Revista de investigación de sistemas e informática
16(1): 21-33 (2023)
DOI: https://doi.org/10.15381/risi.v16i1.25055

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Metodología de comparación de técnicas de corto plazo para el pronóstico de la producción de energía eléctrica de plantas solares fotovoltaicas

Comparison methodology of short-term techniques for forecasting the production of electrical energy from photovoltaic solar plants

César Aristóteles Yajure Ramírez
https://orcid.org/0000-0002-3813-7606
cyajure@gmail.com
Universidad Central de Venezuela. Caracas, Venezuela

Recibido: 24/02/2023 - Aceptado: 26/04/2023 - Publicado: 21/08/2023

RESUMEN

Cada vez hay más conciencia acerca de la necesidad de la utilización de fuentes de energía renovables para la producción de electricidad. Entre estas fuentes, resalta la energía solar fotovoltaica por su gran crecimiento en los últimos años, y por la posibilidad que tienen las plantas solares fotovoltaicas de funcionar en las modalidades *on-grid y off-grid*. En cualquiera de los dos casos, es importante tener un pronóstico de la producción de energía eléctrica de la planta. En esta investigación se propone una metodología para comparar los pronósticos de corto plazo de la energía eléctrica producida por una planta solar fotovoltaica, utilizando algoritmos de aprendizaje automático, el análisis de series de tiempo, y haciendo uso de los datos históricos de la producción de la planta y de las variables climáticas. Se obtuvieron modelos de pronósticos con datos tanto en resolución temporal horaria como con resolución temporal de quince minutos. Las métricas de desempeño utilizadas en la evaluación de los modelos fueron: R^2, RMSE, MAE, y MAPE. Para la resolución de quince minutos, ambos modelos, red neuronal artificial y K vecinos más cercanos, cumplieron con los supuestos estadísticos, pero el modelo de red tuvo los valores más bajos de las métricas. Para la resolución horaria, de nuevo el modelo de red neuronal artificial tuvo el mejor desempeño sobre el modelo ARIMA del análisis de series de tiempo.

Palabras clave: Arima, aprendizaje automático, ciencia de datos, modelos matemáticos, red neuronal artificial, K vecinos más cercanos.

ABSTRACT

There is increasing awareness about the need to use renewable energy sources to produce electricity. Among these sources, photovoltaic solar energy stands out due to its great growth in recent years, and the possibility that photovoltaic solar plants have of operating in on-grid and off-grid modes. In either case, it is important to have a forecast of the plant's electrical energy production. In this research, a methodology is proposed to compare the short-term forecasts of the electrical energy produced by a photovoltaic solar plant, using machine learning algorithms, time series analysis, and making use of the historical data of the production of the plant and climatic variables. Forecast models were obtained with data in both hourly temporal resolution and fifteen-minute temporal

resolution. The performance metrics used in the evaluation of the models were: R2, RMSE, MAE, and MAPE. For the resolution of fifteen minutes, both models, artificial neural network, and K nearest neighbors, fulfilled the statistical assumptions, but the network model had the lowest values of the metrics. For hourly resolution, again the artificial neural network model had the best performance over the ARIMA model of time series analysis.

Keywords: Arima, machine learning, data science, mathematical models, artificial neural network, K nearest neighbors.

I. INTRODUCCIÓN

Las consecuencias del cambio climático han originado un aumento constante del uso de fuentes de energías renovables para la producción de energía eléctrica. Dentro de estas fuentes renovables, resalta la energía solar fotovoltaica por sus bajas emisiones de efecto invernadero de acuerdo con lo planteado en el reporte del 2022 de la IEA (Agencia Internacional de Energía, 2022). Asimismo, es una de las de mayor crecimiento, pues tan sólo en el año 2021 tuvo un aumento del 23% con respecto al año 2020, tal como lo indica el reporte de estatus global de energías renovables (Comunidad global de energías renovables REN21, 2022). Pero, debido a que la fuente primaria de las plantas solares fotovoltaicas, la cual es la energía solar, está disponible en horas específicas del día y que además esta disponibilidad es en ocasiones intermitente, se hace necesario garantizar la estabilidad de la red eléctrica y vigilar exhaustivamente el balance entre el consumo de energía eléctrica y la producción de esa energía.

Una de las herramientas más importantes para contrarrestar la variabilidad y la incertidumbre que introduce el comportamiento de la fuente primaria de las plantas solares fotovoltaicas, es el pronóstico de la producción de dichas plantas. De acuerdo con el reporte PVPS de la Agencia Internacional de Energía (Technology Collaboration Programme - IEA, 2022), el pronóstico de energía eléctrica de la planta es indispensable para poder comercializar su producción en los mercados eléctricos diarios o intradiarios, entre otros posibles usos. Asimismo, en el reporte técnico del Laboratorio Nacional de Energías Renovables (NREL) de los Estados Unidos (National Renewable Energy Laboratory, 2018) consideran que la resolución temporal ideal de los datos para el pronóstico es el horario, que el horizonte de pronóstico pudiera ser de una semana, que los pronósticos se deberían actualizar cada cuatro horas, y que la métrica típica para medir la exactitud del pronóstico pudiera ser, entre otras, la raíz cuadrada del error cuadrático medio (RMSE).

En vista de lo planteado hasta ahora, el objetivo de esta investigación es desarrollar una metodología de comparación de técnicas de pronóstico de corto plazo de la energía eléctrica producida por plantas solares fotovoltaicas, con resolución temporal horaria, pero también con resolución temporal de quince minutos. Las técnicas por utilizar son: red neuronal artificial, algoritmo de K vecinos más cercanos, y el análisis de series de tiempo a través de la metodología Box-Jenkins. La metodología general planteada se aplica a una planta solar específica, considerando tanto variables eléctricas como variables climáticas. La definición de pronóstico de corto plazo se utiliza tal como se propone en el resumen avanzado de pronóstico de energía eléctrica a partir de fuentes renovables de la Agencia Internacional de Energías Renovables (IRENA, 2020), en el que se indica que el horizonte de pronóstico de corto plazo va de una hora a veinticuatro horas, y es útil para incrementar la estabilidad de la red eléctrica.

Por otra parte, se hizo una revisión del estado del arte relacionado con el objetivo de esta investigación, encontrándose una variedad de investigaciones en el área, algunas de las cuales se presentan a continuación. Larson, Nonnenmacher, & Coimbra (2016) presentan el pronóstico de la potencia de salida de plantas solares fotovoltaicas ubicadas en California, utilizando optimización de mínimos cuadrados de la predicción meteorológica numérica (NWP). Para evaluar el desempeño de los modelos utilizaron el MAE, el error medio de sesgo, y el RMSE normalizado. Konstantinou, Peratikou, & Charalambides (2021) realizan el pronóstico de la producción de energía eléctrica de una planta solar fotovoltaica ubicada en Nicosia, Chipre, utilizando datos históricos y con un horizonte de 1,5 horas. Hacen uso de una red neuronal tipo LSTM, y evalúan el modelo a través del RMSE, del cual obtuvieron un valor de 0,11368. Tu, Hong, & Lin (2021) utilizan una red neuronal de regresión basada en optimización GWO para realizar el pronóstico de plantas solares fotovoltaicas. Utilizan como métricas de desempeño de los modelos el MRE, MAE, RMSE, MAPE, MBE, y el coeficiente de determinación R^2. Concluyen que el modelo obtenido tiene

