Prediction of Generation in a Photovoltaic System through the application of Data Mining techniques

Predicción de la Generación para un Sistema Fotovoltaico mediante la aplicación de técnicas de Minería de Datos

C.P. Fabara[1] D.A. Maldonado[1] M.S. Soria[1] A.F. Tovar[1]

[1]*Escuela Politécnica Nacional, Facultad de Ingeniería Eléctrica y Electrónica, Quito, Ecuador*
E-mail: cristian.fabara@epn.edu.ec; diego.maldonado@epn.edu.ec; mauricio.soria@epn.edu.ec; antonio.tovar@epn.edu.ec;

Abstract

This document presents a generation prediction model through data mining techniques for a photovoltaic plant located at Paragachi Community, belonging to Pimampiro (Imbabura), with a total of 14400 solar panels and 3.6 MW nominal power. This system does not have a battery bank for storage, for this reason, it does not provide energy at night, but during the day, it supplies the energy to 2000 households that represent Pimampiro's urban population.

It begins with a univariate and multivariate analysis of the measurement variables, whose objective is to determine the behavior, incidence and the relationship of each variable in the generation of the photovoltaic system. With the variables of higher incidence as input, a learning machine is trained; it uses the technique of decision trees through random forest to predict the generation.

In renewable energies, the photovoltaic system is one of the most implemented and developed nowadays. However, predicting the amount of power it can generate is complicated by the stochastic behavior of the variables, limiting the entry of this technology into a competitive market, which can integrate into the National Interconnected System in an optimal and efficient way.

***Index terms*— Photovoltaic systems, generation prediction, data mining, machine learning, decision trees.**

Resumen

Este documento presenta un modelo de predicción de generación de energía mediante técnicas de minería de datos para la central fotovoltaica ubicada en la Comunidad Paragachi, perteneciente al cantón Pimampiro (Imbabura), con un total de 14400 paneles solares y potencia nominal de 3.6 MW. Este sistema no cuenta con banco de baterías para almacenamiento, debido a esto no aporta durante las noches, pero en el día abastece a más de 2000 familias, que representa toda la población urbana de Pimampiro.

Se inicia con un análisis univariante y multivariante de las variables de medición, cuyo objetivo es determinar el comportamiento, incidencia y la relación de cada variable en la generación de energía de la central. Con las variables de mayor incidencia como entrada, se entrena una máquina de aprendizaje que usa la técnica de árboles de decisión mediante bosques aleatorios (Random Forest) para predecir la generación de energía.

En energías renovables, el sistema fotovoltaico es uno de los más implementados y desarrollados en la actualidad. Sin embargo, predecir la cantidad de generación que puede proveer es complicado por el comportamiento estocástico de las variables, limitando el ingreso de esta tecnología a un mercado competitivo que se integre al Sistema Nacional Interconectado de forma óptima y eficiente.

***Palabras clave*— Centrales fotovoltaicas, predicción de generación, minería de datos, máquina de aprendizaje, árboles de decisión.**

Recibido: 28-05-2019, Aprobado tras revisión: 19-07-2019
Forma sugerida de citación: Fabara, C.; Maldonado, D.; Soria, M.; Tovar, A.; (2019). "Predicción de la Generación para un Sistema Fotovoltaico mediante la aplicación de técnicas de Minería de Datos". Revista Técnica ***"energía".*** No. 16, Issue I, Pp. 64-72
ISSN On-line: 2602-8492 - ISSN Impreso: 1390-5074

1. INTRODUCCIÓN

En la actualidad los combustibles fósiles son fundamentales para asegurar el abastecimiento energético, pero su uso implica elevados impactos ambientales y su ritmo de explotación es tan elevado que se prevé que en un futuro próximo se agote, provocando una crisis energética. El aprovechamiento de energías renovables cobra cada vez más importancia por su característica de recurso ilimitado y sus ventajas ambientales que presenta.

Con base al informe de La Agencia de Regulación y Control de Electricidad emitido el año 2017, la oferta de energía eléctrica determinada por su potencia efectiva proviene de centrales hidroeléctricas en un 60.34% y las centrales fotovoltaicas y eólicas tienen una participación del 0.34% y 0.28%, respectivamente. Las centrales fotovoltaicas reportan 22 concesiones privadas para generar 33.3 GWh de energía bruta al Sistema Nacional Interconectado (SNI) [1].

El Ecuador por encontrarse ubicado en la mitad del mundo, tiene altos niveles de radiación solar, de acuerdo con el mapa de insolación global el Ecuador cuenta con un valor promedio de 4.57 kWh/m^2/día para el aprovechamiento de generación fotovoltaica, este valor se encuentra por encima del promedio a nivel mundial que se estima en 4.2 kWh/m^2/día [2]. Por otra parte, el IIGE (Instituto de Investigación Geológico y Energético) está evaluando el recurso solar y sus investigaciones se enfocan en proyectos como: estimación de potencial de energía renovable mediante la instalación de estaciones meteorológicas, atlas de recurso solar y eólico, modelación de sistemas híbridos con cogeneración para aplicaciones industriales.

Estimar la energía eléctrica de centrales fotovoltaicas es una tarea compleja, por el comportamiento estocástico de las variables. Sin embargo, este estudio es importante para que este tipo de tecnología sea cada vez más competitiva en el mercado eléctrico y se pueda integrar al SNI de manera más eficaz y óptima. La predicción puede ser de corto plazo para prever la cantidad de energía en un tiempo de horas a pocos días, y de largo plazo para planificación y expansión.

