T E S LA
Revista Científica
ISSN: 2796-9320
Vol. 3 Núm. 2 (Julio – Diciembre 2023), e203

Área: Ciencias Exactas
Artículo de Investigación Original

Predicción de radiación solar en Salinas de Urcuquí: Modelo estadístico matemático

Prediction of Solar Radiation in Urcuquí: Mathematical Statistical Model

Iván Patricio Quinteros Campaña [1,2][0009-0006-5351-789X], Paúl Michael Tafur Escanta [3,4][0000-0002-0760-6350]

[1] Universidad Politécnica Estatal del Carchi, Centro de Posgrado, Calle Antisana s/n y Av. Universitaria, Tulcán, Ecuador
[2] Universidad de las Fuerzas Armadas – ESPE, Departamento de Ciencias Exactas, Av. General Rumiñahui s/n y Ambato, Sangolquí, Ecuador
[3] Universidad Técnica del Norte, Facultad de Ingeniería en Ciencias Agropecuarias y Ambientales, Av. 17 de Julio 5-21 y Gral. José María Córdova, Ibarra, Ecuador
[4] Universidad Politécnica de Madrid, Departamento de Ingeniería Energética, C/José Gutiérrez Abascal, 2, 28006, Madrid, España

[1,2] ivan.quinteros@upec.edu.ec, [3,4] pm.tafur@alumnos.upm.es

CITA EN APA:
Quinteros Campaña, I., & Tafur Escanta, P. (2023). Predicción de radiación solar en Salinas de Urcuquí: Modelo estadístico matemático. *Tesla Revista Científica, 3(*2). https://doi.org/10.55204/trc.v3i2.e203

Recibido: 2023-06-17
Revisado: 2023-06-18 al 2023-07-04
Corregido: 2023-07-06
Aceptado: 2023-10-10
Publicado: 2023-10-14

TESLA
Revista Científica
ISSN: 2796-9320

Resumen. Este trabajo se enfoca en plantear modelos de predicción asociados a la radiación solar incidente sobre la superficie terrestre de la parroquia de Tumbabiro en Urcuquí – Imbabura. La metodología está basada en obtener un modelo óptimo donde se utiliza modelos de predicción que responden a distintos métodos estadísticos. Los datos de los parámetros fundamentales que afectan a la radiación solar se obtienen de la base de datos que proporciona el NREL. Se utiliza modelos AR, ARMA, ARMAX y de Redes Neuronales Recurrentes-RNN, donde se realiza una contrastación modelos predictivos estadísticamente significativos, permitiendo observar que, de los primeros, es significativo un ARMAX(1,1), con un error de 9.87%; pero, los mejores pronósticos se obtuvieron del modelo LSTM con error del 0.18%.
Palabras Clave: Modelo matemático, RNA, radiación solar directa, albedo,

Abstract: This work focuses on proposing prediction models associated with incident solar radiation on the earth's surface of the Tumbabiro parish in Urcuquí - Imbabura. The methodology is based on obtaining an optimal model where prediction models that respond to different statistical methods are used. The data of the fundamental parameters that arise from solar radiation are obtained from the database provided by the NREL. AR, ARMA, ARMAX and Recurrent Neural Networks-RNN models are used, where a comparison of statistically significant predictive models is carried out, allowing us to observe that, of the former, an ARMAX(1,1) is significant, with an error of 9.87. % ; but, the best forecasts were acquired from the LSTM model with an error of 0.18%.
Keywords: Mathematical model, RNA, direct solar radiation, albedo

1. INTRODUCIÓN

En la actualidad, las energías renovables adquieren un papel importante en el mercado energético global, su demanda crece ampliamente a nivel mundial, esto debido a los compromisos de los países en los Objetivos del Desarrollo Sostenible (ODS 2030) para la reducción de emisiones contaminantes por la quema de combustibles fósiles, estos compromisos han llevado a que se deben cumplir varios objetivos a futuro, por ejemplo la reducción de las emisiones de CO_2 a la atmósfera hasta en un 40%, aumentar el porcentaje de reciclaje hasta en un 27 % y que el uso de energías renovables para la generación de energía eléctrica se encuentre entre un 27 y 35 % hasta el 2030 en la Unión Europea (Agencia Internacional de la Energía, 2021). Con el fin de que hasta el año 2100 se haya logrado reducir 2°C la temperatura del planeta.

Las afectaciones al medio ambiente por el uso de recursos no renovables han ido creciendo a lo largo de los últimos 50 años. Las políticas de cambio climático han tratado de resolver esta problemática

dando un punto de vista amplio al uso de energías más limpias para la generación de energía eléctrica (Sayed et al., 2021). Una de las energías que se ha vuelto más rentable en los últimos años es la energía solar en sus dos tipos, energía solar térmica y energía solar fotovoltaica (Tembhare et al., 2022; Tafur-Escanta et al., 2023).

En el Ecuador el uso de estas energías es aún reducido a pesar de tener un alto índice de radiación solar en la mayor parte de su territorio. Según el Atlas del Sector Eléctrico ecuatoriano del 2021 las provincias con mayor potencia nominal de generación de energía eléctrica a través de centrales fotovoltaicas son: El Oro y Loja con 5.99 MW, Imbabura con 4.00 MW y Guayas con 3.99 MW. Las políticas implementadas en el país sobre el cambio de la matriz energética han incentivado a varias empresas nacionales e internacionales a invertir en este tipo de energías limpias. El potencial eléctrico fotovoltaico en la región de la Sierra y Galápagos alcanza un total de 3.8 a 4.8 kWh/kWp diarios y alrededor de 1388 y 1753 kWh/kWp anuales.

Según el Ministerio de Energía, considerando la ubicación geográfica privilegiada de Ecuador, se ha identificado una alta radiación que puede ser aprovechada para la generación de energía eléctrica. La disponibilidad del recurso solar, medido como insolación media global, llega a los 4.575 kilovatios hora por metro cuadrado (Wh/m^2/día). Ese nivel es 40% más alto que el promedio de la región.

El gran potencial energético y eléctrico del país está siendo considerado como un factor importante para el crecimiento económico. De esta manera este trabajo de investigación se ve enfocado a plantear modelos matemáticos de predicción de los parámetros fundamentales (presión, temperatura, índice de radiación solar, albedo, etc.) que intervienen en la cantidad de radiación solar que llega a la superficie terrestre. En este sentido se ha escogido a Salinas de Urcuquí ubicado en la Provincia de Imbabura como lugar de análisis para el obtener el modelo matemático debido a su gran potencial energético fotovoltaico que ya cuenta con una primera instalación fotovoltaica con capacidad de 3 MW construida por Gran Solar S.A. La irradiación promedio en el área de influencia es de 5,1 kWh/m^2/día.