mejor desempeño que otros modelos considerados para comparación. Fan, Wei, Chen, & Hong (2022), proponen un modelo combinado de pronóstico de la producción de plantas solares fotovoltaicas basado en análisis de series de tiempo, red neuronal de propagación, y regresión de soporte vectorial, para contrarrestar la baja exactitud de los modelos tradicionales. Se utilizan datos históricos de plantas ubicadas en los Estados Unidos, y evalúan los modelos a través de las métricas MAE, MSE, RMSE, y MAPE, obteniendo valores de 0,53, 0,41, 0,64 and 0,84 respectivamente. Arias & Bae (2021) presentan un modelo de pronóstico para calcular la energía solar fotovoltaica en función de datos históricos de la radiación solar, la eficiencia del sistema fotovoltaico, las características del sistema fotovoltaico y datos meteorológicos, utilizando herramientas de *big data*. De su modelo obtienen 17,57 kWh para el RMSE y 2,80% para el error relativo medio. Fara, Diaconu, Craciunescu, & Fara (2021) utilizan modelos ARIMA y redes neuronales artificiales para el pronóstico de la producción de energía de dos plantas solares fotovoltaicas ubicada en Rumania. Las métricas utilizadas para evaluar los modelos fueron: el RMSE relativo, y el MAE relativo, resultando que el modelo ARIMA tuvo mejor desempeño que el modelo de red neuronal artificial. Jung et al. (2022) proponen un sistema regional de pronóstico de producción de energía eléctrica de plantas solares fotovoltaicas, con horizonte horario y utilizando un modelo de vector autorregresivo (VAR), cuyo desempeño comparan con un modelo ARIMA, utilizando las métricas: RMSE normalizado y el coeficiente de determinación R^2. Concluyen que el modelo VAR tiene mejor desempeño que el modelo ARIMA, de acuerdo con cada una de las métricas consideradas.

El resto del artículo se organiza de la siguiente manera. En la sección 2 se presenta la metodología propuesta, en la sección 3 se discuten los resultados obtenidos luego de aplicar las técnicas consideradas, para finalmente presentar las conclusiones que se derivan de la investigación, y las referencias bibliográficas utilizadas.

II. METODOLOGÍA

La metodología utilizada consta de las etapas que conforman un proyecto de ciencia de datos, mientras que cada una de esas etapas a su vez tiene su propia metodología. Resalta la etapa de modelación que, dependiendo del tipo de algoritmo matemático a utilizar, requerirá de una metodología específica. De acuerdo con lo planteado por Rogel-Salazar (2017, p. 3) “En pocas palabras, la ciencia de datos puede ser entendida como la extracción de conocimiento de varias fuentes de datos y su uso adecuado para hacerlo útil, y las habilidades necesarias para lograrlo varían desde la programación hasta el diseño, y de las matemáticas a la narración.” En ese mismo sentido, (VanderPlas, 2017) plantea que, la combinación de un buen dominio de matemáticas y estadística para modelar conjuntos de datos, con la habilidad sobre el área de investigación del cuál provienen los datos, y la habilidad computacional para diseñar y usar algoritmos que permitan explorar y modelar los datos, define lo que es la ciencia de datos.

Cielen, Meysman, & Ali (2016) proponen seis etapas para un proyecto de ciencia de datos. Estas etapas se podrían aplicar de manera secuencial, pero no necesariamente será así, lo que dependerá del caso particular que se esté tratando. La primera etapa consiste en establecer el o los objetivos de la investigación, para lo cual se requiere conocer el área técnica de la cual provienen los datos a utilizar. La segunda etapa es la recuperación u obtención de los datos, los cuales procederán de una o varias fuentes, internas o externas. Una vez obtenidos los datos, la siguiente etapa consiste en preparar los datos para que puedan ser utilizados en las subsiguientes etapas. Se aplican técnicas de procesamiento de datos como las planteadas en (McKinney, 2018), las cuales incluyen, entre otras, la detección e imputación de datos faltantes, la detección e imputación de datos atípicos, la remoción de datos duplicados, la transformación de datos para que por ejemplo tengan las unidades y formatos adecuados, y la combinación de datos para crear otras variables. La cuarta etapa consiste en hacer un análisis exploratorio de los datos, utilizando herramientas estadísticas analíticas y gráficas, ya sean univariadas o multivariadas. La quinta etapa de la metodología reside en hacer la modelación de los datos, tomando en cuenta los objetivos de la investigación y los resultados de la etapa previa. Para la modelación, por lo general se aplican algoritmos de aprendizaje automático, pero en ocasiones pudieran ser otro tipo de técnicas como por ejemplo el análisis de series de tiempo. Para esta investigación en específico se aplican algoritmos para crear modelos de regresión, y la metodología Box-Jenkins para desarrollar modelos ARIMA. Por último, la sexta etapa radica en presentar los resultados obtenidos, y en ocasiones automatizar la ejecución del proceso.

La metodología Box-Jenkins se emplea en series de tiempo, y es detallada por (Cryer & Chan, 2008), quienes indican que está conformada por tres fases: identificación, estimación y prueba, y aplicación.

Su fortaleza radica en que se puede aplicar a series estacionarias, pero también a series no estacionarias. La primera fase está compuesta por dos pasos: la preparación de los datos y la selección del modelo inicial. En la preparación se aplican las técnicas mencionadas previamente para la ciencia de datos, pero adicionalmente, y en caso de ser necesario, se podrían transforman los datos para estabilizar su varianza y/o se diferencian para convertir la serie de tiempo a estacionaria. La selección de los modelos potenciales se hace inicialmente observando la posible existencia de tendencia y estacionalidad en la gráfica temporal de la serie, y además utilizando las funciones de autocorrelación y autocorrelación parcial. En la segunda fase también se tienen dos pasos, estimación y diagnóstico. Se estiman los parámetros de cada uno de los modelos potenciales, y se utiliza una métrica de comparación para seleccionar el mejor de esos modelos. La métrica que generalmente se utiliza es el AIC (Akaike Information Criteria), ya que de acuerdo con Mills (2019, p. 28), "Hay una variedad de criterios de selección que se pueden utilizar para elegir un modelo apropiado, de los cuales quizás el más popular es el de Akaike". En paralelo debe verificarse que los residuos cumplan con los supuestos estadísticos, lo cual corresponde al diagnóstico. Finalmente, una vez que se tiene el modelo óptimo según el criterio AIC, el cual además cumple con los supuestos estadísticos de los residuos, se llega a la fase de aplicación, la que consiste en utilizar el modelo obtenido para efectuar el pronóstico solicitado.

A continuación, se presentan las etapas de obtención de los datos y preparación de los datos, mientras que en la siguiente sección se presentan las etapas de análisis exploratorio de los datos y la aplicación de los algoritmos para obtener los modelos de pronóstico.

Obtención de los datos

Los datos fueron obtenidos de la plataforma (Kaggle, 2023), la cual se utiliza, entre otros usos, para colocar y extraer set de datos de distintas áreas del conocimiento. El conjunto de datos consta de 68.778 registros correspondientes a las mediciones, con resolución de quince minutos, de las variables eléctricas y climáticas de la planta solar fotovoltaica. Las mediciones se llevaron a cabo entre el 15 de mayo del año 2020 hasta el 17 de junio del mismo año. Las variables eléctricas son: identificación del inversor en el cual se hace la medición (SOURCE_KEY), potencia en corriente continua generada (DC_POWER), y potencia en corriente alterna generada (AC_POWER). Las variables climáticas son: temperatura ambiente (AMB_TEMP), temperatura del módulo (MOD_TEMP), e irradiación solar (IRRADIATION). Adicionalmente, se tiene la variable de fecha y hora (DATE_TIME).

