ASADES

Avances en Energías Renovables y Medio Ambiente
Vol. 22, pp 03.25-03.33, 2018. Impreso en la Argentina
ISSN 2314-1433 - Trabajo selecionado de Actas ASADES2018

PREDICCIÓN DE GENERACIÓN FOTOVOLTAICA CON TÉCNICAS DE APRENDIZAJE SUPERVISADO

Walter Herrera Martinez[1], Analía Moreno[2], Ana Clara Reybet[1], Simón Saint-André[2]
[1]Consejo Nacional de Investigaciones Científicas y Técnicas
[2]Agencia Nacional de Investigaciones Científicas y Tecnológicas
Instituto de Nanociencia y Nanotecnología (INN), CNEA – CONICET
Departamento Energía Solar, Gerencia Investigación y Aplicaciones,
Centro Atómico Constituyentes, CNEA
Av. General Paz 1499 (1650), San Martín, Buenos Aires, Argentina. Tel. (011) 6772-7199
e-mail: herreram@tandar.cnea.gov.ar

Recibido 13/08/18, aceptado 21/09/18

RESUMEN: En el siguiente trabajo se realizó la predicción de la potencia generada por un conjunto de módulos fotovoltaicos ubicado en el edificio 42 del Centro Atómico Constituyentes (CAC) de la Comisión Nacional de Energía Atómica (CNEA). Se efectuó un preprocesamiento de tres años de datos de generación recolectados del inversor fotovoltaico, y posteriormente se determinó a partir de diferentes métodos de aprendizaje supervisado y su análisis, que el método con el algoritmo de *Random Forest* presentó el comportamiento más adecuado para realizar una predicción respecto a los datos de generación fotovoltaica reales. Una vez elegido el método de aprendizaje, se optimizaron los parámetros y se analizaron qué variables características son las más influyentes en los resultados. Finalmente se obtuvo la energía producida por el conjunto de módulos a partir de la generación predicha y se comparó con los datos reales, obteniendo coeficientes de determinación mayores a 0,9.

Palabras clave: Energía solar, generación, aprendizaje supervisado, *Random Forest*.

INTRODUCCION

En las centrales generadoras de energía convencional se utilizan recursos como carbón, gas natural, petróleo, entre otras, que se encuentran asociados con la emisión de gases de efecto invernadero, en especial CO_2 (Yeboah et al., 2015). En los últimos años el cambio climático y la crisis energética han motivado al uso racional de la energía, como también al desarrollo, investigación e implementación de energías renovables. Debido al abundante recurso solar existente en nuestro país, la energía solar fotovoltaica es una de las más prometedoras para la generación energética.

Uno de los principales desafíos a resolver es la variabilidad que se presenta en la generación fotovoltaica. Las fuentes energéticas convencionales, a excepción de cuando presentan fallas técnicas, poseen una generación eléctrica que se puede predecir fácilmente. Sin embargo, la energía solar fotovoltaica, como la mayoría de las energías renovables, presenta una amplia variación con las condiciones meteorológicas, lo que dificulta su predicción. La predicción de la producción renovable puede ser útil para estimar las reservas, programar el sistema de energía, gestionar la congestión en las redes de transmisión, coordinar el almacenamiento o comerciar en los mercados eléctricos. En las últimas décadas se ha investigado y desarrollado modelos para predecir la radiación solar, pero existen pocos desarrollos para evaluar la potencia generada por sistemas fotovoltaicos. Para asegurar una incorporación rentable y económica de energía solar fotovoltaica, la predicción de potencia de sistemas fotovoltaicos se ha convertido en un elemento crítico de la gestión de los sistemas energéticos (Bella Espinar et al, 2010). La motivación para realizar este trabajo fue predecir la potencia generada de una instalación a partir de variables meteorológicas, utilizando técnicas de aprendizaje automático.