Entre las metodologías para la predicción de la radiación solar se tienen aquellas que consideran modelos de series de tiempo de la estadística tradicional, esto es, modelos AR (autorregresivos, por sus siglas en inglés), MA (medias móviles, por sus siglas en inglés), $ARMA$ (modelo autorregresivo de media móvil, por sus siglas en inglés) y el $ARIMA$ (modelo autorregresivo integrado de media móvil, por sus siglas en inglés) y, las provistas por la inteligencia artificial que permiten generar algoritmos capaces de resolver un problema o tomar decisiones a partir de un conjunto de datos conocido (Mazorra, 2016).

En el caso de Clavijo et al. (2019), realizaron un trabajo que permitió la comparación entre los métodos estadísticos tradicionales y el desarrollado usando métodos de la inteligencia artificial y, obtuvieron como resultado que, un mejor desempeño se obtuvo de las Redes Neuronales, que a la hora de estimar el recurso energético cuenta con un error porcentual promedio mucho más bajo (4,2965%) que el mostrado por el método $ARMA$ (15,4560%).

Rangel (2018), manifiesta que, al analizar los resultados de predicción de las variables

meteorológicas del sitio de interés mediante la implementación de técnicas estadísticas y métodos de análisis y transformación de series de tiempo, llegó a la conclusión de que los modelos de redes neuronales son superiores a las técnicas tradicionales $ARIMA$.

Los resultados de la investigación realizada por Villarreal (2020), permiten concluir que, el error en el pronóstico evaluado a través del error medio absoluto porcentual (MAE) fue, en todos los emplazamientos de donde se obtuvieron los datos, de 14% en promedio para predicciones durante tiempo seco y de 29% para predicciones en temporada de lluvias con el modelo de series de tiempo $SARIMA$, mientras que, para el modelo de redes neuronales autorregresivo $ARNN$ se obtuvo un desempeño similar de 19% para temporada seca y de 31,3% para temporada de lluvias.

Mazorra (2016), establece que, el modelo $ARMA$ se mostró más eficaz a la hora de predecir la radiación que el modelo AR. El modelo $ARMA$ consigue mejores resultados de predicción utilizando únicamente dos entradas de datos pasados de radiación solar, mientras que el modelo AR necesitaba de hasta 11 entradas para conseguir resultados óptimos.

En el caso de Ecuador, los trabajos realizados se enfocan únicamente en la propuesta de métodos basados en la inteligencia artificial, uno de ellos Carrillo (2022), que usa un modelo de red neuronal artificial multicapa, señala que, los resultados mostrados para el pronóstico de radiación solar de una estación meteorológica, son muy aceptables ya que tienen un gran desempeño en la red neuronal en cuanto a la estimación de radiación solar. Lalaleo (2021), hizo una comparación de los días en los que se realizó la predicción con los medidos a través del valor medio, verificando que el valor del error del testeo es de 22.8% y, el valor promedio de los datos de los siete días predichos es de 20,9% verificando que los valores son cercanos al entrenamiento de la red.

Finalmente, el principal objetivo de este trabajo es desarrollar un modelo estadístico-matemático mediante la obtención una solución óptima del método estadístico. El análisis de este estudio nos permite realizar una comparación entre los distintos modelos estadísticos utilizados.

Este trabajo se encuentra dividido en secciones de la siguiente manera, en la Sección 1 se encuentra la Introducción, la Sección 2 habla de la metodología que se ha seguido para obtener el modelo estadístico matemático de predicción. En la sección 3 se presentan los resultados y la discusión de los mismos; y finalmente en la Sección 4 se encuentran las conclusiones más importantes de este trabajo.

2. METODOLOGÍA

Para la obtención de una solución óptima que se adapte a los datos de la Irradiación Solar, se utilizarán modelos de predicción que responden a distintos métodos estadísticos que inicialmente partirán del análisis de los modelos Autorregresivos (AR), Autorregresivos de Medias Móviles ($ARMA$); luego, se integra la estacionalidad y variables exógenas con los métodos Autorregresivos Integrados de Medias Móviles Estacional ($SARIMA$) y Autorregresivos Integrados de Medias Móviles Estacional con Variables Explicativas ($SARIMAX$); finalmente se tendrá en cuenta el uso de Redes Neuronales Artificiales (RNA), más específicamente de una extensión de las Redes Neuronales Recurrentes (RNN), denominada *Memoria*

a Largo-Corto Plazo (*LSTM*, *Long-Short Term Memory*).

Este estudio utiliza datos históricos máximos diarios de Irradiancia Solar proporcionados por el *National Renewable Energy Labratory* (*NREL*, en español, *Laboratorio Nacional de Energías Renovables*) para Salinas de Urcuquí, desde el 1 de enero de 1998 hasta el 31 de diciembre de 2020. El *NREL* es ampliamente reconocido por su enfoque riguroso en la recopilación y verificación de datos, así como a su alta calidad y actualización constante de los mismos.

El análisis se enfoca en verificar si los modelos de predicción de irradiancia solar mejoran partiendo de diferentes conjuntos de datos, puesto que dichas predicciones requieren para el entrenamiento de los modelos, una cantidad suficiente de datos históricos. De esta manera el trabajo propuesto se enfoca en evaluar la bondad de cada modelo usando, en primera instancia, los datos de todo el período a partir del 1 de enero de 1998, luego, para diez y cinco años y, finalmente para los últimos dos años, es decir, con los 731 datos a partir del 1 enero de 2019 hasta el 31 de diciembre de 2020, porque este último período es el que mejores resultados arroja para los modelos formulados y con el cual se trabaja en esta propuesta de investigación.

Se considera también que, para el período señalado previamente, el porcentaje óptimo de datos históricos para realizar el entrenamiento es del orden del 67.5%, los demás datos se utilizaron para el conjunto de prueba, con el cual se examina la bondad de ajuste de los modelos obtenidos. Para evaluar y comparar el desempeño de los modelos se realiza el uso de dos métricas ampliamente utilizadas en investigaciones previas (Cutiño et al., 2010; López-García et al., 2022; Pinzón, 2020), la *Raíz del Error Cuadrático Medio* (*RMSE*, *Root Mean Square Error*), el *Error Medio Absoluto* (*MAE*, *Mean Absolute Error*) y el *Error Porcentual Medio Absoluto* (*MAPE*), dados por las expresiones (1) a (3).

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\hat{I}_k - I_k\right)^2} \tag{1}$$

$$MAE = \frac{1}{n}\left|\sum_{k=1}^{n}\left(\hat{I}_k - I_k\right)\right| \tag{2}$$

$$MAPE = \frac{1}{n}\left|\frac{\sum_{k=1}^{n}\left(\hat{I}_k - I_k\right)}{I_k}\right| * 100 \tag{3}$$

Donde, n representa el número de datos históricos del conjunto de prueba, $\hat{I}_k$ es el valor calculado por el modelo de predicción e I_k, el valor observado de Irradiancia Solar, para el día k.

Los resultados para estas métricas, se obtuvieron comparando las predicciones con los datos históricos de prueba.