Preparación de los datos

Se inspeccionaron los datos con el fin de verificar la no existencia de datos faltantes y datos duplicados. Ahora bien, la planta consta de 22 inversores, y cada uno de ellos llegan las señales de un número indeterminado de módulos solares, por lo que se procedió a agrupar los datos de esos 22 inversores. Luego de hacer este agrupamiento de datos, quedaron 3.158 registros de valores de las variables mencionadas previamente, con resolución de 15 minutos. Posteriormente, haciendo uso de la columna de datos de fecha y hora, se procedió a crear columnas para la fecha, la hora, y los minutos. Asimismo, se utiliza la columna de potencia en corriente alterna para crear la columna de energía eléctrica en corriente alterna (AC_ENERGY). Finalmente, se crean los datos horarios, agrupando los cuatro datos con resolución de 15 minutos dentro de cada hora, para obtener 796 registros de datos horarios.

III. RESULTADOS Y DISCUSIÓN

En esta sección se presenta el análisis exploratorio de los datos utilizando la totalidad del set de datos de energía eléctrica. De igual manera, se presenta la modelación de los datos, para la que se utilizaron los valores desde el 15 de mayo hasta el 13 de junio del año 2020, para la creación y evaluación de los modelos. Mientras que los datos desde el 14 de junio hasta el 17 de junio, del mismo año, se utilizan para desarrollar los pronósticos de la producción de energía eléctrica de la planta.

Análisis exploratorio de los datos

Para los modelos de regresión que se utilizan en esta investigación se tiene como variable objetivo a la energía eléctrica generada por la planta solar fotovoltaica, por lo que es importante determinar su correlación con las otras variables del set de datos. Se utiliza el método Pearson, el cual es útil cuando los datos se distribuyen normalmente, asimismo también se utilizan los métodos de Spearman y Kendall los cuales se pueden utilizar cuando no se requiere que los datos sigan alguna distribución específica (Navlani, Fandango, & Idris, 2021). Dado que, para los tres métodos se obtienen resultados similares, en la Figura 1 se presenta la matriz de

correlación de las variables correspondientes, al utilizar el método de Pearson.

De la Figura 1 se puede apreciar que la energía eléctrica generada tiene una alta correlación positiva con la temperatura ambiente, la temperatura del módulo, y la irradiación solar. Según Ratner (2017) una correlación mayor a 0,7 se considera como una fuerte relación positiva entre las variables. Por lo anterior, esas tres últimas variables se pueden utilizar como explicativas para los modelos de regresión.

Por otra parte, es de interés mostrar el comportamiento de la energía eléctrica AC en el tiempo. En primer lugar, en la Figura 2 se presentan los 3.158 valores de energía para cada uno de los períodos de 15 minutos. Se puede ver que los datos tienen un comportamiento estable en el sentido que parten de cero, suben hasta un valor máximo y luego caen a cero nuevamente, y este ciclo se repite durante los 34 días del período de estudio, sin ningún tipo de tendencia ascendente o descendente.

Figura 1

Matriz de correlación de las variables.

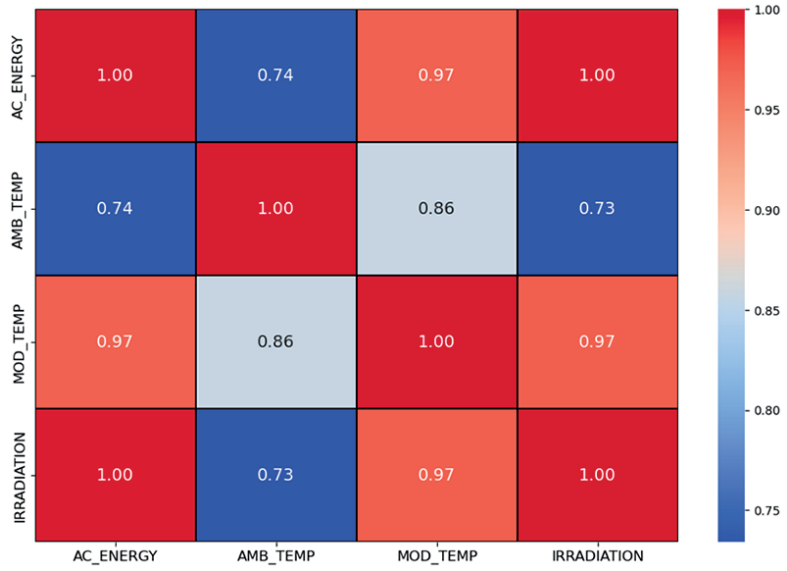
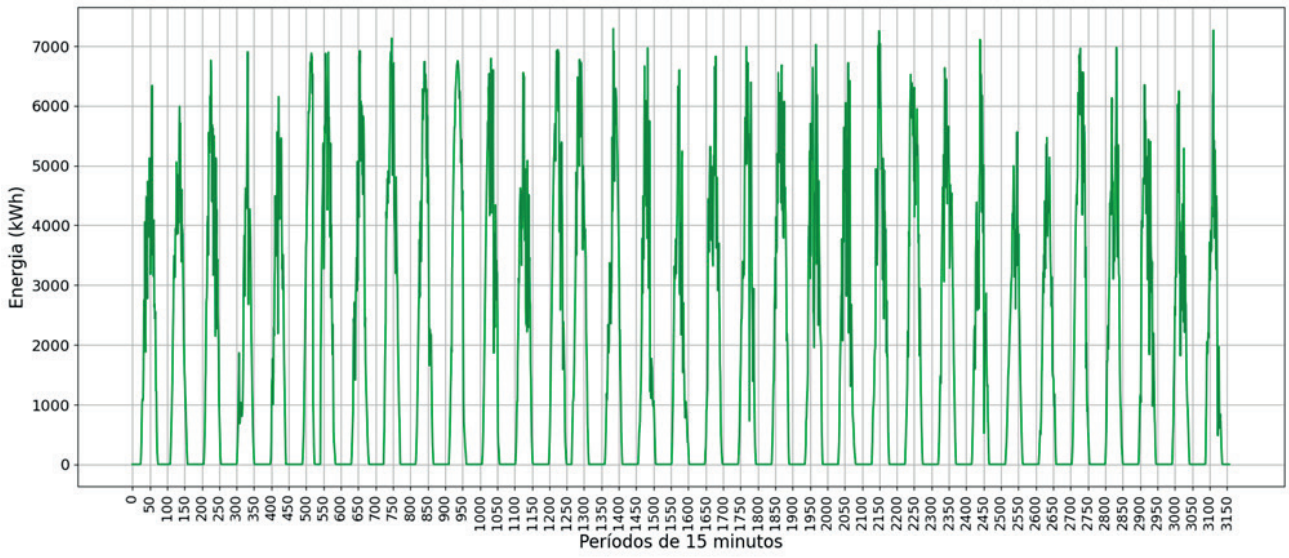


Figura 2

Valores de la energía AC generada.



