hembra que facilitan la conexión y desconexión de los sensores. Además, cuenta con una amplia pantalla LCD que permite visualizar el proceso de la recolección de datos. Por la parte trasera se encuentra el interruptor de encendido, el LED indicador y una fuente de alimentación de 12 V. Aunque algunos componentes se alimentan con 5 V, dentro del circuito se regula el voltaje para evitar usar 2 fuentes de alimentación.

Como se mencionó en la Sección Implementación, el sistema se puso en funcionamiento durante parte de los meses de junio, julio y agosto del 2021. Logrando recopilar más de 20,000 datos de irradiancia y temperatura para formar una base de datos de la zona sur de Tamaulipas, México. Con los datos recopilados a lo largo de un día se generan gráficas, como la presentada en la Fig. 13, que permite observar claramente el comportamiento de la irradiancia. Es posible hacer lo mismo con los datos de temperatura.

Una vez obtenidos los datos se evaluaron estadísticamente, y se observó que tanto los datos de irradiancia como los de temperatura no siguen una distribución normal, debido a su gran variabilidad a lo largo del día. El promedio de irradiancia diaria durante el periodo evaluado fue de 465.46 W/m², valor que es relativamente bajo, pero entendible ya que algunos de los días evaluados fueron nublados.

El promedio de temperatura durante el periodo fue de 33.76 °C, cuyo valor es normal para la zona evaluada. Sin embargo, el máximo aprovechamiento de estos datos se obtiene mediante la implementación de técnicas de minería de datos. Con el



Fig. 12. Fotografía del prototipo del SAD portátil propuesto.

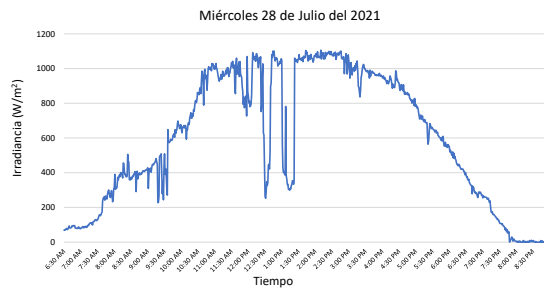


Fig. 13. Gráfica de irradiancia del día 28 de julio del 2021.

algoritmo de k -means se clasificaron los datos en 2 grupos: diodos bypass activados o desactivados. Dicho criterio se tomó en cuenta considerando los valores de irradiancia, temperatura y potencia, ya que si el sistema se encuentra sometido a cierta irradiancia y temperatura, y genera cierta potencia, se podrá determinar si los diodos bypass se encontrarán desactivados o si se activarán automáticamente y disminuirá la potencia generada por el sistema. El grupo 0 que k -means generó representa el sistema con los diodos bypass activados, mientras que el grupo 1 representa el sistema con los diodos bypass desactivados.

Después de aplicar la técnica de ID3 y obtener que el 60 % de los datos fueron clasificados en el grupo 0, y por consiguiente el 40 % de los datos fueron clasificados en el grupo 1, con ayuda de la instrucción `dtree.predict` de Python es posible predecir futuros valores. El usuario tiene la posibilidad de ingresar los valores de irradiancia, temperatura y potencia que desea evaluar, y determinar si se activan o desactivan los diodos bypass.

IX. CONCLUSIONES

A fin de que en los próximos años el porcentaje de energía producida por fuentes renovables sea mayor a la forma tradicional, y adicionalmente se incremente su eficiencia, es necesario evaluar las variables ambientales involucradas. Para el caso de la generación de energía solar fotovoltaica las variables de interés son la irradiancia y la temperatura, ya que dependiendo de su valor, será la cantidad de potencia disponible para ser aprovechada por los algoritmos de seguimiento de potencia en los sistemas FV. Para obtener datos reales de irradiancia y temperatura es necesario construir un SAD, capaz de adquirir en tiempo real el valor de estas variables, y tener un registro que pueda ser consultado para su posterior análisis. Esta propuesta incluye además la interconexión inalámbrica que permite el acceso a los datos de forma segura y remota.