Las técnicas de aprendizaje automático son adecuadas para predecir sistemas complejos, especialmente cuando se utilizan grandes cantidades de datos. Los métodos detectan automáticamente patrones en los datos y luego usan los patrones encontrados para predecir los datos futuros. Estas técnicas pueden ser categorizadas en dos grupos: técnicas supervisadas y técnicas no supervisadas. Las técnicas de aprendizaje automático supervisado (*Supervised Machine Learning*: *SML*) son utilizadas en los modelos predictivos donde los atributos o variables de entradas tienen asociadas variables de salida o respuesta. Si la variable de salida o respuesta es categórica (por ejemplo: si/no, hombre/mujer, etc.) entonces es un problema de clasificación. Cuando la variable de salida es continua entonces se trata de un problema de regresión. Por otro lado, las técnicas de aprendizaje automático no supervisado (*UnSupervised Machine Learning: USML*) son utilizadas para información o descubrimiento de conocimiento (Yeboah et al., 2015).

El foco del siguiente estudio se basa en las técnicas *SML*. Existen varios métodos para resolver problemas de *SML*, como redes neuronales (*Artificial Neural Networks: ANN*), máquinas de vectores de soporte (*Support Vector Machines: SVM*) y métodos de ensamble (*Ensemble methods: ME*) que incluye el método *Random Forest* (*RF*) tanto la clasificación como la regresión. En este caso, la variable de salida, la potencia generada, es continua y por lo tanto, se trata de un problema de regresión. Para realizar las predicciones necesarias se utilizó el método *RF*, debido a que los métodos tradicionales como la regresión lineal no son lo suficientemente robustos para manipular la complejidad y la no linealidad de los valores predichos involucrados en el modelado de la potencia solar (Yeboah et al., 2015).

RF es un algoritmo que se basa en la formación de bosques como conjuntos de árboles de decisión. Los mismos son generados, en la mayoría de los casos, con el método de "*bagging*". Este método combina varios modelos de aprendizaje para encontrar el óptimo. En conclusión, *RF* construye múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable. Una de las ventajas que tiene utilizar este algoritmo es que agrega aleatoriedad adicional al modelo, mientras crecen los árboles. En lugar de buscar la característica más importante mientras se divide un nodo, busca la mejor característica entre un subconjunto aleatorio de características. Esto da como resultado una amplia diversidad que generalmente devuelve un mejor modelo (Breiman, 2001).

RF se considera un algoritmo muy útil y fácil de usar, porque los hiperparámetros predeterminados a menudo producen un buen resultado de predicción. Uno de los grandes problemas en el aprendizaje supervisado es el sobreajuste, lo que no sucede tan fácilmente con el algoritmo *RF*, debido a que, si hay suficientes árboles en el bosque, el clasificador no sobreajustará el modelo. La principal limitación de *RF* es que una gran cantidad de árboles puede hacer que el algoritmo sea lento e ineficaz para las predicciones en tiempo real. En general, estos algoritmos son rápidos de entrenar, pero bastante lentos para crear predicciones una vez que están entrenados (Svetnik, 2003).

El objetivo del trabajo es poder predecir la potencia generada por una instalación fotovoltaica a partir de variables meteorológicas como temperatura, humedad, radiación y nubosidad, por un método de aprendizaje supervisado y demostrar que el algoritmo de *RF* es aquel que mejor ajusta el modelo.

DATOS UTILIZADOS

La instalación fotovoltaica donde se registraron los datos de generación utilizados para este trabajo está compuesta por 23 paneles solares de silicio policristalino y Tedlar transparente, marca Brandoni de 215 Wp cada uno, inclinados 34° con respecto a la horizontal y con orientación 20° al Este del Norte (aproximadamente NNE). Estos paneles están instalados en una pérgola en el tercer piso del edificio 42 del CAC de CNEA (Eyras, Durán, 2013). El inversor conectado a dicha instalación es de 4,6 kW, marca AEG modelos Protect PV 4600. El mismo cuenta con una interfaz de comunicación serie RS232 y el fabricante proporciona el

software propietario PV MONITOR para monitorear el estado de múltiples inversores, además permite definir el intervalo de muestreo y exportar los datos adquiridos.