Los primeros métodos de predicción utilizan variables atmosféricas para predecir la radiación solar y de forma indirecta la generación de energía mediante la metodología Box-Jenkins (modelo ARIMA), modelo que se basa en el análisis de las propiedades probabilísticas o estocásticas de las series de tiempo económicas, donde una variable Yt puede ser expresada en función de sus valores pasados, donde no existe relación causal alguna a diferencia de los modelos clásicos de regresión, siendo útil para la predicción referente a generación de energía solar fotovoltaica, por su comportamiento [3].

En [4] se presenta un modelo ARIMA que predice la radiación solar media diaria considerando como datos la radiación solar media mensual de siete años en diferentes localidades, se consigue un error absoluto medio de 16% aproximadamente.

En [5] se determina la radiación media horizontal y de plano inclinado con horizonte de predicción de 10, 20, 30 min y 1h. Se utilizan las mediciones reales de radiación horizontal y de plano inclinado de 4 períodos de 15 días en dos inviernos y dos veranos diferentes de la región. El aporte de esta investigación radica en demostrar que la eficiencia de cada modelo crece cuando es más pequeño el horizonte de predicción.

En [6] se utiliza la radiación horizontal global medida localmente durante año y medio para predecir la potencia media diaria generada por una central fotovoltaica de 2.1 kW compuesto por paneles fijos. Este artículo es de gran importancia para determinar la relación existente entre la radiación y la potencia eléctrica y presenta buenos resultados por ser un horizonte de predicción corto.

Modelos más avanzados utilizan técnicas de inteligencia artificial y variables de entrada tanto eléctricas como meteorológicas para mejorar la predicción. En [7] se utiliza un modelo gris y un modelo de red neuronal para predecir la potencia eléctrica media horaria de las próximas 12 horas (de 7 am – 7 pm), los resultados muestran que la combinación de ambas técnicas aumenta la precisión del modelo.

En [8] se utiliza un modelo de red neuronal para predecir la radiación media horaria de las próximas 24 horas, normalizando las variables de entrada y calculando el parámetro $TOD_{máx}$ para encontrar variaciones bruscas de radiación y NDD que clasifica los días en nublados o claros. Las variables de entrada son la radiación solar y la temperatura ambiente media diaria del día previo a la predicción.

En [9] se presenta una metodología basada en máquinas de soporte vectorial (SVMs) combinado con el modelo del vecino más próximo (k-NN) en la selección de parte de las variables de entrada. Es un modelo de predicción de la potencia media horaria generada de las próximas 24 horas en una central fotovoltaica de 20 MW formada por paneles fijos. Se obtiene un MAE del 8.64% y un RMSE próximo al 11%.

En [10] se propone un modelo de predicción de potencia media generada en tiempo real, utilizando como entradas la temperatura ambiente y radiación horizontal medidas localmente. Se desarrollan dos redes neuronales diferentes que se aplican según el día (nublado o soleado) obteniendo factores de correlación de 0.96 al 0.97.

En [11] se desarrolla un modelo de predicción a corto plazo con redes neuronales y se lo optimiza mediante algoritmos genéticos, ya sea por medio de nube de partículas u enjambre de partículas, donde el lado del PSO permite guardar información del mejor individuo, mientras que las mutaciones permiten mantener la diversidad de la población de la PSO.

En [12] se predice la potencia eléctrica media horaria para el horizonte de unos pocos días con datos de cada 10 minutos de una planta fotovoltaica de 20 kW, con variables de irradiancia, temperatura del módulo, temperatura ambiente, tensión, intensidad y potencia. Se desarrollan tres modelos, uno basado en la técnica SARIMA, otro en SVM y un tercero híbrido combinando ambos modelos, donde el modelo híbrido muestra un mejor comportamiento.

En [13] se predice la potencia eléctrica media horaria de la próxima hora utilizando históricos de potencia, temperatura máxima y mínima diaria, velocidad del viento media diaria, presión atmosférica media del aire y el tipo de día (soleado, nublado, cubierto). El modelo utiliza el algoritmo Fuzzy C-Means para clasificar a los datos históricos en seis tipos y a cada grupo implementar un modelo de predicción con red neuronal. Los resultados muestran una buena exactitud predictiva del modelo.

A partir de la revisión bibliográfica es posible determinar que la capacidad de generación de energía de una central fotovoltaica conectada a una red de distribución de energía eléctrica depende de la incidencia del efecto directo de variables meteorológicas como: temperatura ambiente, velocidad del viento, humedad relativa, presencia de nubes, entre otros; a las que se encuentran expuestos los paneles fotovoltaicos. Por tanto, a través de un análisis de minería de datos es posible definir un modelo de predicción de generación de energía con el cual el gestor de la central fotovoltaica puede planificar anticipadamente el uso de esa energía.

Sobre la base de lo mencionado, este artículo presenta el análisis de datos de la central fotovoltaica ubicada en la Comunidad de Paragachi, que consta de 14400 paneles solares y una potencia nominal de 3.6 MW. Las variables de medición local de dos años de operación consideradas son: temperatura ambiente (°C), temperatura del panel (°C), humedad relativa (%), velocidad del viento (m/s), dirección del viento (°), mediciones de piranómetros horizontal, inclinado y dos generales (W/m^2); además se cuenta con mediciones de potencia activa (kW), reactiva (kVAr) y aparente (kVA) total generada por la central. Posteriormente, se define un algoritmo de predicción de la generación de energía basado en máquinas de aprendizaje, que usa como entradas las variables que poseen la mejor información referente a la producción de la central.