El análisis posterior puede llevarse a cabo con estadística para evaluar su valor promedio, y también con técnicas de minería de datos, que permitan clasificar los datos obtenidos, y datos futuros. Para mejorar el uso de la minería de datos, además de la irradiancia y la temperatura, se agrega el atributo de potencia, para evaluar la energía eléctrica que el sistema sería capaz de generar. Por lo que es necesario resolver el modelo que representa el comportamiento del panel. Con la técnica de k -means los datos adquiridos por medio del SAD construido fueron clasificados en grupo 0 y grupo 1 (diodo

bypass activado o desactivado). Es importante no solo saber cuándo el sistema FV se encuentra generando la máxima potencia, sino también predecir cuándo disminuirá la energía generada por el sistema para así implementar estrategias que permitan suplir en todo momento la demanda energética.

APÉNDICE

PARÁMETROS DE MODELO MATEMÁTICO FOTOVOLTAICO

T_j : Temperatura incidente en el panel fotovoltaico

T_n : Temperatura nominal (25°C)

I_{ph} : Corriente fotogenerada

I_d : Corriente del diodo

G_n : Irradiancia nominal (1000 W/m^2)

I_n : Corriente de cortocircuito a condiciones nominales

I_{os} : Corriente de saturación inversa del diodo

q : Carga del electrón ($1.602 \times 10^{-19} \text{ C}$)

V_{fv} : Voltaje fotovoltaico

R_s : Resistencia parasitaria en serie

A : Factor de idealidad del diodo

K : Constante de Boltzmann ($1.381 \times 10^{-23} \text{ J/K}$)

AGRADECIMIENTOS

Los autores agradecen al TECNOLÓGICO NACIONAL DE MÉXICO por el apoyo otorgado para la realización de este proyecto bajo el registro: 13431.21-P.

REFERENCIAS

- [1] Á. A. Bayod Rújula, *Energías renovables: Sistemas fotovoltaicos*. Zaragoza: Prensas Universitarias de Zaragoza, 2009.
- [2] J. Choi, M. Choi, Y. Shin, and I.-W. Lee, "Design of Web-based Monitoring System for Solar Photovoltaic Power Plants," in *2020 International Conference on Information Networking (ICOIN)*, pp. 784–786, 2020.
- [3] M. Mutillo, T. de Rubeis, D. Ambrosini, G. Barile, and G. Ferri, "Sensor monitoring system for PV plant with active load," in *2019 IEEE 8th International Workshop on Advances in Sensors and Interfaces (IWASI)*, pp. 124–127, 2019.
- [4] A. Shakya, S. Michael, C. Saunders, D. Armstrong, P. Pandey, S. Chalise, and R. Tonkoski, "Solar Irradiance Forecasting in Remote Microgrids Using Markov Switching Model," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 3, pp. 895–905, 2017.
- [5] A. C. Killam, J. F. Karas, A. Augusto, and S. G. Bowden, "Monitoring of Photovoltaic System Performance Using Outdoor Suns-VOC," *Joule*, vol. 5, no. 1, pp. 210–227, 2021.
- [6] J. Walters, S. Guo, E. Schneller, H. Seigneur, and M. Boyd, "PV module loss analysis using system in-situ monitoring data," in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC)*, pp. 2204–2208, 2018.
- [7] A. Eschenbach, G. Yepes, C. Tenllado, J. I. Gómez-Pérez, L. Piñuel, L. F. Zarzalejo, and S. Wilbert, "Spatio-Temporal Resolution of Irradiance Samples in Machine Learning Approaches for Irradiance Forecasting," *IEEE Access*, vol. 8, pp. 51518–51531, 2020.
- [8] E. Basha, R. Jurdak, and D. Rus, "In-Network Distributed Solar Current Prediction," *ACM Trans. Sen. Netw.*, vol. 11, Dec. 2015.
- [9] M. Nahidan, M. Niroomand, and B. Mirzaeian Dehkordi, "Power enhancement under partial shading condition using a two-step optimal pv array reconfiguration," *International Journal of Photoenergy*, vol. 2021, pp. 1–19, 01 2021.
- [10] S. Ghosh, V. K. Yadav, and V. Mukherjee, "Impact of environmental factors on photovoltaic performance and their mitigation strategies—a holistic review," *Renewable Energy Focus*, vol. 28, pp. 153–172, 2019.
- [11] L. Fialho, R. Melicio, V. Mendes, J. Figueiredo, and M. Collares-Pereira, "Effect of shading on series solar modules: Simulation and experimental results," *Procedia Technology*, vol. 17, pp. 295–302, 2014. Conference on Electronics, Telecommunications and Computers – CETC 2013.