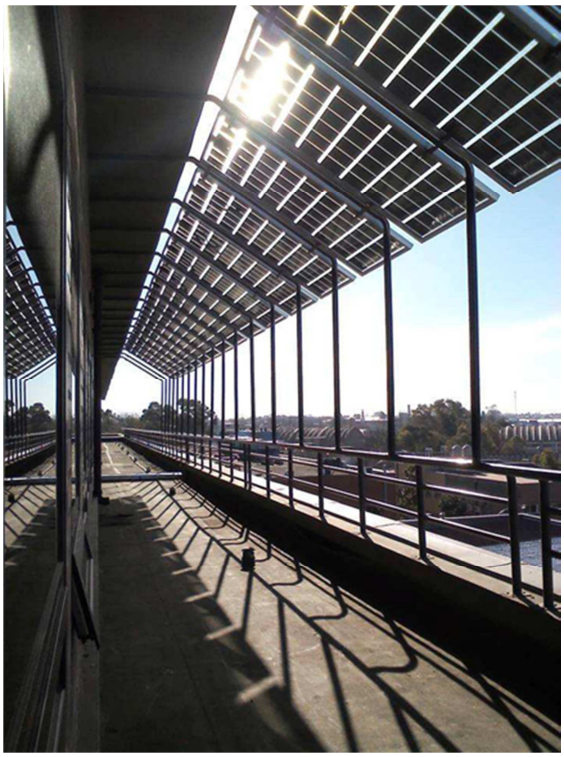


Figura 1: Instalación fotovoltaica.

Figura 2: Inversor AEG.

Las mediciones de radiación solar fueron obtenidas de un piranómetro CMP21 calidad Level 1.0 de la red AERONET (*Aerosol Robotic Network, NASA*) instrumento instalado en el CEILAP (CITEFA - CONICET), Villa Martelli, Buenos Aires. Los datos de temperatura, humedad relativa y nubosidad total para los tres años estudiados fueron proporcionados por el Servicio Meteorológico Nacional (SMN) de la estación Villa Ortúzar, Ciudad Autónoma de Buenos Aires.

METODOLOGIA

Para el análisis, limpieza, procesamiento y transformación de los datos en este trabajo se ha elegido como lenguaje de programación de código abierto Python v.3.6.4. Se aprovecharon las facilidades que brinda la librería *sklearn (http://scikit-learn.org/stable/index.html)* para aplicar métodos de aprendizaje supervisado.

Inicialmente se realizó un preprocesamiento de los datos, debido a que estos no fueron proporcionados por las mismas fuentes, se igualaron intervalos de muestreo en las mediciones y se realizaron las conversiones de unidades correspondientes. Para realizar la limpieza de datos se tuvo en cuenta que el inversor entra en modo de espera cuando la generación está por debajo de un límite (100 VDC), también se consideraron fallas técnicas informadas que produjeron la desconexión de los equipos, tanto en el inversor como en el piranómetro. Inicialmente el criterio que se adoptó para descartar un día completo fue tener 8 horas sin generación.

Para realizar el entrenamiento de los datos se utilizaron dos métodos: Regresión lineal y *Random Forest*, un importante aspecto a tener en cuenta es que los árboles permiten mejores predicciones con variables que se encuentran poco correlacionadas entre sí. En los modelos se

utilizan hiperparámetros y parámetros, los primeros son una configuración externa al modelo y su valor no puede ser estimado de los datos; en contraste los parámetros son internos al modelo y su valor se obtiene de los datos procesados. Se realizaron ambos métodos con el objetivo de comparar y demostrar cómo aumenta el coeficiente de determinación utilizando un algoritmo más complejo y confiable como es *RF*. El coeficiente de determinación indica la proporción de la variación total que está siendo explicada por la regresión, además ofrece una idea de la calidad del ajuste del modelo a los datos. Los hiperparámetros en *RF* se usan para aumentar el poder predictivo del modelo o para hacer que el modelo sea más rápido. Dentro de los más importantes está el hiperparámetro "*n_estimators*", que es el número de árboles que construye el algoritmo, en general una mayor cantidad de árboles aumenta el rendimiento y hace que las predicciones sean más estables, pero también ralentiza el cálculo. Otro hiperparámetro importante es "*max_depth*", que se refiere a la máxima profundidad del árbol, la variación del mismo permitiría optimizar el modelo.