El procedimiento a seguir para la elección del modelo de predicción que mejor se ajuste a los datos observados sugerido por Capa (2022), inicia con un análisis exploratorio de la serie temporal para comprender sus características y patrones, para después, aplicar de manera general, los siguientes pasos:

1. Escoger el modelo;
2. Dividir los datos históricos del período seleccionado, en datos de *entrenamiento* y de *prueba*, a fin de evaluar el rendimiento de los modelos en datos no vistos (Collantes et al., 2019) y, verificar la capacidad de generalización y evitar el sobreajuste.

3. Ajustar el modelo a los datos de *entrenamiento*;
4. Evaluar el modelo en los datos de prueba;
5. Predecir los datos futuros.

Una consideración esencial en el análisis exploratorio, es la condición de estacionariedad de la serie de tiempo, que se usa con el propósito de realizar análisis robustos y obtener resultados confiables. La estacionariedad de una serie temporal es un concepto fundamental en el análisis de datos y modelización de series de tiempo. Castro et al. (2019), mencionan que una serie estacionaria exhibe propiedades estadísticas que se mantienen constantes a lo largo del tiempo, lo que implica que la media, la varianza y la autocorrelación no varían significativamente. Además, la estacionariedad es un requisito común en muchos modelos y técnicas de series temporales, como lo señala Capa (2022), ya que facilita la identificación de patrones y la predicción de futuros valores.

Estadísticamente, para comprobar si hay evidencia significativa de que una serie de tiempo es estacionaria, es decir, para evaluar si los coeficientes de autorregresión son significativos y si existe una correlación entre los valores pasados y presentes de la serie, se utiliza la prueba de *Dickey-Fuller* (*prueba DF*). Una de las principales características de esta prueba, de acuerdo a Hassani (2018), es que su uso permite verificar la hipótesis nula de no estacionariedad en una serie de tiempo, lo cual implica la presencia de raíces unitarias. Esta prueba ha sido ampliamente utilizada en el análisis de series temporales y se ha demostrado su robustez y eficiencia en diferentes contextos (Enders, 2018).

Para el caso particular de este estudio, se empieza con un modelo simple y luego se complejiza introduciendo más retrasos, hasta obtener el modelo más apropiado; sin embargo, el proceso genera nuevos parámetros (coeficientes) a estimar, por lo que, una de las cosas que se debe hacer es comprobar que los nuevos coeficientes sean significativamente diferentes de cero.

Una vez que se tiene varios posibles modelos, se requiere compararlos entre sí, para, usando la inferencia, determinar si al añadir complejidad al modelo, las predicciones de este último son significativamente mejores que las del modelo menos complejo. Con este propósito, se utilizó la *prueba de log-verosimilitud* (*LL*: *Log Likelihood test*), que nos permite escoger el "mejor" modelo como el que tiene un mayor valor de probabilidad en esta prueba y, valiéndonos de ésta, usar la prueba de *razón de log-verosimilitud* (*LLR*: *Log Likelihood ratio*) entre dos *modelos anidados* (*nested models*) o jerarquizados, es decir, aquellos que tienen las mismas variables, para construir otro modelo con un número menor de las mismas variables (Gómez-Mejía, 2020).

Por otro lado, a fin de asegurar que el modelo elegido era adecuado y confiable para hacer predicciones, se hace necesario comprobar la condición de ruido blanco para los residuos, pues eso confirmaría, de acuerdo a Box et al. (2015), la no existencia de patrones sistemáticos o información relevante en ellos que no haya sido capturada por el modelo, es decir, que las observaciones eran independientes entre sí y que no había un efecto de tendencia que no haya sido tomado en cuenta. Para hacerlo, se utiliza, por un lado, la *prueba DF* de estacionariedad para los residuos, a fin de asegurar que

éstos cumplían con la condición de ser ruido blanco y, por otro, la función de autocorrelación de los residuos, en la que se debía comprobar que no hay correlaciones significativas.

3. RESULTADOS Y DISCUSIÓN

Las predicciones en los modelos estadísticos, se realizaron con los datos del valor máximo diario de la serie de Irradiancia Solar Global UV ($GHUVI$, por sus siglas en inglés). Una vez que se hayan realizado los procesos de entrenamiento de cada uno de los modelos, el cálculo y discusión de los errores se llevarán a efecto en términos de la irradiancia medida en W/m^2.

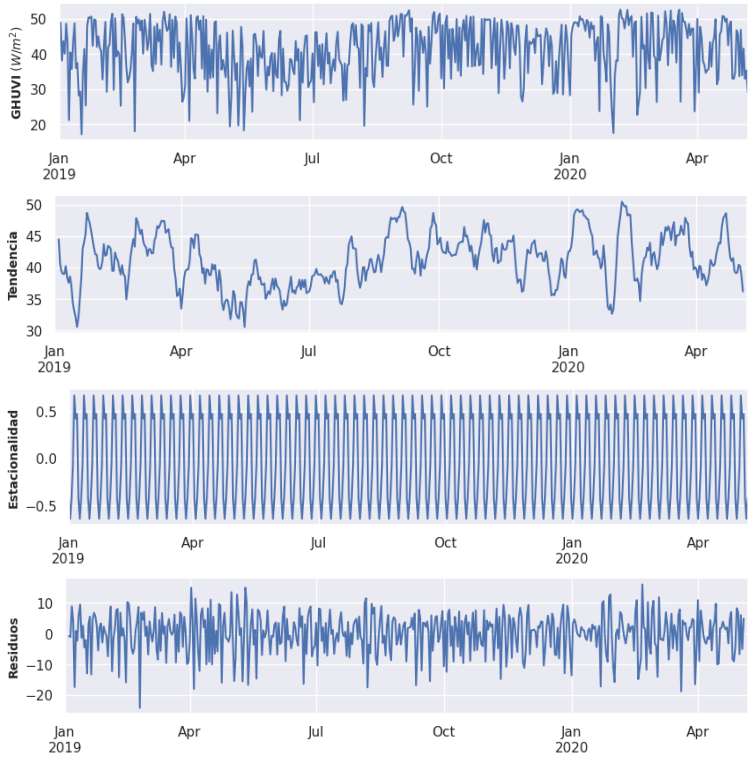
3.1. Análisis descriptivo y exploratorio

En todo el análisis que se sigue para el procesamiento y graficación de los datos y resultados se utiliza la versión gratuita de Colaboratory o "Colab", un producto de Google Research, que permite escribir y ejecutar un código arbitrario de Python en el navegador. Los resultados que se presentan, se obtuvieron usando los datos de entrenamiento. De la evaluación previa de los datos, no hubo necesidad de hacer ningún pretratamiento.

Los datos diarios del valor máximo de la serie de *Irradiancia Solar Global UV* desde el 1 de enero de 2019 hasta el 7 de mayo de 2020, así como, su descomposición en tendencia, estacionalidad y los residuos, se muestran en la figura 1.

Figura 1

Descomposición de la serie de Radiación Solar UV



Fuente: Elaboración autor

La gráfica sugiere que la serie $GHUVI$ cumple con la condición de estacionariedad. En efecto, usando la *prueba DF* para un nivel de confianza del 95%, el $p - valor$ es igual a 6.89×10^{-10}, menor a

0.05, por lo que se rechaza la hipótesis nula que afirma la presencia de raíces unitarias y se acepta que hay evidencia significativa para confirmar la estacionariedad de la serie.