En las siguientes secciones se detalla la metodología y el análisis de resultados en base al error medio cuadrático (RMSE) y al error porcentual absoluto medio (MAPE), que son los coeficientes más utilizados en las investigaciones revisadas del estado del arte para valorar la precisión de los estimadores y comparar las diferentes técnicas de predicción.

2. METODOLOGÍA PARA LA PREDICCIÓN DE GENERACIÓN

Uno de los temas de estudio en centrales fotovoltaicas es determinar un modelo de predicción que minimice las diferencias en sus predicciones respecto al valor real de generación de energía de la central, permitiendo que este tipo de tecnología se siga desarrollando y sea considerada de manera óptima y eficaz en el mercado eléctrico.

Varios modelos analizados en el estado del arte utilizan diferentes variables tanto meteorológicas, eléctricas o ambas para entrenar sus modelos de predicción, sin determinar la relación entre variables y tampoco cuál de ellas tienen mayor incidencia en la generación de energía, este criterio es importante para que el predictor no procese información innecesaria y pueda dar la predicción lo más exacto y en el menor tiempo posible.

En este trabajo se propone utilizar técnicas de minería de datos para analizar el comportamiento, incidencia y participación de las variables de medición local en una central fotovoltaica de la región y encontrar cuál de ellas deben ser utilizadas como entrada de la máquina de aprendizaje, usando la siguiente metodología:

- Análisis univariante y multivariante de variables de medición local.
- Determinar el número óptimo de variables y cuál de ellas son las más incidentes para la predicción de generación de potencia.
- Entrenar la máquina de aprendizaje con las variables de mayor incidencia y validar el modelo.

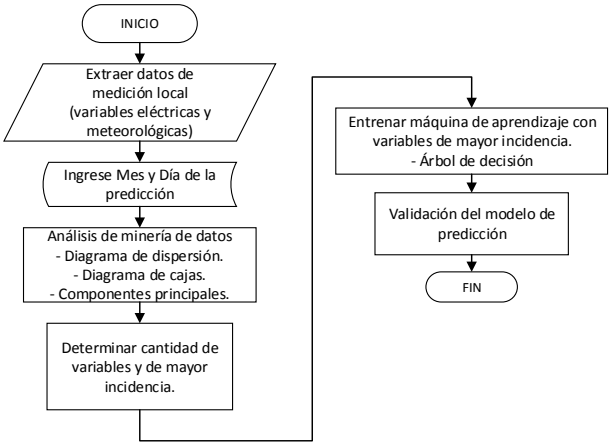

Figura 1: Estructura de la Metodología Aplicada

Las variables de medición corresponden a dos años de operación de la central fotovoltaica tomados en un lapso de 15 minutos. Se inicia con las expresiones matemáticas y definiciones para determinar la potencia generada de un sistema fotovoltaico, con el objetivo de tener una visión general y validar la relación de las variables de entrada referente a la generación de energía.

Las técnicas de minería de datos utilizadas son diagramas de dispersión, diagrama de cajas, componentes principales y la máquina de aprendizaje que usa la técnica de árboles de decisión mediante bosques aleatorios (Random Forest).

2.1. Sistemas Fotovoltaicos (PV)

Las características eléctricas de un módulo PV están dadas por las células fotovoltaicas que lo conforman mediante las curvas P-V e I-V (Fig. 2), y se definen con los parámetros Isc (corriente de corto circuito), Voc (tensión de circuito abierto), Vmpp (tensión máxima), Impp (corriente máxima), Pm (potencia máxima) y FF (factor de forma). [14]

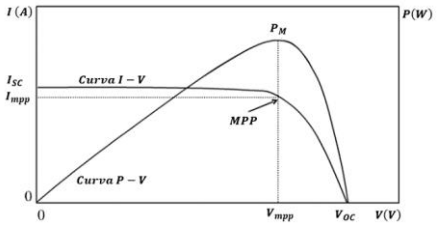

Figura 2: Curva P – V / I – V

El fabricante proporciona los valores de Isc, Voc, Vmpp, Impp calculados a condiciones estándar de una celda de prueba (STC – Standard Test Conditions), cuyas condiciones son las siguientes: Irradiancia G= 1kW/m^2, distribución espectral AM=1.5, temperatura de célula Tc= 25 °C. La máxima potencia del módulo en base a STC se conoce como potencia pico y está dada en vatios-pico (Wp) [15].

Las condiciones en operación real son muy diferentes respecto a los valores estándar e influyen fuertemente en el rendimiento eléctrico de la celda, causando pérdida de eficiencia respecto al valor del STC, esta pérdida puede ser dada por cuatro factores: distribución angular de la luz, contenido espectral de la luz, nivel de irradiancia, temperatura de la celda [15].

La influencia de la temperatura y la irradiancia en el rendimiento de la celda es de gran cuidado para la predicción y enfocarlo en un modelo físico sería poco práctico. En su lugar, se usan métodos matemáticos para trasladar la curva I – V en STC a diferentes puntos de operación, pero dentro de ciertos rangos de temperatura e irradiancia cercanos a los valores de prueba y haciendo uso de otros parámetros como se verá a continuación [15].

Bajo este concepto se tiene que la potencia generada $\eta_{(STC)}$, está dada por la ecuación:

$$\eta_{(STC)} = P_{(STC)} * \frac{G}{1(kW/m^2)} \quad (1)$$

Donde $P_{(STC)}$ es la potencia nominal de la planta en STC y **G** la irrandiancia solar.