Inicialmente se analizó la variación de las predicciones mediante regresión lineal y RF, utilizando como característica de entrenamiento sólo la variable radiación global. Luego se agregaron a los datos de entrenamiento las variables meteorológicas y temporales como humedad relativa, nubosidad total, temperatura, mes del año y hora del día, con el fin de estimar su influencia en la generación de energía.

Una vez obtenido el modelo óptimo para predecir los datos de potencia generada se procedió al cálculo de la energía producida utilizando la ecuación 1:

$$E(t) = \int_a^b P(t)dt \tag{1}$$

Para comparar con los datos proporcionados por el inversor se establecen los límites de integración para cada día. Se resolvió la integral empleando el método de los trapecios que aproxima la ecuación (1) con la siguiente ecuación 2 (Larson, 2006):

$$\int_a^b P(t)dt \sim \frac{b-a}{n}\left[\frac{P(a)-P(b)}{2} + \sum_{k=1}^{n-2} a + k\frac{b-a}{n}\right] \tag{2}$$

RESULTADOS Y DISCUSIÓN

Se llevaron a cabo las predicciones de la potencia generada con las dos técnicas de regresión antes descriptas (Regresión lineal y *Random Forest*) modificando las variables de entrenamiento, los coeficientes de determinación (R^2) que se obtuvieron en cada uno de los modelos simulados se muestran en la tabla 1:

	Radiación global	Radiación global + Variables características
Regresión lineal	0,694	0,764
Random Forest	0,557	0,869

Tabla 1: Métodos de predicción utilizados y valores de R^2 para cada uno de ellos.

A partir de estos resultados se eligió utilizar como método de predicción *RF* y considerar todas las variables de entrenamiento: radiación global total, variables meteorológicas y temporales, ya que presenta el máximo R^2 de los casos analizados.

Como se mencionó anteriormente, en el algoritmo de *RF* existen parámetros variables para maximizar la eficiencia del método. En la Figura 3 se observa el aumento de R^2 al variar la cantidad de árboles (*n_estimators*) y la profundidad de los árboles *(max_depth*). El resultado de este estudio permite conocer cuáles son los hiperparámetros óptimos para realizar la predicción con los datos utilizados para este trabajo.