Seguidamente, en la Figura 3 se presentan los valores promedio de energía AC generada para cada uno de esos períodos de 15 minutos. Se puede observar cómo en promedio comienza la generación de energía a partir de las 06:15 am, se llega a un valor máximo promedio alrededor del mediodía, y cae a producción nula nuevamente alrededor de las 06:30 pm. Es importante aclarar que para cada punto de la curva se presenta el intervalo de confianza al 5% de significancia estadística, en la que la línea oscura es la curva de los valores medios, mientras que la parte sombreada representa a las bandas del intervalo de confianza. Se puede decir entonces, que a medida que aumenta la producción de energía, también hay mayor variabilidad en los datos.

Luego, se suman los valores de los registros de energía de cada 15 minutos para formar los valores horarios para todo el período. En la Figura 4 se presentan los 796 valores horarios de energía eléctrica

Figura 3

Valores promedio de la energía AC generada.

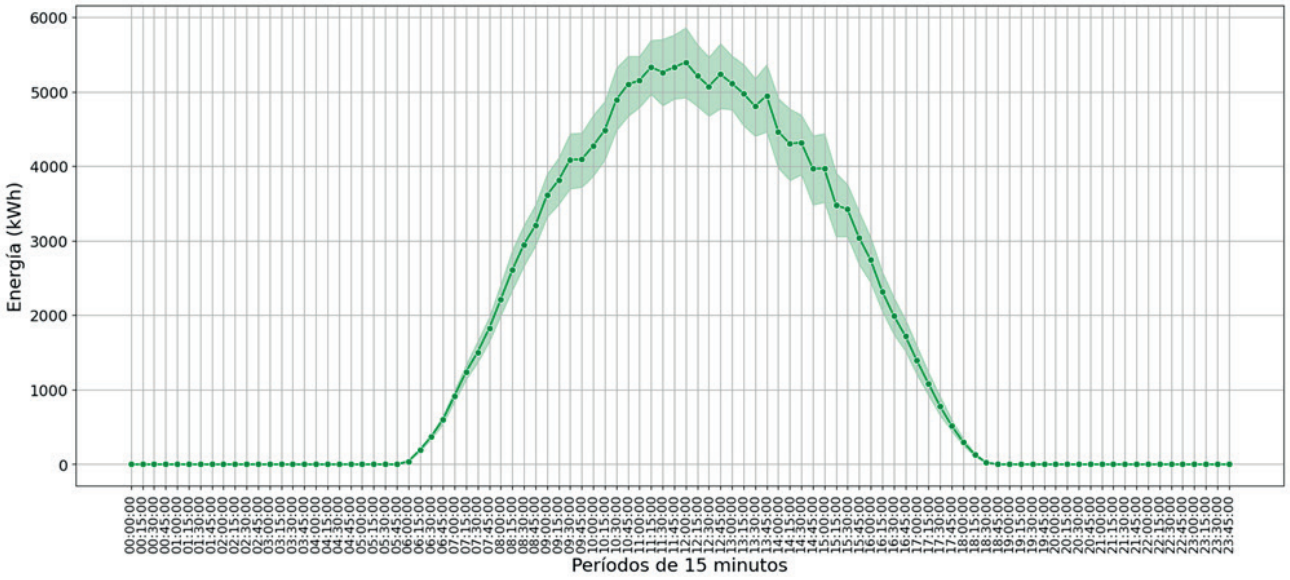
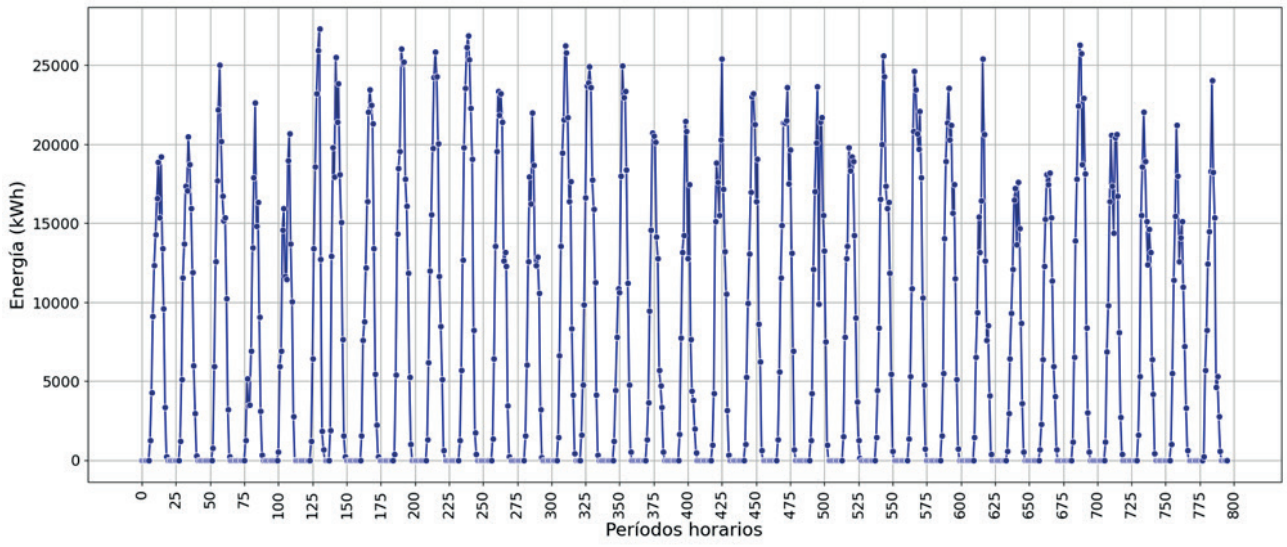


Figura 4

Valores horarios de la energía AC.



AC. Se puede ver que los datos presentan un comportamiento similar a los de la Figura 2, pues no se visualiza ningún tipo de tendencia, pero sí estacionalidad diaria.

Los datos horarios se promediaron por hora, y su gráfica se presenta en la Figura 5, en la que realmente se presenta el intervalo de confianza al 5% de significancia estadística. Como era de esperarse, la planta produce energía a partir de alrededor de las 6 am hasta aproximadamente las 6 pm, teniendo su valor máximo promedio alrededor de las 11 am. Se repite la presencia de mayor variabilidad de los datos alrededor de los valores máximos de energía.

Modelación de los datos

A continuación, se aplican los algoritmos necesarios para obtener los modelos de pronóstico. Para los datos con resolución temporal de quince minutos, se obtienen modelos a través de una red neuronal artificial, y con el algoritmo de K vecinos más cercanos. Para los datos con resolución horaria, se obtiene un modelo a través de una red neuronal artificial, y un modelo del análisis de series de tiempo.