Por otra parte, el balance de potencia en estado estable determina la temperatura de la celda y la misma aumenta linealmente en base a la irradiancia (2), cuyo coeficiente depende de la instalación del módulo, velocidad del viento, humedad, entre otros. Esta información se encuentra contenida en la temperatura de operación normal de la celda (NOCT), cuya definición es la temperatura de la celda a temperatura ambiente Tamb=20°C, Irrandiancia G= 0,8 kW/m^2 y velocidad de viento a 1 m/s; este valor es cercano a 45°C [15].

$$T_{cell} = T_{amb} + G * \frac{NOCT - 20°C}{0.8(kW/m^2)} \quad (2)$$

La corriente de corto circuito y la tensión de circuito abierto está dada por la ecuación (3) y (4), respectivamente, donde se observa la influencia de la irradiancia y la temperatura de la celda.

$$I_{SC(Tcell,G)} = I_{SC(STC)} * \frac{G}{1(kW/m^2)} * (1 + \alpha(T_{cell} - 25°C)) \quad (3)$$

$$V_{OC(Tcell,G)} = V_{OC(STC)} - \beta(T_{cell} - 25°C)) \quad (4)$$

Donde los coeficientes α y β, representan el incremento relativo de corriente y voltaje, respectivamente. El primero tiene un valor normalmente de 0.4 %/°C y el segundo de 2 mV/°C.

Con este estudio se observa la relación entre las variables de medición y se da una mejor idea de los resultados esperados en la presente investigación. Además, se puede notar la dificultad para calcular la generación de un PV debido al comportamiento estocástico de las variables que dependen bastante del medio donde se encuentra el sistema y su equipamiento.

2.2. Diagramas de cajas

Permiten visualizar y comparar la distribución y la tendencia central de valores numéricos mediante cuartiles los cuales son una forma de dividir valores numéricos en cuatro grupos iguales basados en cinco valores clave: mínimo, primer cuartil, mediana, tercer cuartil y máximo [16].

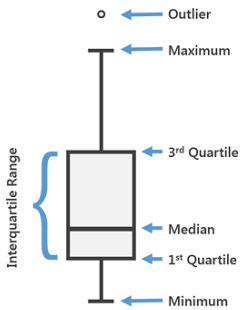

Figura 3: Diagramas de Cajas

Ilustra los valores mínimos y máximos de los datos mediante bigotes que se extienden desde la caja y a su vez presenta como puntos los datos atípicos (outliers). Cuando se crea un diagrama de caja a partir de campos numéricos, se aplica una estandarización que permite que las variables con diferentes unidades de medida sean comparables [16].

En el estudio se aplica esta técnica para determinar valores atípicos del sistema y analizar su origen para eliminarlos o considerarlos dentro del predictor, por otra parte, se podrá observar la distribución y relación de variables sin importar la unidad de medida que tengan, dando un enfoque más significativo y de gran ayuda al análisis de datos.

2.3. Análisis de Componentes Principales (PCA)

El propósito de análisis de componentes principales es reducir un espacio de dimensión *p* a un nuevo espacio de dimensión *d*, donde *d* es mucho menor que *p*, mientras que sigue representando la variación de los datos tanto como sea posible. Con este análisis, se transforman los datos en un nuevo conjunto de coordenadas o variables que son una combinación lineal de las variables originales [17].

Para el estudio se utiliza está técnica para determinar una relación multivariable y la cantidad de variables que se deben utilizar para el predictor sin perder información relevante y de suma importancia de la central.

2.4. Máquina de Aprendizaje – Árboles de Decisión – Random Forest

Los árboles de decisión incorporan un enfoque de clasificación supervisada, que se estructura en base a un árbol que se compone de una raíz, nodos, ramas y hojas. En definitiva, es un conjunto de condiciones organizadas en una estructura jerárquica cuya decisión final a considerar se determina siguiendo las condiciones que cumplen desde la raíz hasta sus hojas. Esta estructura se desarrolla en función de los valores de las variables y atributos disponibles [18].

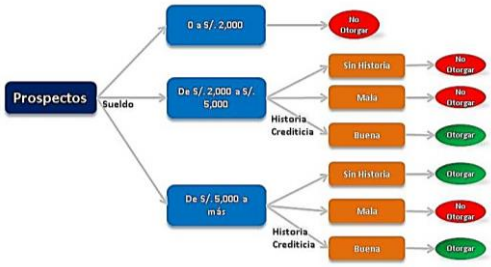

Figura 4: Árbol de decisión

Sus resultados son muy favorables: la clasificación final es simple y eficiente, robusto respecto a los outliers y puntos mal clasificados, se puede visualizar gráficamente; por otra parte, las desventajas principales de esta técnica son: clasificación aleatoria de valores perdidos, árboles grandes tienen tendencia a sobre ajustar los datos y son inestables (pequeños cambios en los datos iniciales producen árboles muy distintos) [19]

Random Forest es una metodología de árboles de decisión (clasificadores débiles) que trabaja con una colección de árboles incorrelacionados y los promedia, cada árbol depende de valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque, esto reduce considerablemente la inestabilidad presente en los árboles de decisión [18].

Es un algoritmo compuesto por numerosos árboles de decisión, en el que se definen una cantidad de árboles a desarrollar y una cantidad de atributos *m* tal que sea menor a la cantidad total de atributos. Entre los árboles se reparten *k* patrones con reemplazo y se desarrollan los árboles, el resto de los patrones son usados para calcular el error. Al desarrollar cada nodo se eligen *m* atributos y se determina el mejor atributo para desarrollar el nodo, en el entrenamiento los patrones son repartidos aleatoriamente con repetición entre cada árbol y se determina la mejor partición del conjunto de entrenamiento. Para la predicción, el nuevo caso es empujado hacia abajo por cada uno de los árboles y se le asigna la etiqueta del nodo terminal. Todos los árboles dan una etiqueta y la mayor cantidad de incidencias es reportada como la predicción [20].