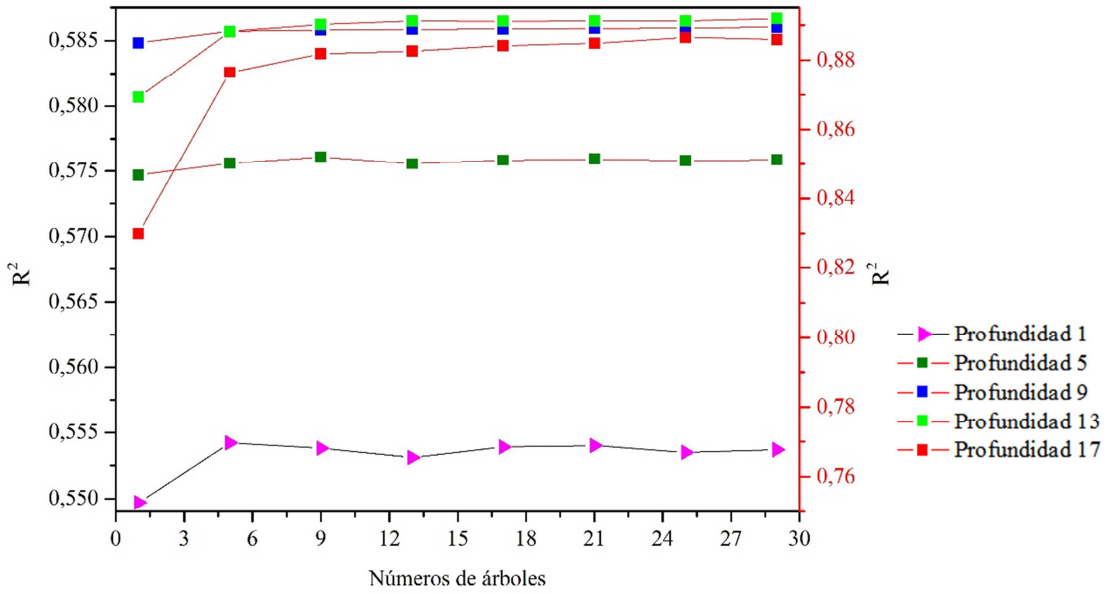


Figura 3: Optimización de hiperparámetros para el método RF y su correspondiente valor de R^2. El eje de la izquierda solo corresponde a los datos de profundidad 1.

	Profundidad 1	Profundidad 5	Profundidad 9	Profundidad 13	Profundidad 17
Nº de árboles 1	0,5497	0,8468	0,8850	0,8693	0,8300
Nº de árboles 15	0,5542	0,8502	0,8882	0,8882	0,8765
Nº de árboles 19	0,5538	0,8520	0,8886	0,8902	0,8818
Nº de árboles 13	0,5531	0,8501	0,8888	0,8912	0,8826
Nº de árboles 17	0,5539	0,8511	0,8889	0,8911	0,8842
Nº de árboles 21	0,5540	0,8514	0,8890	0,8912	0,8848
Nº de árboles 25	0,5535	0,8510	0,8891	0,8912	0,8865

Tabla 2: Optimización de hiperparámetros para el método RF y su correspondiente valor de R^2.

Aumentar el número de árboles y su profundidad es sinónimo de más espacio computacional y tiempo de procesamiento, por esto es importante encontrar los valores óptimos de los hiperparámetros de modo que se maximice el valor de R^2. A partir de la Tabla 2 se puede inferir que el ajuste más favorable fue aquel cuyos parámetros eran 13 árboles y una profundidad de 13, lo que proporcionó un R^2 de 0,8912, demostrando una alta calidad del modelo de ajuste.

Un aspecto importante en la evaluación del método escogido fue determinar la influencia de cada variable en el modelo (Figura 4), se debe tener en cuenta que la influencia de las variables

depende del algoritmo utilizado y se refiere a cuáles variables tuvieron mayor influencia para la predicción obtenida en base a los datos usados en este trabajo.

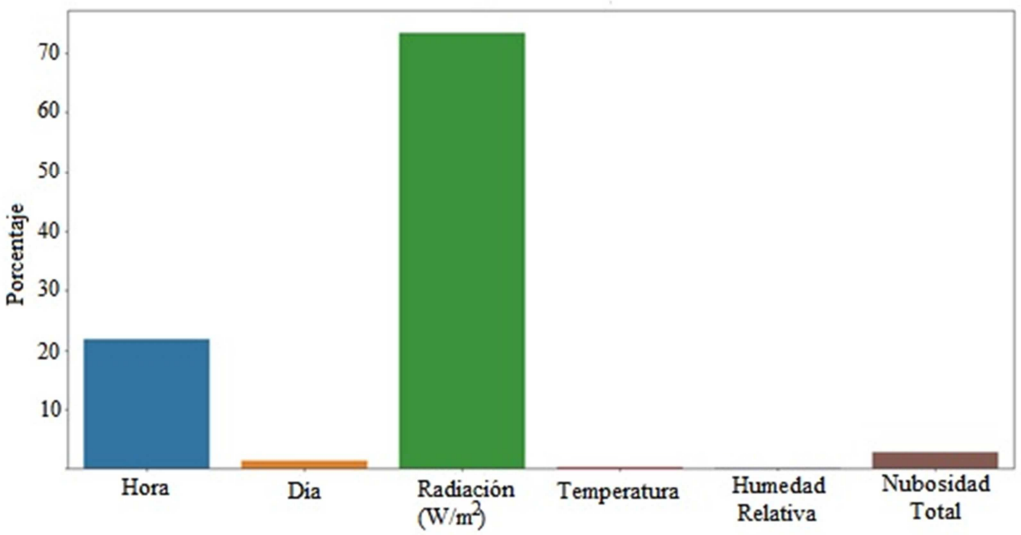


Figura 4: Influencia de cada variable en el modelo empleado.