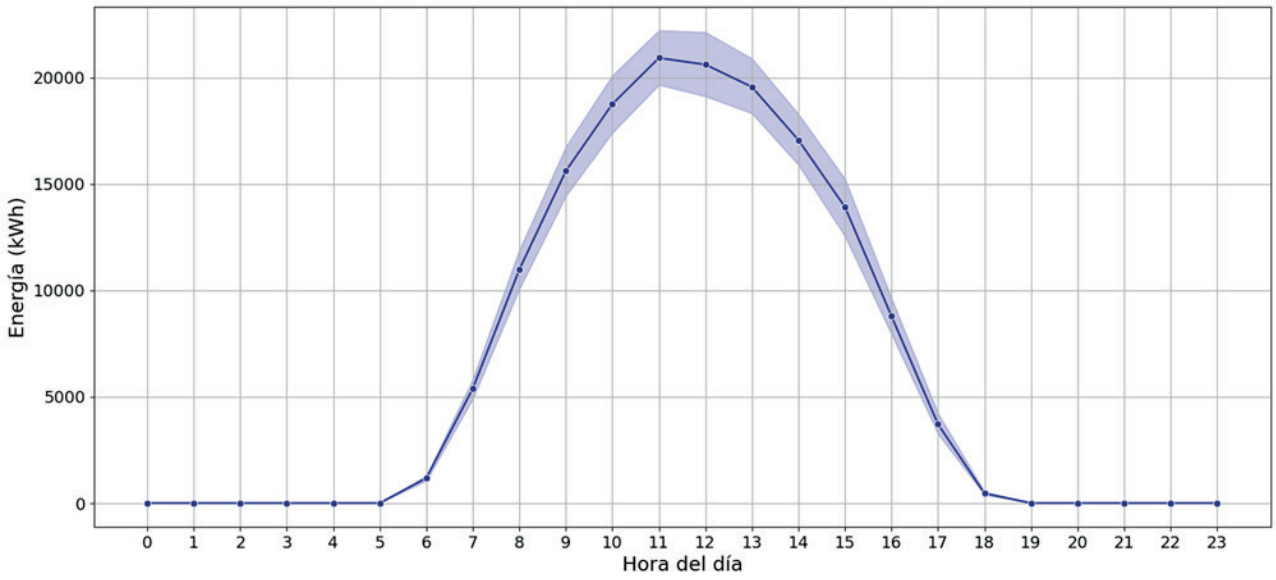
Modelo de Red Neuronal Artificial (RNA) – Resolución 15 minutos

Para el desarrollo del modelo se tienen 1.298 registros, correspondientes a los intervalos de 15 minutos, de las horas con irradiación solar diferente de cero. Como variables explicativas se tienen a la irradiación solar, la temperatura ambiente, y la temperatura del módulo solar. Como variable objetivo se tiene a la energía eléctrica AC producida por la planta. Los datos se dividen en dos partes. Una parte correspondiente al 70% de los datos, para el entrenamiento del modelo, y una segunda parte para la prueba y validación del modelo, correspondiente al restante 30% de los datos. Ese 30% se divide a su vez en dos partes iguales, una para la prueba del modelo, y la otra para la validación del modelo.

Se trabaja con un modelo conocido como “perceptrón multicapa”, que está conformado por varias capas: una capa de entrada, una capa de salida, y un grupo de capas ocultas ubicadas entre la entrada y la salida (Kapoor, Gulli, & Pal, 2022). Se consideró una RNA de tres capas, todas las cuales son del tipo “densa”, lo que significa que la capa respectiva conecta todas sus neuronas con todas las neuronas de la capa previa (Moolayil, 2019). Se tiene la capa de entrada con 256 neuronas, a la cual se conectan las entradas del sistema, una capa oculta también de 256 neuronas, y al final una capa de salida que, al ser de regresión, sólo tiene una neurona. Tanto la capa de entrada como la oculta tienen una función de activación tipo ReLU, con el fin de limitar los valores de sus respectivas salidas, a sólo valores positivos, mientras que la

Figura 5

Valores promedio horarios de la energía AC.



capa de salida tiene una función de activación tipo "lineal" con el fin de no limitar el pronóstico. Dado que se trabaja un caso de regresión, se utilizan el MAE y el MSE como funciones de pérdida, las cuales según Chollet (2018) son útiles para controlar la desviación del pronóstico cuando es comparado con el valor deseado. Cuando la desviación es significativa, se realimenta la salida hacia la entrada a través del optimizador, la que como indica Cholett, actualiza los pesos de las entradas y por consiguiente se repite el ciclo. En esta investigación se hace uso del optimizador de propagación de raíz cuadrática media (RMSProp). Para la historia de la red se trabaja con un total de 400 épocas. Las épocas son el número de iteraciones del proceso de entrenamiento de la red a través del set de datos (Brownlee, 2016).

Se utilizan varias métricas de desempeño para evaluar el modelo. El coeficiente de determinación R^2, que mide la calidad de ajuste del modelo, y la que según lo planteado por Walpole, Myers, Myers, & Ye (2012, p. 407) "indica la proporción de la variabilidad que es explicada por el modelo obtenido". Cuando toma un valor cercano a "1" implica que el ajuste del modelo es casi perfecto, pero cuando toma un valor cercano a "0" es indicativo de un ajuste deficiente del modelo. Las otras métricas son el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE), y el error porcentual absoluto medio (MAPE), los cuáles de acuerdo con (Makridakis, Wheelwright, & Hyndman, 1997) son medidas estadísticas estándar para evaluar modelos de pronóstico. Para el análisis de los residuos se utiliza el test Shapiro-Wilk, el que de acuerdo con Arnastauskaite, Ruzgas, & Braženas (2021) tiene como hipótesis nula que los datos se distribuyen normalmente. El estadístico de prueba varía entre 0 y 1, y cuando está cercano a 1 es un indicativo de que los datos se distribuyen normalmente.

Los resultados obtenidos en la fase de evaluación del modelo RNA se presentan en la Tabla 1. Se puede observar que el modelo tiene un buen coeficiente de determinación de casi el 99%, el RMSE representa alrededor del 5% del valor medio de los datos de prueba de la energía, el MAE representa alrededor del 4% de dicho valor medio, y el MAPE fue de 5,27%. Del análisis de los residuos se puede ver que el estadístico de prueba es cercano a la unidad y el p-valor es mayor al 5%, por lo que se puede decir que los residuos están normalmente distribuidos.

Tabla 1

Métricas del modelo RNA.

Indicador	Resultado
R^2	0,988
RMSE (kWh)	194,19
MAE (kWh)	154,75
MAPE(%)	5,27
Test Shapiro-Wilk a los residuos	
Estadístico de prueba	0,995
p-valor	0,867

Modelo de regresión K-NN – Resolución 15 minutos

El algoritmo de K vecinos más cercanos (K-NN) es una técnica no paramétrica que se puede utilizar tanto para clasificación como para regresión, pero el principio básico es el mismo en ambos enfoques, considerar los elementos que son similares entre sí. Para el caso de la regresión, el valor a pronosticar se estima a través de un estadístico, que usualmente es el valor medio, que se obtiene al agrupar o resumir las características de los vecinos más cercanos (Fenner, 2020). Para este modelo, se trabajó con las mismas variables explicativas y objetivo que las utilizadas para el modelo RNA, así como con el mismo set de datos. De igual forma, se hizo la división 70% - 30% para los datos de entrenamiento y los datos de prueba.

Luego de aplicar el algoritmo de regresión K-NN, utilizando el valor óptimo 9 para el número de vecinos más cercanos, se obtuvieron los resultados de la evaluación del modelo con los datos de prueba, los cuales se presentan en la Tabla 2. Se puede ver que casi el 98% de la variabilidad del modelo es explicada por las variables exógenas, el RMSE representa el 7% del valor medio de los datos de energía eléctrica del set de prueba, el MAE corresponde al 5,44% del valor medio recién mencionado, y el MAPE es del 6,72%. En cuanto al análisis de los residuos, se podría decir que se distribuyen normalmente, puesto que el estadístico de la prueba de Shapiro-Wilk es cercano a la unidad y el p-valor es mayor al 5% de significancia estadística.