- [12] A. López-Vargas, M. Fuentes, M. V. García, and F. J. Muñoz-Rodríguez, "Low-Cost Datalogger Intended for Remote Monitoring of Solar Photovoltaic Standalone Systems Based on Arduino™," *IEEE Sensors Journal*, vol. 19, no. 11, pp. 4308–4320, 2019.
- [13] P. Ayala, C. Muñoz, N. Osorio, C. Hernández, F. Zurita, V. Gutierrez, G. Ramirez, F. Mancilla, P. Valdivia, F. Cuevas, and P. Ferrada, "Bifacial Technology Performance Compared With Three Commercial Monofacial PV Technologies under Outdoor High Irradiance Conditions at the Atacama Desert," in *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC)*, pp. 0672–0675, 2018.
- [14] K. Achuthan, J. D. Freeman, P. Nedungadi, U. Mohankumar, A. Varghese, A. M. Vasanthakumari, S. P. Francis, and V. K. Kolil, "Remote Triggered Dual-Axis Solar Irradiance Measurement System," *IEEE Transactions on Industry Applications*, vol. 56, no. 2, pp. 1742–1751, 2020.
- [15] M. Dhaoui and H. Najjari, "Remote control and monitoring of the Agureb-Jemna Kébili photovoltaic pumping station," in *2017 International Conference on Green Energy Conversion Systems (GECS)*, pp. 1–6, 2017.
- [16] M. Sikimić, M. Amović, V. Vujović, B. Suknović, and D. Manjak, "An Overview of Wireless Technologies for IoT Network," in *2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–6, 2020.
- [17] DIGI, "Explore the Digi XBee Ecosystem 2021," 2021.
- [18] E. Kabalci and Y. Kabalci, "Remote monitoring system design for photovoltaic panels," in *2017 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pp. 888–891, 2017.
- [19] A. Avila, P. R. Vizcaya, and R. Diez, "Daily irradiance test signal for photovoltaic systems by selection from long-term data," *Renewable Energy*, vol. 131, pp. 755–762, 2019.
- [20] A. Prastianto, M. Fadhill, A. Rahardjo, F. H. Jufri, and F. Husnayain, "Design of Solar Irradiance Measurement Based On Analytical Data Using Microcontroller," in *2019 IEEE Conference on Energy Conversion (CENCON)*, pp. 137–141, 2019.
- [21] M. Rana, I. Koprinska, and V. G. Agelidis, "Univariate and multivariate methods for very short-term solar photovoltaic power forecasting," *Energy Conversion and Management*, vol. 121, pp. 380–390, 2016.
- [22] X.-m. Gao, S.-f. Yang, and S.-b. Pan, "Optimal parameter selection for support vector machine based on artificial bee colony algorithm: A case study of grid-connected pv system power prediction," *Computational Intelligence and Neuroscience*, vol. 2017, pp. 1–14, 08 2017.
- [23] M. Colak, M. Yesilbudak, and R. Bayindir, "Very Short-Term Estimation of Global Horizontal Irradiance Using Data Mining Methods," in *7th International IEEE Conference on Renewable Energy Research and Applications, ICRERA 2018*, vol. 7, (Paris), pp. 1472–1476, IEEE, 2018.
- [24] T. Stoffel and A. Andreas, "http://www.osti.gov/servlets/purl/1052547/, 6 de Abril de 2021."
- [25] B. Zhou, W. Zhao, X. Su, S. Lu, T. Wang, W. Yao, P. Xie, T. Mao, L. Guan, and Y. Lv, "PV Power Characteristic Modeling based on Multi-scale Clustering and Its Application in Generation Prediction," *IEEE Power and Energy Society General Meeting*, vol. 2018-August, pp. 1–5, 2018.
- [26] X. Wu, V. Kumar, Q. J. Ross, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, *Top 10 algorithms in data mining*. Springer, 2008.
- [27] M. Alfonso Galipienso, M. Cazorla Quevedo, O. Colomina Pardo, F. Escolano Ruiz, and M. Lozano Ortega, *Inteligencia artificial: Modelos, Técnicas y Áreas de Aplicación*. Ediciones Paraninfo S.A., 2003.
- [28] D. Rekioua and E. Matagne, *Optimization of photovoltaic power systems: Modelization, Simulation and Control*, vol. 102. London: Springer, 2012.
- [29] L. F. L. Villa, D. Picault, B. Raison, S. Bacha, and A. Labonne, "Maximizing the power output of partially shaded photovoltaic plants through optimization of the interconnections among its modules," *IEEE Journal of Photovoltaics*, vol. 2, no. 2, pp. 154–163, 2012.