En la Figura 4 se observa que la variable radiación solar global es la que más influye en la predicción de la generación, algo que era esperable. Sin embargo, la temperatura y humedad relativa son dos variables casi sin efecto en esta predicción, este es un punto que se podría estudiar y analizar en trabajos futuros donde se comparen otros métodos de aprendizaje.

En las Figuras 5 y 6 se presentan resultados de la predicción contrastados con los datos originales en un día soleado y un día con gran nubosidad, respectivamente. Estas gráficas ilustran la capacidad que presenta el modelo para predecir la potencia generada con las variables de entrenamiento consideradas.

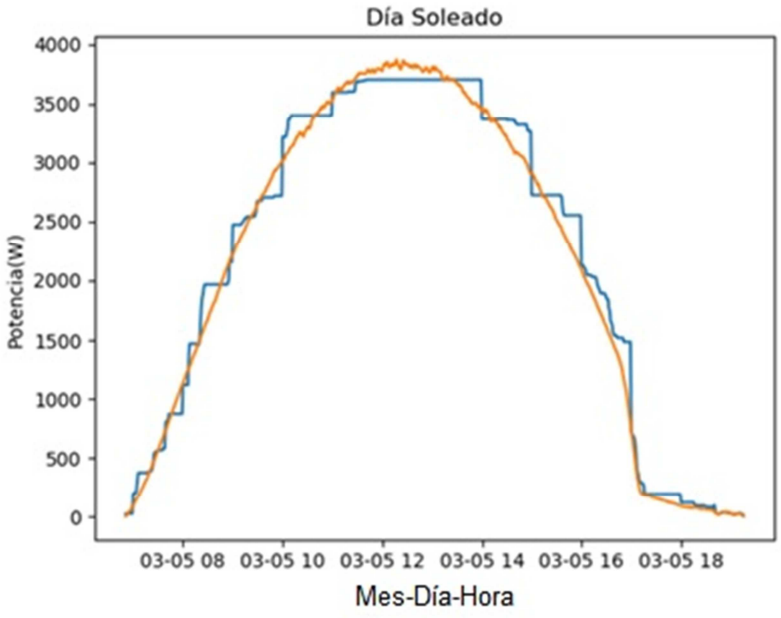


Figura 5. Potencia generada y predicción para un día soleado.