Tabla 2

Indicadores del modelo K-NN

Indicador	Resultado
R^2	0,978
RMSE (kWh)	263,08
MAE (kWh)	203,72
MAPE(%)	6,72
Test Shapiro-Wilk a los residuos	
Estadístico de prueba	0,995
p-valor	0,277

Comparación de los pronósticos con resolución de 15 minutos

Luego de la evaluación de los modelos, se procedió a realizar el pronóstico para los días 14, 15, 16, y 17 de junio, con resolución de 15 minutos, y considerando las horas con irradiación solar diferente de cero. Las métricas de desempeño para cada uno de los modelos se presentan en la Tabla 3. Dado que las métricas utilizadas son del tipo "mientras menos mejor", se puede decir que el modelo RNA presenta mejor desempeño que el modelo K-NN, pues presenta los menores valores de cada una de las métricas de desempeño consideradas.

Tabla 3

Métricas de los pronósticos – Resolución 15 min.

Indicador	RNA	K-NN
RMSE (kWh)	172,55	218,50
MAE (kWh)	113,63	160,36
MAPE(%)	3,81	6,12

Sin embargo, ambos modelos tienen métricas aceptables y cumplen con los supuestos estadísticos. En ese sentido, en la Figura 6 se presentan las gráficas de los pronósticos obtenidos, más los valores reales de energía eléctrica producida, en el período considerado. Se puede ver que los pronósticos de ambos modelos siguen casi perfectamente a los valores reales de energía, lo que confirma que ambos modelos tienen un buen desempeño.

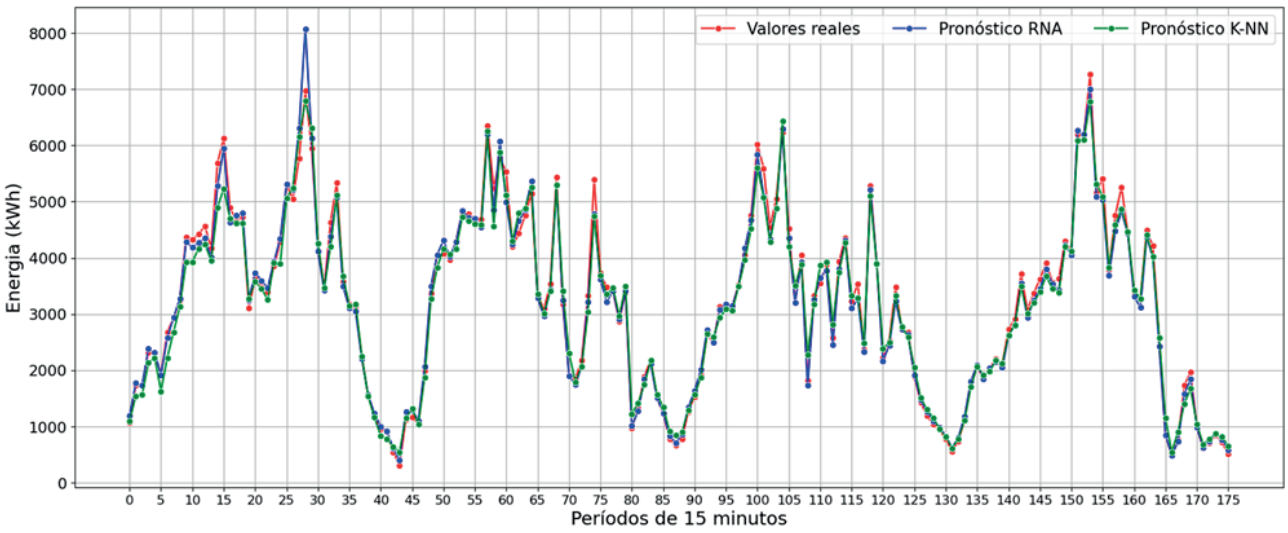
Modelo de Red Neuronal Artificial – Resolución horaria

En lo que respecta a este caso, se tienen 356 registros, correspondientes a los intervalos horarios de las horas con irradiación solar diferente de cero. Las variables irradiación solar, temperatura ambiente, y temperatura del módulo solar, se seleccionan como variables explicativas del modelo. Como variable objetivo se tiene a la energía eléctrica AC producida por la planta.

Al igual que en el modelo RNA anterior, los datos se dividen en dos partes: 70% para el entrenamiento del modelo, y 30% de los datos para la prueba y validación del modelo, por lo que este 30% se divide a su vez en dos partes iguales, una para la prueba del modelo, y la otra para la validación del modelo. Asimismo, la RNA es de tres capas densas, una de entrada, una oculta, y una de salida. La capa de entrada tiene 256 neuronas, a la cual se conectan las entradas del sistema, la capa oculta de 256 neuronas, y al final la capa de salida con una neurona. Tanto la capa de entrada como la oculta tienen una función de activación tipo ReLU, mientras que la capa de salida tiene una función de activación tipo "lineal". La red utiliza un optimizador tipo RMSProp, y las funciones de pérdida son el error absoluto medio y el error cuadrático medio, utilizados usualmente cuando se tiene un caso de regresión. Para la historia de la red se trabaja con un total de 1000 épocas.

Figura 6

Valores reales más pronósticos – resolución 15 minutos.



Luego de evaluar el modelo, los resultados se presentan en la Tabla 4. Se puede observar que el modelo tiene un buen coeficiente de determinación de casi el 99%, el RMSE representa alrededor del 3,89% del valor medio de los datos de prueba de la energía, el MAE representa alrededor del 3,16% de dicho valor medio, y el MAPE fue de 5,03%. Del análisis de los residuos se puede ver que el estadístico de prueba es cercano a la unidad y el p-valor es mayor al 5%, por lo que se puede decir que los residuos están normalmente distribuidos.

Tabla 4

Indicadores del modelo RNA

Indicador	Resultado
R^2	0,994
RMSE (kWh)	515,11
MAE (kWh)	418,61
MAPE(%)	5,03
Test Shapiro-Wilk a los residuos	
Estadístico de prueba	0,962
p-valor	0,09

Modelo ARIMA – Resolución horaria

A la serie de tiempo de los valores de energía eléctrica generada por la planta solar fotovoltaica, compuesta por los 356 registros correspondientes a irradiación solar diferente de cero entre el 15 de mayo del año 2020 hasta el 13 de junio del año 2020, se le aplicó la metodología Box-Jenkins para encontrar el mejor modelo para realizar el pronóstico. Los datos horarios entre el 14 de junio del 2020 y el 17 de junio del 2020 se utilizaron para comparar con el pronóstico obtenido.

El primer paso consiste en verificar la estacionariedad de los datos, aplicando la prueba de Dickey-Fuller ampliada con el fin de detectar posibles raíces unitarias. Esta prueba tiene como hipótesis nula que la serie de tiempo tiene al menos una raíz unitaria, lo que implicaría que no sería estacionaria, tal como lo plantean Afriyie, Twumasi-Ankrah, Gyamfi, Arthur, & Pels (2020). Se aplica la prueba a la serie en nivel, es decir, sin diferenciarla, y a la serie diferenciada estacional. Los resultados se presentan en la Tabla 5.