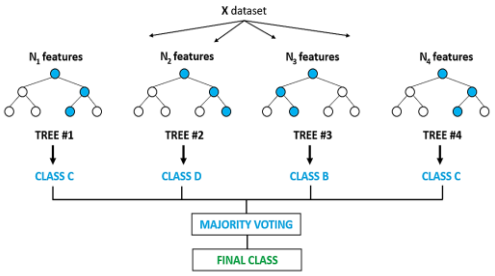

Figura 5: Bosques Aleatorios – Random Forest

Cada uno de los árboles que conforman el random forest ha sido diseñado para ajustarse a diferentes escenarios, pero todos parecidos al escenario real a aprender, esto aporta coherencia al random forest y la capacidad de ajustarse adecuadamente a nuevos escenarios desconocidos [20]

Entre las ventajas principales de este algoritmo se encuentra: buenos resultados en estudios empíricos, se ejecuta de forma eficiente sobre grandes bases de datos, trata miles de variables sin eliminar ninguna, proporciona estimaciones de la importancia de cada variable, dispone de un método efectivo de estimación de valores perdidos. Por otra parte, las desventajas encontradas en el estado de arte citan: el algoritmo sobre ajusta en ciertos grupos de datos con tareas de predicción ruidosas, la clasificación es difícil de interpretar y si los datos contienen grupos de atributos correlacionados con similar relevancia para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes produciendo incoherencia en los resultados [19].

En [21] se hace un estudio comparativo entre diferentes máquinas de aprendizaje que utilizan las técnicas de regresión lineal, random forest y redes neuronales para la predicción de generación de energía en paneles fotovoltaicos. Se puede observar que la técnica de random forest presenta excelentes resultados

con un coeficiente de correlación de 0.986, pero no alcanza a superar a la técnica de redes neuronales con un valor de 0.997. En el estudio se utilizan variables como: irradiancia global, elevación del panel, temperatura ambiente, velocidad del viento, humedad relativa, dirección del viento; no se hace una clasificación de cuál tiene mayor incidencia sobre la salida, generación de energía.

Para esta investigación, con el análisis de variables se utilizará el predictor Random Forest únicamente con las variables de mayor incidencia sobre la generación de energía, esto permite que el predictor sea más preciso y su tiempo de respuesta sea menor. Además, con el estado del arte se observa que este algoritmo brinda una adecuada respuesta a sistemas con parámetros de comportamiento estocástico.

3. ANÁLISIS DE RESULTADOS

Para comprobar la efectividad de la capacidad de predicción de generación de energía de la central fotovoltaica mediante la máquina de aprendizaje Random Forest, es necesario previamente realizar una serie de análisis de los datos recolectados tanto univariante como multivariante con el fin de detectar los posibles outliers e identificar las relaciones existentes entre las variables.

La correlación existente entre variables se resume en la Tabla 1, donde si el valor de correlación entre las variables es cercano a 1, la correlación es alta. Por ejemplo, las variables con mayor influencia con la generación de energía es la radiación directa sobre los paneles, dato medido por los piranómetros (horizontal, inclinado y general). De la misma manera se puede observar la relación de la velocidad del viento y humedad relativa con respecto a la potencia reactiva, evidenciando que, para mejorar el factor de potencia los paneles deben ser enfriados paulatinamente para lo cual se utiliza la corriente del viento como un extractor natural de calor.

Tabla1: Correlación entre variables de medición.

Variables Medidas	kVAr	kVA	kW
Humedad relativa (%)	0.473	-0.611	-0.612
Piranómetro general (W/m^2)	-0.780	0.950	**0.951**
Piranómetro horizontal (W/m^2)	-0.784	0.953	**0.953**
Piranómetro inclinado (W/m^2)	-0.786	0.951	**0.952**
Temperatura Panel (°C)	-0.007	0.043	0.044
Temperatura ambiente (°C)	-0.544	0.767	**0.771**
Weather Station Pyr (W/m^2)	-0.849	0.988	**0.987**
Velocidad del viento (m/s)	-0.383	0.490	0.499
Dirección del viento (°)	0.314	-0.503	-0.509
Potencia Reactiva (kVAr)	1.000	-0.823	-0.818
Potencia Aparente (kVA)	-0.828	1.000	0.999
Potencia Activa (kW)	-0.819	0.999	1.000

El análisis de outliers se presenta para las variables con mayor grado de correlación referente a la Tabla 1, y mediante diagrama de cajas son presentadas en la Fig. 6 y Fig. 7.

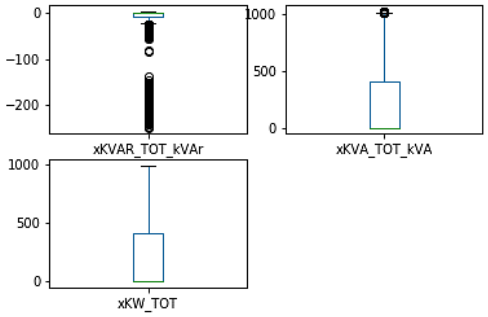

Figura 6: Diagrama de cajas de variables de potencia

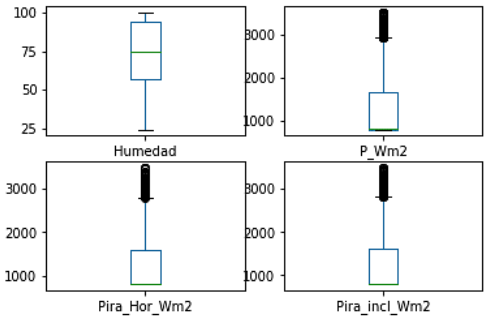

Figura 7: Diagrama de cajas de variables de correlación

Se observa la gran cantidad de valores atípicos que poseen los datos recolectados, en especial para la potencia reactiva, lo cual corresponde a la característica propia de los paneles de ser un modelo no lineal. Otro factor importante para la aparición de valores atípicos es debido a los baches en la generación causados por el cambio drástico de la radiación directa sobre los paneles fotovoltaicos a causa de la aparición esporádica de nubes, absorción de agua, CO_2 y gases en el ambiente, factores que provocan incertidumbre en la predicción. Mediante técnicas de estandarización fue posible limitar la aparición de estos valores para que no tengan un alto grado de significancia y por ende no puedan alterar el entrenamiento del predictor.