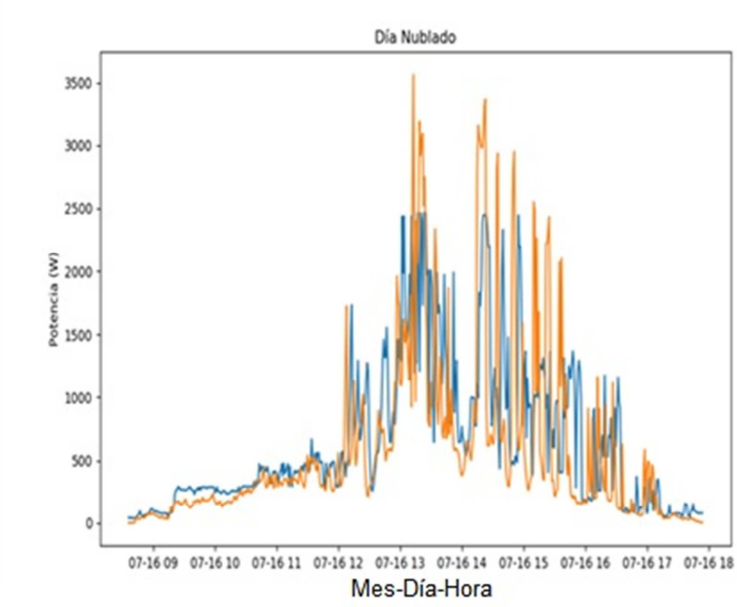
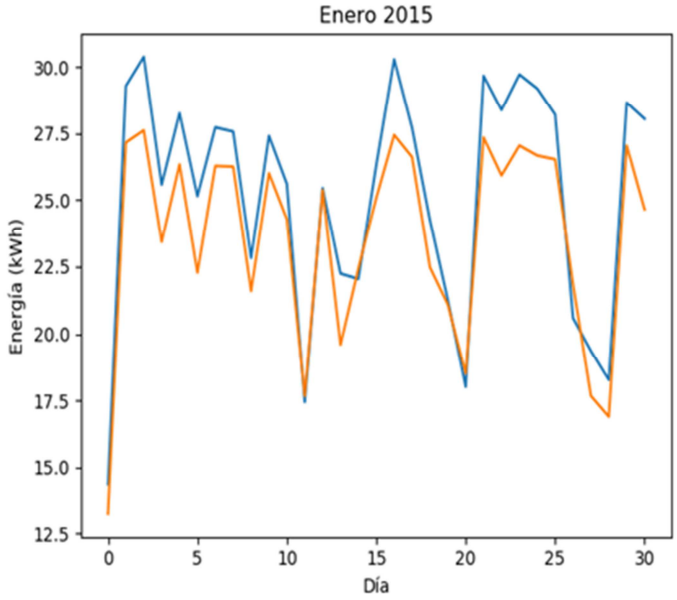


Figura 6. Potencia generada y predicción para día nublado.

Las figuras anteriores, muestran cómo el modelo predicho tiene un comportamiento similar al presentado por los datos reales, las mayores diferencias se encuentran en los mínimos de potencia generada que corresponden a momentos nubosos del día, esta diferencia se relaciona con la dificultad que presenta el predecir la nubosidad instantánea.

Por último, se procedió al cálculo de la energía generada con los datos predichos y se comparó con los datos energéticos que entrega el inversor. Para ello se realizó la integral numérica diaria utilizando el método del trapecio como se describió anteriormente; se consideró únicamente los datos que se encontraban equiespaciados y los días que presentaban más del 5% de los valores totales diarios. A modo ilustrativo se presenta la gráfica del mes enero de 2015 (Figura 7). En adición, en la Figura 8 se presenta la correlación existente entre la energía calculada y la energía que es medida por el inversor, en la cual se observó una correlación entre los valores calculados a partir de los datos predichos y los datos reales. Esta correlación presentó un coeficiente de



determinación de 0,935.

Figura 7. Energía medida por el inversor (curva azul) y calculada (curva naranja) para el mes de enero de 2015.