Tabla 5

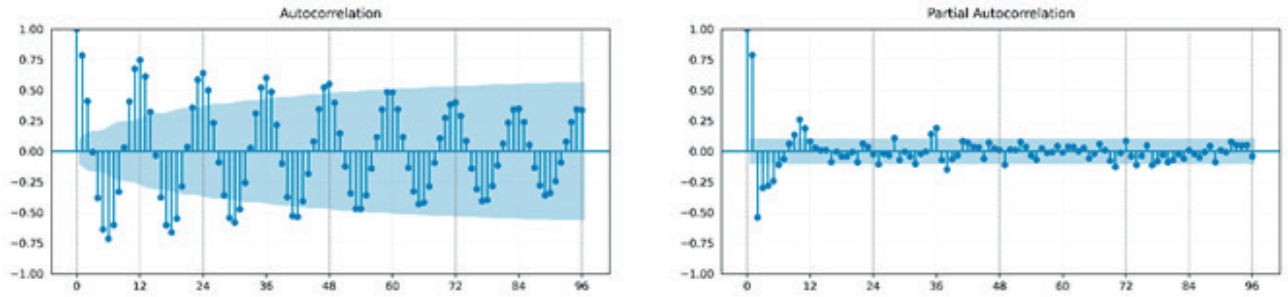
Resultados de prueba de Dickey-Fuller ampliada.

Test Dickey-Fuller ampliado	Nivel	Diferenciada estacional
Estadístico de prueba	-3.579964	-5.006641
p-valor	0.006156	0.000022
Valor crítico (1%)	-3.449503	-3.450445
Valor crítico (5%)	-2.869979	-2.870392
Valor crítico (10%)	-2.571266	-2.571486

De la Tabla 5 se puede ver que para la serie de nivel el p-valor es menor al 5% de significancia estadística, y además el estadístico de prueba es menor a cada uno de los valores críticos, por lo que se puede decir que es estacionaria. El mismo análisis aplica para la serie diferenciada estacional. Seguidamente, se grafican las funciones de autocorrelación y autocorrelación parcial, para así determinar un modelo inicial, las cuales se presentan en la Figura 7. Se puede ver que la serie en nivel tiene varios valores autorregresivos, así como componentes de promedio móviles. Se confirma una marcada estacionalidad cada 12 períodos.

A partir de los resultados mostrados en la Figura 7, se prueban de manera iterativa varios modelos, y se selecciona el que minimiza el AIC, pero que además cumpla con los supuestos estadísticos para los residuos. El modelo seleccionado es ARIMA(4,0,2)(1,2,0)[12]. Es decir, se tiene un modelo con cuatro componentes autorregresivas y dos de promedio

Figura 7



Funciones de autocorrelación y autocorrelación parcial.

móvil en la parte de nivel, y una componente autorregresiva en la parte estacional, además de considerar dos raíces unitarias estacionales.

En cuanto al análisis de los residuos, se aplica la prueba de Ljung-Box, pues según lo planteado por Mahan, Chorn, & Georgopoulos (2015), esta prueba es útil para probar la autocorrelación de los residuos, considerando más de un rezago. La hipótesis nula de la prueba de Ljung-Box indica que las autocorrelaciones entre los residuos son nulas. Además, para verificar la normalidad de los residuos se aplica la prueba de Jarque-Bera cuya hipótesis nula dice que la asimetría y el exceso de curtosis de los datos son nulos, es decir se distribuyen normalmente. De acuerdo con Ahmad & Khan Sherwani (2015), la prueba de Jarque-Bera es muy popular para probar normalidad, y tienen buen desempeño para tamaños de muestras similares a las de esta investigación. El p-valor obtenido al aplicar la prueba Ljung-Box fue de 37% superior al 5% de significancia estadística por lo que no se rechaza la hipótesis nula de que las autocorrelaciones entre los residuos son nulas, es decir, se puede decir que los residuos son independientes. En lo que respecta a la prueba de Jarque-Bera, se obtuvo un p-valor de 0,32 por lo que no se rechaza la hipótesis nula de esta prueba, y entonces se puede decir que los residuos están normalmente distribuidos.

Adicionalmente, luego de evaluar el modelo, en la Tabla 6 se muestran las métricas de desempeño obtenidas. Aunque el modelo cumple con los supuestos estadísticos, se puede ver que los valores de las métricas de desempeño son mucho mayores que los obtenidos con el modelo RNA anterior, destacando un error porcentual del 47,17%.

Tabla 6

Métricas del modelo ARIMA

Indicador	Resultado
RMSE (kWh)	5327,94
MAE (kWh)	4013,55
MAPE(%)	47,17

Comparación de los pronósticos con resolución horaria

Luego de obtener y evaluar los modelos, se procedió a utilizarlos para pronosticar la energía eléctrica para las horas con irradiación solar diferente de cero de los días 14, 15, 16, y 17 de junio del año 2020, y obtener las métricas de desempeño, al comparar los pronósticos con los valores reales de energía. En la Tabla 7 se presentan los resultados obtenidos, en la que se puede ver que el modelo ARIMA tiene métricas muy deficientes, mientras que el modelo RNA obtiene resultados aceptables.

Tabla 7

Métricas de los pronósticos – Datos horarios.

Indicador	RNA	ARIMA
RMSE (kWh)	536,58	14.002,64
MAE (kWh)	387,49	10.842,62
MAPE(%)	4,84	115,37

Posteriormente, se grafican los pronósticos obtenidos, junto con los valores de energía, lo que se muestra en la Figura 8. Se confirma lo obtenido con las métricas, es decir, la curva del pronóstico obtenido con el modelo RNA sigue con buena precisión a la curva de los valores reales. Mientras que la curva del pronóstico del modelo ARIMA, se aleja considerablemente de los valores reales, a medida que se incrementan los períodos del pronóstico. Se puede ver que en las primeras doce horas, correspondientes al día 14, el pronóstico del modelo ARIMA está relativamente cerca de los valores reales, aunque el pronóstico con el modelo RNA sigue siendo mejor.

IV. CONCLUSIONES

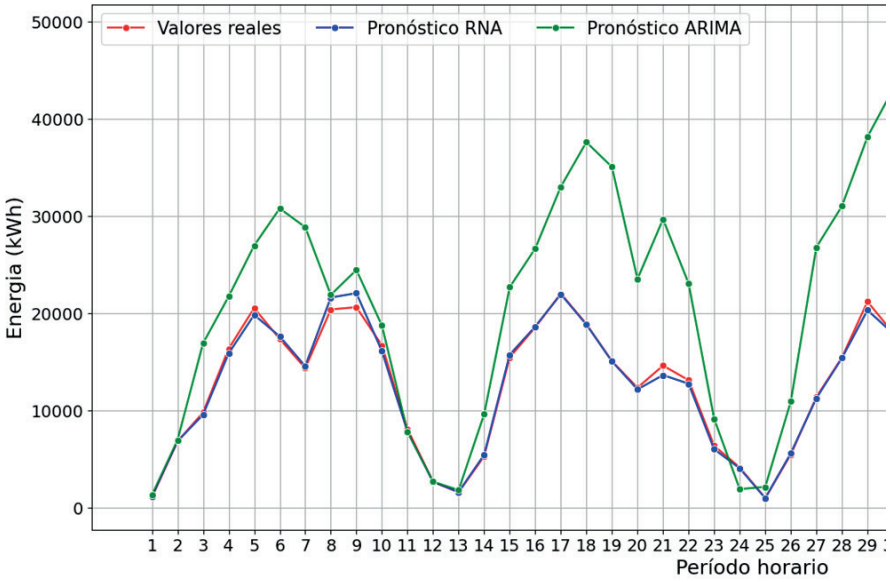
Los datos, tanto con resolución de 15 minutos como con resolución horaria, no presentan tendencia de ningún tipo, pero si presentan estacionalidad diaria. La producción de energía eléctrica de la planta inicia alrededor de las 6 de la mañana y culmina alrededor de las 6 de la tarde, alcanzando su nivel máximo entre las 11 de la mañana y el mediodía.