Se observa que no hay una tendencia. Con respecto a la estacionalidad, parece darse cada 28 días y, aplicando la *prueba DF* al 95% de confianza, hay evidencia significativa para afirmar que los residuos, tanto gráfica como estadísticamente, se comportan como ruido blanco (el $p - valor = 1.168 \times 10^{-19}$ es menor que 0.05).

3.2. Modelo autorregresivo de orden p ($AR(p)$)

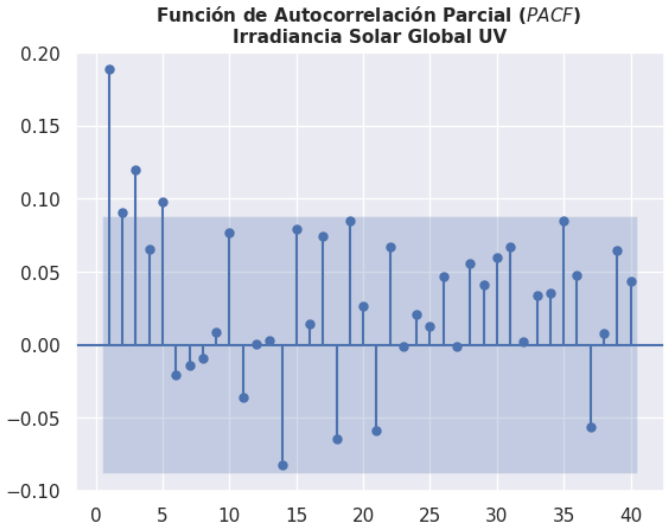
El modelo autorregresivo de orden $\boldsymbol{p}$ para predecir la irradiancia solar está definido por la ecuación (4).

$$I_t = c + \sum_{k=1}^{p} \varphi_k I_{t-k} + \epsilon_t \qquad (4)$$

Donde I_t es el valor de interés, del período actual t de la serie de tiempo, c es la constante, $\{\varphi_k\}_{k=1,2,\ldots,p}$ son los coeficientes que se deben estimar, ϵ_t es el residuo en el período actual, e I_{t-k}, es el valor de la serie en un período anterior. Este modelo se utiliza debido a la condición de estacionariedad de la serie de datos de Irradiancia Solar que se evaluó en este estudio. Para tomar una decisión del orden p para el modelo óptimo a utilizar, se utiliza la gráfica de la Función de Autocorrelación Parcial ($PACF$) para identificar el máximo orden del modelo AR. La figura 2, representa a la $PACF$ de la serie $GHUVI$.

Figura 2

Función de Autocorrelación Parcial de la Irradiancia Solar Global



Fuente: Elaboración autor

Como la $PACF$ establece la correlación entre dos instantes de tiempo y nos interesa un modelo eficiente, solo se considerará aquellos retrasos que tengan un efecto directo y significativo sobre el período presente a un nivel de confianza del 95%. Al examinar la $PACF$ se puede observar que, los retrasos significativos son $p = 1, 3, 5$, los demás se los puede considerar prácticamente nulos.

La Tabla 1 resume los resultados para las pruebas de significancia de la constante c y los coeficientes

$\{\varphi_k\}_{k=1,2,\dots,p}$, así como, la prueba de significancia LL para cada modelo $AR(p)$ y la prueba LLR, razón de log-verosimilitud, que permite la comparación entre modelos con el propósito de elegir un modelo significativamente mejor que otro.

Tabla 1

Pruebas de Significancia de la constante c, los coeficientes $\{\varphi_k\}$ y la prueba de significancia LL

Modelo $AR(p)$	p-valor de c	Significancia coeficientes φ_k (p-valor)						Prueba LL	Prueba LLR entre modelos
		φ_1	φ_2	φ_3	φ_4	φ_5	φ_6		
$AR(1)$	0,000	0,000						-1715,866	
$AR(2)$	0,000	0,000	0,590					-1713,841	0,044
$AR(3)$	0,000	0,000	0,140	0,010				-1710,321	0,008
$AR(4)$	0,000	0,000	0,164	0,023	0,135			-1709,272	0,148
$AR(5)$	0,000	0,001	0,242	0,033	0,254	0,025		-1706,947	0,031
$AR(6)$	0,000	0,001	0,235	0,031	0,248	0,026	0,660	-1706,847	0,665

Fuente: Elaboración autor

Se observa que, en todos los modelos, para un nivel de confianza del 95%, la constante c es significativa; para los coeficientes φ_k, cuando $p > 1$, los coeficientes φ_2, φ_4 y φ_6 en los modelos en los que están presentes, son estadísticamente no significativos. La prueba de log-verosimilitud LL, da evidencia estadística que, aumentar la complejidad en un modelo, determina que las predicciones del modelo más complejo, son significativamente mejor que las del anterior. En el caso de la prueba LLR entre modelos, se observa que cuando se complejiza un modelo al aumentar el número de retardos, el modelo obtenido es significativamente mejor que el anterior, excepto para los modelos $AR(4)$ y $AR(6)$. Al realizar la comparación entre los modelos $AR(5)$ y $AR(1)$, la prueba LLR dio un valor de 0,001 que, estadísticamente nos dice que el modelo $AR(5)$ es significativamente mejor que el modelo $AR(1)$.

3.3. Modelo autorregresivo de medias móviles $ARMA(p, q)$

En el modelo Autorregresivo de Medias Móviles $ARMA(p, q)$, las predicciones se obtienen al combinar los modelos: autorregresivo $AR(p)$ y, el de Medias Móviles $MA(q)$, como una combinación lineal de una cantidad determinada de valores pasados de la serie y de los errores, de acuerdo con la ecuación (5).

$$I_t = c + \sum_{k=1}^{p} \varphi_k I_{t-k} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \qquad (5)$$

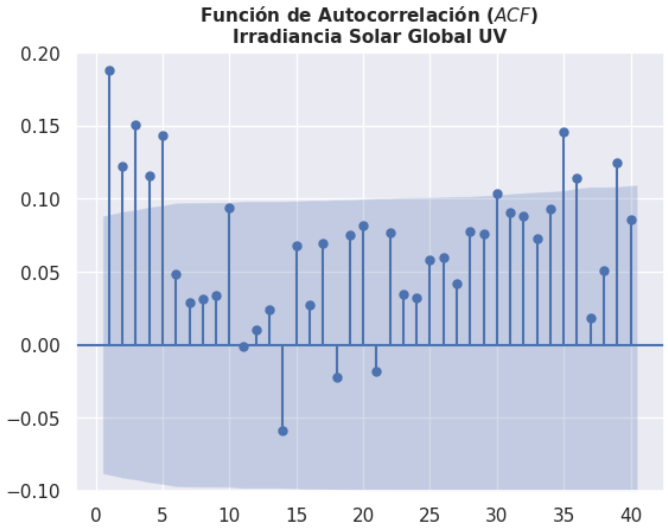
Donde I_t es el valor de interés de la serie de tiempo a predecir en el período actual t, c es la constante, $\{\varphi_k\}_{k=1,2\dots,p}$ son los coeficientes de I_{t-k} que se deben estimar, ϵ_t es el residuo en el período actual, $\{\theta_j\}_{j=1,2,\dots,q}$ son los coeficientes a estimar de la serie de errores ϵ_{t-j}. Este modelo se usó, debido a la condición de estacionariedad de la serie de datos de Irradiancia Solar que se evalúa en este estudio.