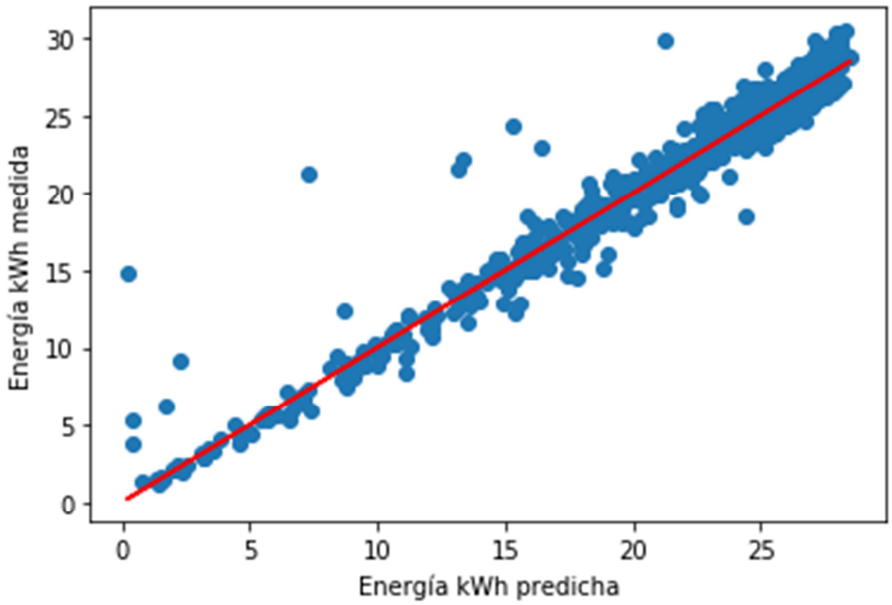


Figura 8. Correlación entre la energía medida por el inversor y la energía calculada.

CONCLUSIONES

La generación fotovoltaica depende de factores complejos para su predicción, por esa razón los métodos tradicionales de modelado, como regresión lineal, no conducen a resultados apropiados. Otros métodos, como *RF*, demuestran ser más adecuados para modelar sistemas con dichas características.

La predicción de potencia generada se realizó utilizando *RF* con todas las variables de entrenamiento meteorológicas y temporales obteniendo un coeficiente de determinación (R^2) de 0,892 con los valores de los hiperparámetros optimizados. Se encontró que al realizar la predicción para días muy nublados, esta no era tan buena, debido a la falta de datos instantáneos de nubosidad.

Se puede concluir, además, como era de esperarse, la potencia generada se ve en gran medida relacionada con la radiación global y en menor medida con las demás variables que se utilizaron para el aprendizaje. Esto conduce a pensar que se podrían estudiar casos futuros de predicción tanto de generación sólo con variables meteorológicas como humedad, nubosidad y temperatura.

La comparación entre la energía generada real vs la energía calculada mostró un comportamiento lineal, con un R^2 de 0,935. Esto indicara que la predicción de la potencia nos permitió obtener valores de energía con gran precisión. Se pretende en el futuro explorar otras técnicas más complejas como redes neuronales o máquinas de vectores de soporte para realizar predicciones de generación fotovoltaica.

Finalmente, en base a este trabajo se podría predecir aproximadamente las pérdidas energéticas debido a fallas de desconexión del inversor.

REFERENCIAS

Bella Espinar, José-Luis Aznarte, Robin Girard, Alfred Mbairadjim Moussa, Georges Kariniotakis. (2010). Photovoltaic Forecasting: A state of the art. 5th European PV-Hybrid and Mini-Grid Conference, (pp. 250-255). Tarragona, Spain.

Breiman, L. (2001). Random forests. Machine learning, (pp. 5-32).

Eyras, R. D. (2013). Proyecto Iresud:" Interconexión De Sistemas Fotovoltaicos A La Red Eléctrica En Ambientes Urbanos. In Primer Encuentro Lationamericano de Uso Racional y Eficiente de la Energía.

Larson, R. H. (2006). Cálculo I. Octava edición. Ed. McGraw-Hill.

Svetnik, V. L. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences, (p. 1).

Yeboah, Frank & Pyle, Robert & Bock Hyeng, Christian. (2015). Predicting Solar Radiation for Renewable Energy Technologies - A Random Forest Approach. International Journal of Modern Engineering, 100-107.

ABSTRACT

In this work was carried out the power prediction generated by a set of photovoltaic modules located in 42´s Constituyentes Atomic Center (CAC) building of the National Atomic Energy Commission (CNEA). A preprocessing of three years of photovoltaic generation data, collected from the inverter to which the modules are connected, was conducted. Later it was determined that the method with the Random Forest algorithm presented the most suitable behavior to make a prediction regarding the real photovoltaic generation data. After choosing the learning method, parameters were optimized and analyzed to determine which characteristic variables are the most influential in the results. Finally, the energy produced by the modules was obtained from the predicted generation and it was compared with the real data, obtaining coefficients of determination greater than 0.9.

Keywords: Solar energy, generation, machine learning, Random Forest.