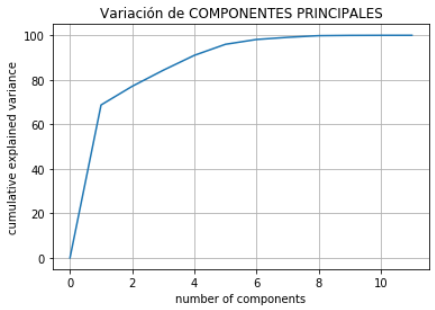

Figura 8: Variación de componentes principales

Consecutivamente se aplica el análisis de componentes principales con el fin de minimizar de manera eficiente la cantidad de variables iniciales que se tienen en estudio, todo esto sin que exista el riesgo de

afectar la capacidad de predicción. En la Fig. 8, se observa el número de componentes vs. la varianza explicativa acumulada, donde se llega a la conclusión de que para caracterizar al sistema de aprendizaje es necesario utilizar un máximo de cinco variables con una efectividad del 95.95% en similitud, Tabla 2.

Tabla 2: Variabilidad Explicada

N° PCA	1	2	3	4	5	6	7
Varianza (%)	68.7	77.0	84.2	91.0	95.9	98.1	99.0

Previo a la implementación de este modelo es indispensable estandarizar las variables debido a sus unidades como ya se mencionó anteriormente. Para entrenar al sistema se utiliza la característica del "randomforestregressor", ingresándose las variables más características de la central fotovoltaica, para la validación se utiliza el Cross-Validation en un valor seteado de 10 y con los datos de cada mes del año 2015 se entrena al modelo para predecir los valores del año 2016.

En la Fig. 9, se presenta la predicción obtenida para el mes de enero donde se obtiene un RMSE de 31.65 y un MAPE de 4.27%. En el esquema la gráfica de color verde representa la predicción y la de color rojo es la curva real de potencia de valores históricos.

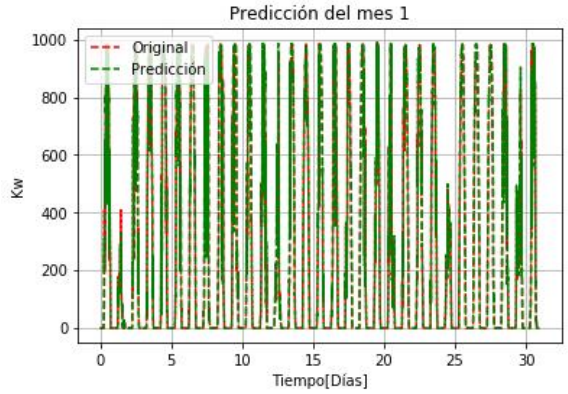

Figura 9: Testing mes Enero

En la Fig. 10a, se muestra la predicción para el día 12 del mes de enero con un RMSE de 23.96 y en la Fig.10b el día 15 del mes de agosto con un RMSE de 26.7.

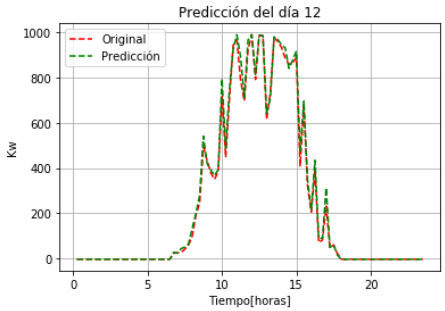

Figura 10a: Día nublado o condiciones ambientales variables

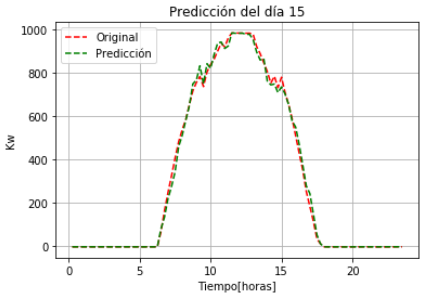

Figura 10b: Día con condiciones ambientales uniformes

Se observa que el modelo elaborado en base a la máquina de aprendizaje RandomForest es capaz de predecir la capacidad de generación de energía ante eventos que la afectan, como la absorción de agua, CO_2 y gases que se presentan en el ambiente. Estos efectos producen baches que afectan a la característica del modelo base de una central fotovoltaica, estos baches se pueden visualizar en la Fig. 8, mientras que en la Fig. 9 se presenta un comportamiento que se asemeja al modelo base de una central fotovoltaica.

En la Tabla 3, se hace un resumen de los valores RMSE y MAPE obtenidos con el modelo de Random Forest para cada uno de los meses del año.