El proceso para decidir cuál es el modelo óptimo a utilizar, es decir, para establecer qué valores de p y q determinan el modelo $ARMA$ óptimo, al igual que para el modelo AR, se utiliza la gráfica de la función de autocorrelación parcial $PACF$ para identificar el máximo orden p de la parte autorregresiva AR del modelo (ver Figura 4), como se mencionó anteriormente, el máximo orden para este caso fue para $p = 5$. Mientras que para decidir el orden máximo q de la parte de media móviles MA, se utilizará la gráfica de

la función de autocorrelación ACF, cuya gráfica consta en la Figura 3.

Figura 3

Función de Autocorrelación de Irradiancia Solar Global



Fuente: Elaboración autor

En vista de que la función de autocorrelación ACF es una medida de la relación lineal que existe entre dos instantes cualesquiera dentro de la serie de datos, los valores significativos de ACF representan el número de valores pasados que pueden ser necesarios para generar un modelo lineal. En el caso que nos ocupa, se puede considerar que, para un nivel de confianza del 95%, la gráfica de la figura 5, muestra que el valor de q debe ser igual a 5.

Una vez establecido el modelo $ARMA(5,5)$ como posible modelo óptimo, se procede a obtener otros modelos para órdenes $p, q < 5$. Comparando los resultados obtenidos a través de las pruebas de log-verosimilitud LL y de razón de log-verosimilitud LLR, para la comparación entre los modelos. Se observa que el modelo más relevante fue el $ARMA(1,1)$ más sencillo, que tenía un mejor comportamiento aún que el modelo $ARMA(5,5)$ establecido al principio; sin embargo, al realizar la comparación respectiva LLR con el modelo $AR(5)$ el $p - valor$ resultó ser igual a 0.044, menor que el valor de significancia de 0.05.

3.4. Modelos $ARMAX(p, q)$

Los modelos $ARMAX(p, q)$ son modelos que consideran información exógena para explicar la variable endógena. La expresión matemática para un modelo $ARMAX(p, q)$ con una variable exógena, puede ser escrita de la siguiente manera ecuación (6),

$$I_t = c + \beta Y + \sum_{k=1}^{p} \varphi_k I_{t-k} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \qquad (6)$$

Para tomar una decisión del modelo óptimo a utilizar en este caso, es decir, los órdenes p y q del mismo, así como, la variable o variables exógenas, se analiza la matriz de correlaciones de las variables que podrían estar más relacionadas con la irradiancia solar y se obtuvo los resultados que se consignan en la Tabla 2.

Tabla 2.
Variables exógenas de la Irradiancia Solar Global

	GHUVI	DNI	Albedo	Temp	Presion
GHUVI	1.000000	0.750333	0.094512	0.295878	-0.036080
DNI	0.750333	1.000000	0.046419	0.328789	-0.146964
Albedo	0.094512	0.046419	1.000000	0.079394	0.102156
Temp	0.295878	0.328789	0.079394	1.000000	-0.185378
Presion	-0.036080	-0.146964	0.102156	-0.185378	1.000000

Fuente: Elaboración autor

Se distingue claramente que la única variable que significativamente está más relacionada con la variable $GHUVI$ es la Irradiación Directa Normal, con una correlación del 75%. DNI es variable exógena para definir el o los modelos más óptimos para predecir a $GHUVI$. La prueba DF para un nivel de confianza del 95% da un $p - valor = 2.97 \times 10^{-10}$, por lo que se descarta la presencia de raíces unitarias y se acepta la evidencia significativa de estacionariedad para la serie de la variable DNI.

Se realizaron, además, las pruebas necesarias de log-verosimilitud y de razón de log-verosimilitud entre varios modelos con diferentes órdenes de p y q y, se pudo establecer que los modelos mejor comportados fueron un $ARMAX(3,0)$ y un $ARMAX(1,1)$.

3.5. Modelos *Long-Short Term Memory* (*LSTM*)

Las Redes Neuronales Artificiales (RNA), son un mecanismo que nos permite obtener modelos predictivos bajo un procedimiento similar a los métodos empleados antes, esto es, entrenar el modelo, obtenerlo y luego usarlo para predecir el futuro. Las RNA, son capaces de aprender, descubrir, la regla denominada ecuación actualizadora, que relaciona el futuro del sistema con su pasado (Rodríguez, 2019),

$$x_{t+1} = f(x_1, x_2, \dots, x_t) \qquad (7)$$

Donde, en f se codifican las relaciones dinámicas del sistema para predecir el futuro. Sin embargo, en la práctica (Rodríguez, 2019), es muy difícil establecer una función que dependa de todas las observaciones pasadas. Por esto, la red asume que la información sobre las observaciones $x_1, x_2, \dots, x_t$ se puede codificar en un vector h_t, en cuyo caso, bajo el supuesto que de la serie temporal es estacionaria, la ecuación actualizadora se simplificaría así

$$x_{t+1}, h_{t+1} = f(x_1, h_t) \qquad (8)$$

De acuerdo a Torres (2019), las Redes Neuronales Recurrentes (RNN), son una clase de RNA para analizar datos de series temporales permitiendo tratar la dimensión de "tiempo"; sin embargo, un problema común que enfrentan las RNN es que después de un tiempo, especialmente si está entrenando la red en una secuencia muy grande o larga, esta red comenzará a olvidar las primeras entradas. Las redes *Long-Short Term Memory* ($LSTM$) son una extensión de las RNN, que básicamente amplían su memoria para aprender de experiencias importantes que han pasado hace mucho tiempo.

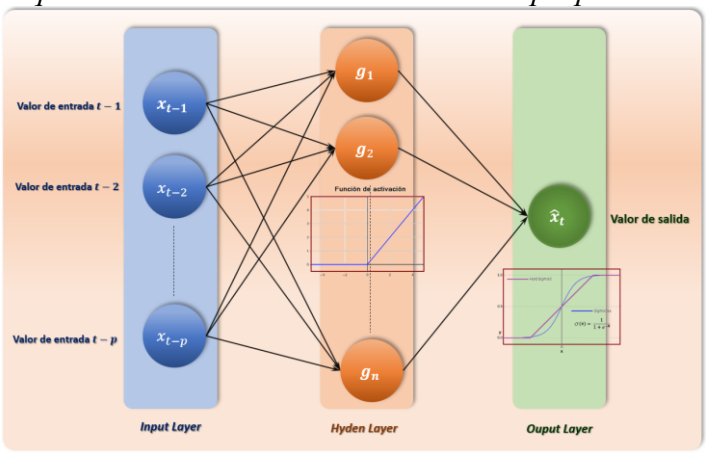
La arquitectura de las redes neuronales, tiene tres componentes básicos:

- *input layer* (capa de entrada)
- *hyden layer* (capa/s oculta/s)
- *ouput layer* (capa de salida)

Un ejemplo de una posible configuración, es el que se muestra en la Figura 4.