Aracely Zapién-Castillo received the B.S. degree in industrial engineering from Instituto de Estudios Superiores de Tamaulipas in 2019. She is currently studying the M.S. degree in engineering science at Instituto Tecnológico de Cd. Madero. Her research interests include data analysis and modeling of photovoltaic systems.



José Angel Zumaya-García received the A.S. degree in informatic from Centro de Bachillerato Tecnológico Industrial y de Servicios 103 and the B.S. degree in mechatronics from Instituto de Estudios Superiores de Tamaulipas in 2014 and 2018, respectively. He is currently studying the M.S. degree in engineering science at the Instituto Tecnológico de Cd. Madero. His research interests include monitoring, analysis, optimization and modeling of photovoltaic systems.

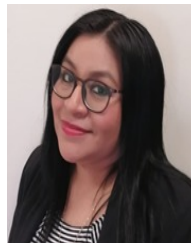


utility scale power systems.

Pedro M. Garcia-Vite received the B.S. degree in electronics and the M.S. degree in electrical engineering from Instituto Tecnológico de Cd. Madero in 2004 and 2006, respectively, and his Ph.D. degree in electrical engineering from the Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Guadalajara, México, in 2012. He is currently with the Tecnológico Nacional de México-Instituto Tecnológico de Cd. Madero. His research interests include modeling and control of power electronics converters and their application to



Luciano Aguilera-Vázquez is originally from Veracruz, Mexico and currently works as a full-time professor at the Instituto Tecnológico de Cd. Madero, Tamaulipas, Mexico. He holds a Ph. D. in Engineering Sciences from University of Stuttgart, in Baden-Württemberg, Germany, where he developed research at the institute for biochemical engineering processes. He has several publications in international conferences. His research interests include energy metabolism, microalgae biotechnology, data analysis, and mathematical programming.



analysis.

Francisca Hernández-Angel received the B.S. degree in industrial engineering from the Instituto Tecnológico de Cd. Madero Tamaulipas, México, and the M.S. degree in administration science in the field of industrial relationships, currently she is a Ph. D. student in strategic business management, at the School of Business Administration in the Universidad Autónoma de Tamaulipas, UAT, she is full-time professor at the Universidad Politécnica de Altamira. Her research interests include: innovation and entrepreneurship, as well as statistical data