Tabla 3: Valores de RMSE y MAPE en cada mes de predicción

Mes	**Accurancy**	**RMSE**	**MAPE**
1	98.87%	31.65	4.27%
2	98.76%	35.85	5.37%
3	99.45%	22.40	4.56%
4	99.25%	27.93	4.42%
5	98.77%	33.02	5.14%
6	98.78%	30.67	5.33%
7	98.52%	34.88	5.27%
8	99.16%	29.14	4.58%
9	98.66%	33.18	4.71%
10	98.30%	31.68	4.65%
11	98.93%	28.01	4.21%
12	98.25%	31.42	4.43%
PROMEDIO	98.80%	30.81	4.75%

La predicción de generación de energía de una central fotovoltaica se ha comparado con diferentes métodos propuestos por varios autores, Tabla 4. En [22] se realiza el análisis con dos técnicas SVM y PCA – SVM obteniendo errores del 15.48% y 13.48%, respectivamente. En [23] se aplica el modelo BP neural network y el TS fuzzy neural machine con errores del 10.4% y 5.6%. Finalmente, en [24] se presentan modelos combinados con máquinas de soporte vectorial obteniendo errores del 8.68% para el LS – SVM y del 5.11% para el CS – LSSVm.

Tabla 4: Comparación con modelos de predicción

MODELO DE PREDICCIÓN	MAPE (%)
PCA – RANDOM FOREST (Metodología aplicada en el artículo)	4.75
Curkoo Search – Least Square Support Vector Machine (CS – LSSVM)	5.11
Takagi – Sugeno (TS) Fuzzy Neutral Network	5.61
Least Squares – Support Vector Machine (LS – SVM)	8.68
Back Propagation (BP) neutral network	10.4
PCA – SVM	13.48
SVM	15.48

Con el método propuesto PCA – RANDOM FOREST se ha logrado alcanzar resultados positivos ya que el error de predicción se reduce a un valor de 4.75%. Por otra parte, aunque no se considere en el estudio el tiempo de respuesta del predictor es bajo y en futuros trabajos se podría aplicarlo en tiempo real.

4. CONCLUSIONES Y RECOMENDACIONES

La minería de datos en la actualidad es de vital importancia para analizar sistemas que tienen variables con comportamiento estocástico, como en el presente estudio con sistemas fotovoltaicos. Con la ayuda de esta herramienta se logra reconocer patrones que caracterizan el comportamiento de las variables y también los posibles atípicos que se presentan en las mediciones locales reales. Por otra parte, es de gran ayuda para observar las relaciones existentes entre diferentes variables de un sistema sin necesidad de implementar un modelo matemático.

La predicción de la capacidad de generación de energía eléctrica de una central fotovoltaica es de gran interés, ya que al ser una fuente de energía renovable con una alta prioridad de despacho su operación incide en gran manera sobre los costos de mercado, dicha premisa conlleva a la necesidad de predicción del momento oportuno de la integración de una central fotovoltaica a la red. Para este caso la predicción se logra de manera eficiente por medio de la determinación de un modelo basado en aprendizaje de máquina en el cual se alcanza una precisión promedio del 98.7%.

Las máquinas de aprendizaje si son adecuadamente entrenadas son capaces de predecir el comportamiento del sistema con un bajo error. Es de suma importancia que las variables que se ingresan en la máquina sean las necesarias y no emplear datos sin utilidad, porque esto hace que el sistema se entrene de forma errónea y se demore en procesar la información, además es necesario estandarizar todas las variables en una sola representación, por unidad, para que se encuentre de la mejor manera las variables de entrada con mayor incidencia sobre las variables de salida.

AGRADECIMIENTOS

Al Doctor Jaime Cepeda por los conocimientos impartidos en la Maestría de Redes Eléctricas Inteligentes de la Escuela Politécnica Nacional, Quito, Ecuador.

REFERENCIAS BIBLIOGRÁFICAS

[1] Agencia de Regulación y Control de Electricidad (ARCONEL), “Estadística anual y multianual del sector eléctrico ecuatoriano.” 2017.

[2] CONELEC, «Atlas Solar del Ecuador con fines de Generación Eléctrica», ago-2008. [En línea]. Disponible en: https://www.ariae.org/servicio-documental/atlas-solar-del-ecuador-con-fines-de-generacion-electrica. [Accedido: 22-jul-2019].

[3] D. J. Bartholomew, G. E. P. Box, and G. M. Jenkins, “Time Series Analysis Forecasting and Control.,” Oper. Res. Q., vol. 22, no. 2, p. 199, 1971.

[4] J. M. Santos, J. M. Pinazo, and J. Canada, “Methodology for generating daily clearness index index values Kt starting from the monthly average daily value K̄t. Determining the daily sequence using stochastic models,” Renew. Energy, vol. 28, no. 10, pp. 1523–1544, 2003.

[5] C. Craggs, E. M. Conway, and N. M. Pearsall, “Stastitical investigation into optimal averaging time for solar irradiance,” Sol. Energy, vol. 68, no. 2, pp. 179–187, 2000.

[6] Y. Li, Y. Su, and L. Shu, “An ARMAX model for forecasting the power output of a grid connected photovoltaic system,” Renew. Energy, vol. 66, pp. 78–89, 2014.

[7] S. Wang, N. Zhang, Y. Zhao, and J. Zhan, “Photovoltaic system power forecasting based on combined grey model and BP neural network,” 2011 Int. Conf. Electr. Control Eng. ICECE 2011 - Proc., no. 2, pp. 4623–4626, 2011.

[8] F. Wang, Z. Mi, S. Su, and H. Zhao, “Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters,” Energies, vol. 5, no. 5, pp. 1355–1370, 2012.