Figura 4

Arquitectura de la Red Neuronal. Fuente propia



Fuente: Elaboración autor

El *input layer* consiste de un vector de regresores o *features* $(x_{t-1}, x_{t-2}, \dots, x_{t-p})$ que ingresan a la capa *hyden layer* que contiene n neuronas y cada una aplica una transformación activadora, lineal o no lineal, que genera el output g_i, con

$$g_i = h(w_i x + b_i) \tag{9}$$

Donde los w_i y los b_i son, respectivamente, los pesos y sesgos de la transformación activada por la *función activadora* h, la misma que permite a la red modelizar las complejas relaciones entre los regresores, *features* y la variable objetivo. Las dos funciones h más utilizadas son la *sigmoide* y la *tangente hiperbólica* (tanh), dadas por las ecuaciones (10) y (11).

$$h = \frac{1}{1 - e^{-x}}, sigmoide \tag{10}$$

$$h = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}, \tanh \tag{11}$$

De una manera muy general y sucinta, el proceso que realiza la RNN, se describe así: la red recibe las variables *input* $(x_{t-1}, x_{t-2}, \dots, x_{t-p})$, luego, realiza una serie de operaciones con las funciones activadoras h y, finalmente, transforma el *input* en la predicción $\hat{x}_t$.

Al utilizar la librería *Keras* de *Python*, se analizaron varios modelos, resultando el más óptimo, el modelo *LSTM* caracterizado por los parámetros que se observan en las líneas de código de la Figura 5. El objetivo es minimizar errores en el entrenamiento de la red, a fin de encontrar los pesos adecuados de cara a la convergencia del modelo, asegurando una buena generalización, lo que se logra a través de métodos iterativos, conocidos como métodos de optimización u optimizadores.

Entre las instrucciones que se usaron en este modelo, se tienen *model* = *Sequential*( ), que establece la arquitectura que permite agregar las capas que se consideren necesarias; en este caso, son las capas *model.add*(*LSTM*(*8, …*)) y *model.add*(*Dense*(*1*)), así también, se usa como optimizador el algoritmo *adam* (*adaptive moment estimation*) y, para estimar el error en la *función de pérdida* (*loss*), el MAE. Luego, se ajusta el modelo con la instrucción que arranca el procedimiento de aprendizaje, *model.fit*( ), usando los

datos de entrenamiento y, una vez ajustado, se procede con este modelo, a predecir y comparar los resultados con los datos de prueba.

Figura 5

Código del diseño de la Red Neuronal

```
# Diseño de la Red
model = Sequential()
model.add(LSTM(8, input_shape=(train_X.shape[1], train_X.shape[2])))
model.add(Dense(1))
model.compile(loss='mae', optimizer='adam')
print(model.summary())
# Ajuste de la red
history = model.fit(
    train_X,
    train_y,
    epochs=40,
    batch_size=8,
    validation_data=(test_X, test_y),
    verbose=1,
    shuffle=False)
```

Fuente: Elaboración autor

3.6.Discusión de resultados

Los resultados de la evaluación y comparación del desempeño de cada uno de los modelos entrenados para la predicción, usando las métricas $RMSE$ y MAE, se describen en la Tabla 3.

Tabla 3

Comparación del desempeño de los modelos estadísticos.

Modelo	*RMSE*	*MAE*	*MAPE* (%)
$AR(5)$	7.508	6.297	17.05
$ARMA(1,1)$	7.504	6.296	17.05
$ARMAX(3,0)$	4.732	3.723	9.95
$ARMAX(1,1)$	4.684	3.689	9.87
$SARIMAX(5,0,0)(3,1,0)_{[28]}$	5.719	4.329	11.68
LSTM	0.109	0.061	0.18

Fuente: Elaboración autor

Se puede advertir que, los modelos con los errores más altos, son el $AR(5)$ y $ARMA(1,1)$, en los que el *error porcentual medio absoluto* ($MAPE$) es del orden del 17.05% y, al observar la opción (a) de la Figura 6, los resultados evidencian que estos modelos no son una buena opción para predecir la serie de irradiancia solar $GHUVI$.

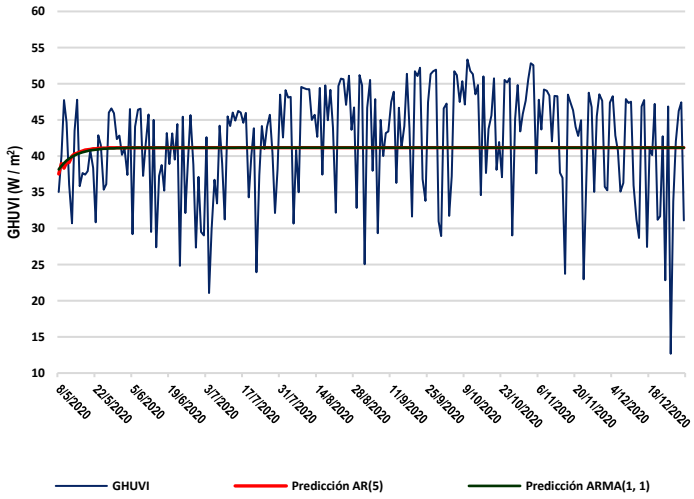
Cuando se introdujo en los modelos la variable exógena DNI, se comprueba que los mejores modelos para predecir con una mayor precisión a la variable $GHUVI$, son el $ARMAX(3,0)$ y el $ARMAX(1,1)$, el $MAPE$ obtenido para cada uno de éstos es 9.5% y 9.87%, respectivamente. El literal (b) de la Figura 6, muestra gráficamente lo mencionado.

Con el propósito de evaluar si introducir la estacionalidad permitiría encontrar modelos que mejoren las predicciones, se procedió a estimar aquellos que consideren esta opción. De los modelos evaluados, el que resultó con mejores predicciones fue el $SARIMAX(5,0,0)(3,1,0)_{[28]}$ que tuvo un $MAPE$ de 11.68%, pero que gráficamente mejora sustancialmente las predicciones de los dos primeros modelos (ver Figura 6, (c)), aunque la precisión medida es menor que las obtenidas para los modelos $ARMAX(3,0)$ y

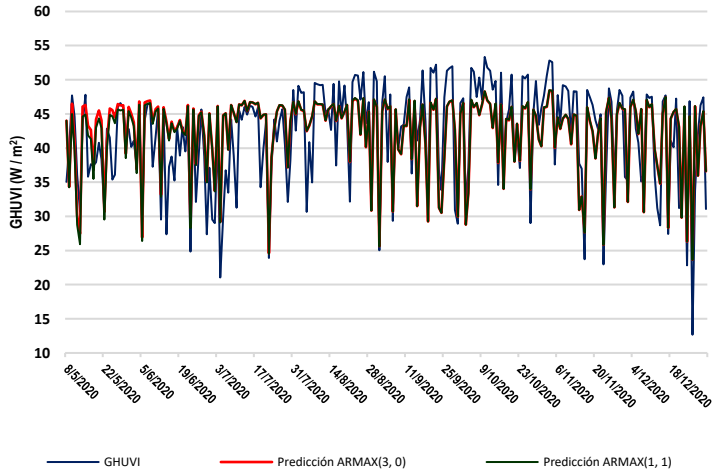
$ARMAX(1,1)$.