Los dos modelos de pronóstico para los datos con resolución de quince minutos tuvieron valores aceptables de las métricas de desempeño, tanto en el entrenamiento del modelo como en el pronóstico de la energía eléctrica. Sin embargo, el modelo RNA fue el que alcanzo los mejores valores de las métricas de desempeño, con un MAPE de 5,27% en la fase de entrenamiento del modelo, y un MAPE de 3,81% en la fase de pronóstico. También obtuvo los valores mínimos del MAE y el RMSE, en ambas fases.

En cuando a la resolución horaria, el modelo RNA de nuevo tuvo los mejores valores de las métricas de desempeño tanto en el entrenamiento del modelo como en el pronóstico de la energía eléctrica. En la fase de pronóstico tuvo un MAPE de 4,84%, un RMSE de 536,58 kWh, y un MAE de 387,49 kWh. El modelo ARIMA del análisis de series de tiempo,

Figura 8

Valores reales de energía más pronósticos – Resolución horaria.



no sólo tuvo valores mayores en las métricas, sino que esos valores se consideran inaceptables en un pronóstico, pues el MAPE en el entrenamiento fue de 47,17% y en el pronóstico subió hasta alrededor de 115%.

REFERENCIAS

[1] Afriyie, J., Twumasi-Ankrah, S., Gyamfi, K., Arthur, D., & Pels, W. (2020). Evaluating the Performance of Unit Root Tests in Single Time Series Processes. *Mathematics and Statistics*, 656-664. DOI: 10.13189/ms.2020.080605.

[2] Agencia Internacional de Energía IEA. (2022). *World Energy Outlook 2022.* París: IEA.

[3] Ahmad, F., & Khan Sherwani, R. (2015). Power comparison of various normality tests. *Pakistan Journal of Statistics and Operation Research*, 331-345. DOI: 10.18187/pjsor.v11i3.845.

[4] Arias, M., & Bae, S. (2021). Solar Photovoltaic Power Prediction Using Big Data Tools. *Sustainability*, https://doi.org/10.3390/su132413685.

[5] Arnastauskaite, J., Ruzgas, T., & Braženas, M. (2021). An Exhaustive Power Comparison of Normality Tests. *Mathematics*, https://doi.org/10.3390/math9070788.

[6] Brownlee, J. (2016). *Deep Learning With Python - Develop Deep Learning Models On Theano And TensorFlow Using Keras.* Melbourne, Australia: Machine Learning Mastery.

[7] Chollet, F. (2018). *Deep Learning with Python.* Shelter Island, NY: Manning Publications Co.

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

[9] Comunidad global de energías renovables REN21. (2022). *Global Status Report 2022.* Paris: REN21.

[10] Cryer, J., & Chan, K.-S. (2008). *Time Series Analysis with Applications in r.* New York: Springer Science+Business Media, LLC.

[11] Fan, G.-F., Wei, H.-Z., Chen, M.-Y., & Hong, W.-C. (2022). Photovoltaic Power Generation Forecasting Based on the ARIMA-BPNN-SVR Model. *Global Journal of Energy Technology Research Updates*, 18-38. DOI: https://doi.org/10.15377/2409-5818.2022.09.2.

[12] Fara, L., Diaconu, A., Craciunescu, D., & Fara, S. (2021). Forecasting of Energy Production for Photovoltaic Systems Based on ARIMA and ANN Advanced Models. *International Journal of Photoenergy*, https://doi.org/10.1155/2021/6777488.

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

[14] IRENA. (2020). *Advanced Forecasting of Variable Renewable Power Generation - Innovation Landscape Brief.* Abu Dhabi. https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2020/Jul/IRENA_Advanced_weather_forecasting_2020.pdf.: International Renewable Energy Agency.

[15] Jung, A.-H., Lee, D.-H., Kim, J.-Y., Kim, C., Kim, H.-G., & Lee, Y.-S. (2022). Regional Photovoltaic Power Forecasting Using Vector Autoregression Model in South Korea. *Energies*, https://doi.org/10.3390/en15217853.

[16] Kaggle. (12 de Marzo de 2023). *Your Machine Learning and Data Science Community*. Obtenido de https://www.kaggle.com/

[17] Kapoor, A., Gulli, A., & Pal, S. (2022). *Deep Learning with TensorFlow and Keras.* Birmingham: Packt Publishing Ltd.

[18] Konstantinou, M., Peratikou, S., & Charalambides, A. (2021). Solar Photovoltaic Forecasting of Power Output Using LSTM Networks. *Atmosphere*, https://doi.org/10.3390/atmos12010124.

[19] Larson, D., Nonnenmacher, L., & Coimbra, C. (2016). Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest. *Renewables Energy*, 11-20. http://dx.doi.org/10.1016/j.renene.2016.01.039.

[20] Mahan, M., Chorn, C., & Georgopoulos, A. (2015). White Noise Test: detecting autocorrelation and nonstationarities in long time series after ARIMA modeling. *PROC. OF THE 14th PYTHON IN SCIENCE CONF. (SCIPY 2015)* (págs. 97-104). Scipy2015.

[21] Makridakis, S., Wheelwright, S., & Hyndman, R. (1997). *Manual of Forecasting: Methods and Applications.*

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

[23] Mills, T. (2019). *Applied Time Series Analysis - A Practical Guide to Modeling and Forecasting.* London, United Kingdom: Academic Press - Elsevier.

[24] Moolayil, J. (2019). *Learn Keras for Deep Neural Networks - A Fast-Track Approach to Modern Deep Learning with Python.* Vancouver, BC, Canada: Apress.

[25] National Renewable Energy Laboratory, Sandia National Laboratory, SunSpec Alliance, and the SunShot National Laboratory Multiyear Partnership (SuNLaMP) PV O&M Best Practices Working Group. (2018). *Best Practices for Operation and Maintenance of Photovoltaic and Energy Storage Systems.* NREL/TP-7A40-73822. https://www.nrel.gov/docs/fy19osti/73822.pdf.: NREL.

[26] Navlani, A., Fandango, A., & Idris, I. (2021). *Python Data Analysis.* Birmingham, UK: Packt Publishing Ltd.

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

[28] Rogel-Salazar, J. (2017). *Data Science and Analytics with Python.* Boca Raton, FL: CRC Press Taylor & Francis Group.

[29] Technology Collaboration Programme - IEA. (2022). *Task 13 Reliability and Performance of Photovoltaic Systems - Guidelines for Operation and Maintenance of Photovoltaic Power Plants in Different Climates.* Alzenau, Germany: Ulrike Jahn, VDE Renewables.

[30] Tu, C.-S., Hong, C.-M., & Lin, W.-M. (2021). Short-Term Solar Power Forecasting via General Regression Neural Network with GreyWolf Optimization. *Energies*, https://doi.org/10.3390/en15186624.

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

[32] Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probabilidad y estadística para ingeniería y ciencias. Ciudad de México: Pearson Educación de México, S.A. de C.V.

Fuentes de financiamiento:

Propia.

Conflictos de interés:

El autor declara no tener conflictos de interés.

Contribución de los autores

El autor desarrolló el preprocesamiento de los datos, el análisis exploratorio, así como todos los modelos de pronóstico de la energía eléctrica. Asimismo, utilizó los modelos para realizar los pronósticos de energía eléctrica generada por la planta solar fotovoltaica, y redactó las conclusiones. De igual manera, participó en la redacción y revisión final del artículo.