[9] J. Shi, W. J. Lee, Y. Liu, Y. Yang, and P. Wang, “Forecasting power output of photovoltaic systems based on weather classification and support vector machines,” IEEE Trans. Ind. Appl., vol. 48, no. 3, pp. 1064–1069, 2012.

[10] A. Mellit, S. Sağlam, and S. A. Kalogirou, “Artificial neural network-based model for estimating the produced power ofaphotovoltaic module,” Renew. Energy, vol. 60, pp. 71–78, 2013.

[11] N. Zhang, P. K. Behera, and C. Williams, “Solar radiation prediction based on particle swarm optimization and evolutionary algorithm using recurrent neural networks,” SysCon 2013 - 7th

Annu. IEEE Int. Syst. Conf. Proc., pp. 280–286, 2013.

[12] M. Bouzerdoum, A. Mellit, and A. Massi Pavan, "A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant," Sol. Energy, vol. 98, no. PC, pp. 226–235, 2013.

[13] Z. Yang, Y. Cao, and J. Xiu, "Power generation forecasting model for photovoltaic array based on generic algorithm and BP neural network," CCIS 2014 - Proc. 2014 IEEE 3rd Int. Conf. Cloud Comput. Intell. Syst., pp. 380–383, 2014.

[14] R. Paola and L. Pérez, "Diseño y simulación de un sistema de generación fotovoltaica para una cocina de inducción," vol. 5, 2007.

[15] Gray, J. L., Luque, A., & Hegedus, S. (2003). Handbook of photovoltaic science and engineering. Luque and S. Hegedus, Eds. West Sussex, England: John Wiley & Sons, 14, Sección 7.9.

[16] K. Potter, H. Hagen, A. Kerren, and P. Dannenenmenn, "Methods for Presenting Statistical Information: The Box Plot," Vis. Large Unstructured Data Sets, vol. S-4, pp. 97–106, 2006.

[17] A. Sánchez López, V. Cruz-Gutiérrez, M. Alberto Posada-Zamora, M. M. Teresa Torrijos, and M. Auxilio Osorio Lama, "A Study of Principal Components Analysis of Air Quality Databases," Res. Comput. Sci., vol. 120, pp. 9–19, 2016.

[18] R. F. Medina-Merino and C. I. Ñique-Chacón, "Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python," Interfases, vol. 0, no. 010, p. 165, 2017.

[19] P. Pino, C. Tutora, E. Gutiérrez, and M. P. Somet, "Evaluación del riesgo crediticio mediante árboles de clasificación y bosques aleatorios," 2017.

[20] E. Enrique et al., "Métodos de clasifiación para identificar lesiones en piel a partir de espectros de reflexión difusa," Rev. Ing. Biomédica, vol. 4, no. 8, pp. 34–40, 2010.

[21] M. Kayri, I. Kayri, and M. T. Gencoglu, "The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data," 2017 14th Int. Conf. Eng. Mod. Electr. Syst. EMES 2017, pp. 1–4, 2017.

[22] S. Qijun, L. I. Fen, Q. Jialin, and Z. Jinbin, "Photovoltaic Power Prediction Based on Principal Component Analysis and Support Vector Machine," 2016.

[23] L. Kaiju, L. Xuefeng, M. Chaoxu, and W. Dan, "Short-Term Photovoltaic Power Prediction Based on T-S Fuzzy Neural Network," 2018 33rd Youth Acad. Annu. Conf. Chinese Assoc. Autom., pp. 620–624, 2018.

[24] M. Aidil, A. Aziz, Z. M. Yasin, and Z. Zakaria, "Prediction of Photovoltaic System Output using Hybrid Least Square Support Vector Machine," vol. 3, no. October, pp. 2–3, 2017.

Cristian Fabara. - Nació en Quito en 1989. Sus estudios universitarios los realizó en la Escuela Politécnica Nacional, obteniendo su título de Ingeniero en Electrónica en Control y Automatización. Actualmente, se encuentra cursando sus estudios de maestría en la Escuela Politécnica Nacional en el área de Eléctrica, Mención Redes Eléctricas Inteligentes. Sus campos de investigación están relacionados con el desarrollo de energías renovables y sistemas inteligentes para su integración con el Sistema Eléctrico de Potencia.

Diego Maldonado. - Nació el 12 de noviembre de 1990 en la ciudad de Tulcán. Graduado Cum Laude de la carrera de Ingeniería Electrónica y Control en la Escuela Politécnica Nacional en octubre 2014. Actualmente, se encuentra cursando sus estudios de maestría en la Escuela Politécnica Nacional en el área de Eléctrica, Mención Redes Eléctricas Inteligentes Áreas de Interés: Control Industrial, SmartGrids, Sistemas eléctricos de potencia.

Mauricio Soria Colina. - Nació en Ambato, Ecuador en 1991. Recibió su título de Ingeniero Eléctrico de la Escuela Politécnica Nacional en el 2016. Actualmente, se encuenta cursando sus estudios de Maestría en la Escuela Politécnica Nacional con mención en Redes Eléctricas Inteligentes. Sus campos de investigación están relacionados con el desarrollo de Protecciones Eléctricas y Algoritmos Inteligentes para Smart Grids.

Antonio Tovar Arboleda. - Nació en La Maná, Ecuador en 1987. Recibió su título de Ingeniero en Electrónica y Control de la Escuela Politécnica Nacional en abril del 2014. Actualmente se encuentra cursando sus estudios de posgrado en la Maestría en Electricidad mención Redes Eléctricas Inteligentes de la Escuela Politécnica Nacional. Sus áreas de interés son: Automatización de Subestaciones, Interoperabilidad de Redes Industriales, Sistemas SCADA, Sistemas OMS, ADMS.