Figura 6

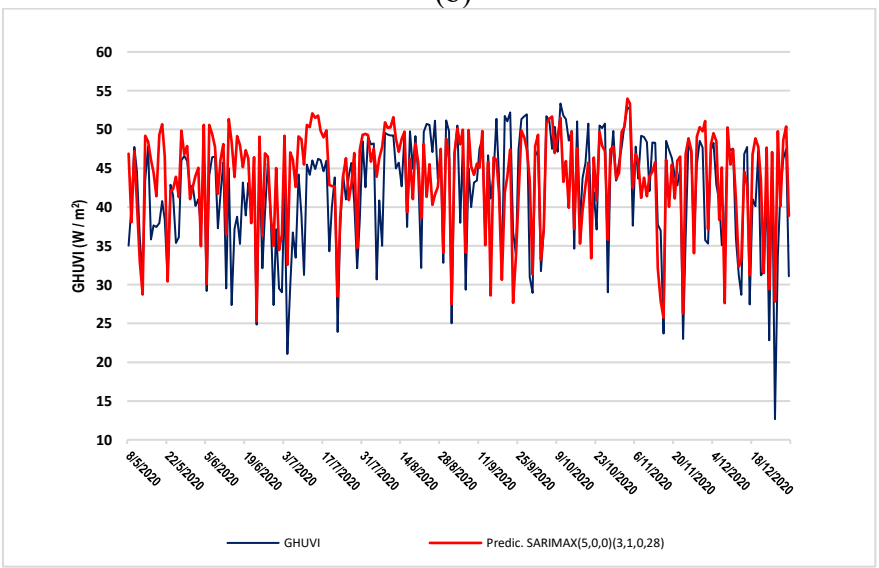
Modelos Estadísticos con mejores predicciones. (a) Modelos $AR(5)$ *y* $ARMA(1,1)$*, (b) Modelos* $ARMAX(3,0)$ *y* $ARMAX(1,1)$ *y (c) Modelo* $SARIMAX(5,0,0)(3,1,0)_{[28]}$



(a)



(b)



(c)

Fuente: Elaboración autor

Para el caso del modelo generado para predecir la variable $GHUVI$, usando el conjunto de

entrenamiento a partir de las redes neuronales *LSTM*, los datos observados para el conjunto de prueba y los pronosticados muestran un ajuste casi perfecto.

La complejidad del modelo se resume en la Tabla 4 y la precisión con la que el modelo de red neuronal realiza las predicciones para una entrada determinada, usando la *función de pérdida* (*Loss function*), se observa en la figura 7; la misma permite concluir que, a partir del *ciclo* (*epoch*) 20, todos los datos de entrenamiento que pasan por la red neuronal para que ésta aprenda sobre ellos, producen una mejora significativa en el desempeño del modelo (Pérez et al., 2021).

Tabla 4

Resumen del Modelo de Red Neuronal

Model: "sequential_2"

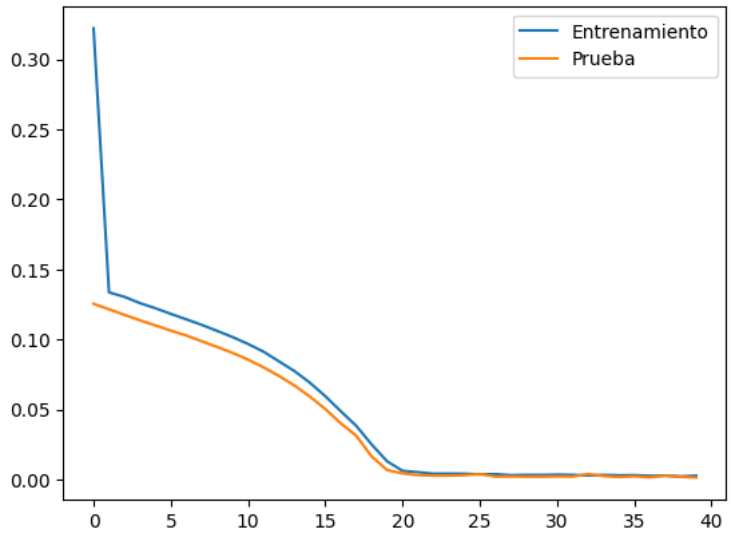
Layer (type)	Output Shape	Param #
lstm_2 (LSTM)	(None, 8)	320
dense_2 (Dense)	(None, 1)	9

Total params: 329
Trainable params: 329
Non-trainable params: 0

Fuente: Elaboración autor

Figura 7

Función de pérdida



Fuente: Elaboración autor

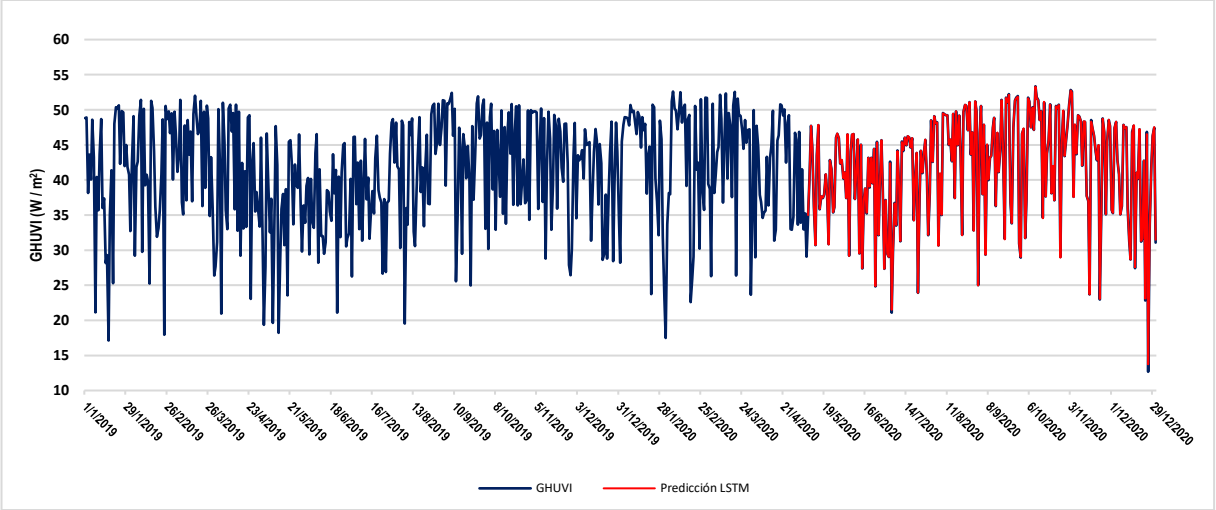
De acuerdo a los resultados de la Tabla 3 para el modelo *LSTM*, los valores obtenidos para el *RSME* y el *MAE* son 0.109 y 0.061, respectivamente, mientras que el *MAPE* es del orden de 0.18%. Gráficamente, el resultado del ajuste se presenta en la figura 8.

Al realizar un análisis comparativo con los resultados obtenidos por otras investigaciones, se encontró que para Clavijo et al. (2019) el método *ARMA* tuvo un desempeño del 15.4560% y de 4,2965% usando redes neuronales; así también, para Villarreal (2020) el pronóstico evaluado a través del error medio absoluto porcentual (*MAE*) fue, de 14% en promedio para predicciones durante tiempo seco y de 29% para

predicciones en temporada de lluvias con el modelo de series de tiempo $SARIMA$, mientras que, para el modelo de redes neuronales se obtuvo un desempeño de 19% para temporada seca y de 31,3% para temporada de lluvias. Si bien es cierto que, el desempeño logrado para predecir por los modelos AR y $ARMA$, en el caso de esta investigación, son de 17,05%; al introducir en los modelos una variable exógena, el desempeño tiene una mejora notable, pues el error para los modelos ARX, $ARMAX$ y $SARIMAX$ es menor que las investigaciones mencionadas, con valores de 9.95%, 9.97% y 11.68%, respectivamente; pero, cuando comparamos con el que se obtuvo para redes neuronales, el desempeño fue considerablemente mejor con un 0.18% de error promedio.

Figura 8

Entrenamiento y Predicción de la Red Neuronal



Fuente: Elaboración autor

A vista de lo expresado anteriormente, parece obvio que la manera más óptima de obtener modelos de predicción, para el caso de la serie de tiempo de la variable $GHUVI$, es a través de redes neuronales $LSTM$.

CONCLUSIONES

El desarrollo de la presente investigación tuvo como principal objetivo contrastar los modelos estadísticos tradicionales de predicción, comparados entre sí y, entre el modelo de redes neuronales recurrentes $LSTM$.

De esta manera, se logró obtener varios modelos de predicción, los mismos que respondían a ciertas consideraciones que, en principio, se hicieron sobre la base del análisis de las funciones de autocorrelación (ACF) y de autocorrelación parcial ($PACF$) de la serie $GHUVI$, con la mira de establecer el modelo $ARMA$ que mejor se ajustaba a los datos de la variable en cuestión.

La prueba LLR entre los modelos $AR(5)$ y $ARMA(1,1)$ dio un valor de 0,044, muy próximo al valor de significancia de 0.05, lo que plantea la similitud del ajuste de los dos modelos, afirmación que se ve respaldada cuando comparamos los resultados obtenidos para el $RMSE$ y el MAE que, a la décima más cercana son 7.5 y 6.3, respectivamente, así como, por un $MAPE$ igual a 17% en ambos modelos; sin embargo, las predicciones no describen de manera precisa, el comportamiento de la serie para los datos de

prueba como se observa en la figura 6 (a).

Al introducir en el modelado la variable DNI, correlacionada con la variable $GHUVI$ en el orden del 75%, el análisis de los modelos $ARMAX(3,0)$ y $ARMAX(1,1)$, con un $MAPE$ de 9.95% y 9.87%, respectivamente, permitió observar una mejora sustancial relativa del 71% en el $MAPE$ con respecto a los modelos antes mencionados. Con estos modelos se consigue una buena aproximación a los datos de prueba de la serie, como se observa en la figura 6 (b).

Si bien es cierto que, los modelos basados en procesos $ARMA$ son más fáciles de entender, pues, por ejemplo, los valores de los coeficientes y el número de retrasos necesarios, pueden ser observados, en las redes neuronales, en cambio, ese tipo de información permanece técnicamente oculta en algún lugar de la red; no obstante, en contraste con los modelos anteriores, el modelo $LSTM$ permitió modelar un proceso que se ajusta de manera casi perfecta a los datos de prueba. Para hacerlo, se controló un umbral de *epochs* a probar, 40 en este caso, con el fin de no llegar a un sobreajuste del modelo, pero también, no sacrificar su precisión; así mismo, la configuración de la red tampoco fue tan compleja, ya que la capa intermedia no cuenta con más de 8 neuronas (8 *hyden layer*) de memoria larga-corta y al no ser tan compleja, responde pronto con el ajuste del modelo de red neuronal. El porcentaje de error medio absoluto medido fue de 0.18%.

FINANCIACIÓN

Los autores no recibieron financiación para el desarrollo de la presente investigación.

CONFLICTO DE INTERESES

Los Autores declaran que no existe conflicto de intereses

CONTRIBUCIÓN DE AUTORÍA

En concordancia con la taxonomía establecida internacionalmente para la asignación de créditos a autores de artículos científicos (https://credit.niso.org/). Los autores declaran sus contribuciones en la siguiente matriz:

	Quinteros I.	*Tafur P.*
Participar activamente en:		
Conceptualización	X	X
Análisis formal	X	X
Adquisición de fondos	X	
Investigación	X	X
Metodología	X	X
Administración del proyecto	X	X
Recursos	X	
Redacción –borrador original	X	X
Redacción –revisión y edición	X	X
La discusión de los resultados	X	X
Revisión y aprobación de la versión final del trabajo.	X	X

RECONOCIMIENTO A REVISORES:

La revista reconoce el tiempo y esfuerzo del editor Juan Santillán, y de revisores anónimos que dedicaron su tiempo y esfuerzo en la evaluación y mejoramiento del presente artículo.

REFERENCIAS:

Capa, H. (2022-). *Modelación de series temporales* (2da. Ed.). Escuela Politécnica Nacional

Carrillo Andrade, F. A. (2022). *Pronóstico del recurso solar a corto plazo para Distritos industriales basado en redes neuronales artificiales* (Bachelor's thesis). https://dspace.ups.edu.ec/bitstream/123456789/21913/1/UPS%20-%20TTS619.pdf

Castro, C., Lima, V., & Figueiredo, M. (2019). Series temporais. In A. C. Medeiros, C. L. Ribeiro, & D. C. Medeiros (Eds.), Introdução à estatística aplicada à engenharia: conceitos, métodos e aplicações (pp. 215-236). Springer.

Clavijo Galvis, F. S., & Pinto Pérez, C. A. Implementación de métodos de predicción de radiación solar para una zona particular de la geografía colombiana. https://repository.udistrital.edu.co/handle/11349/29465

Collantes Duarte, J., Colmenares La Cruz, G., Orlandoni Merli, G., & Rivas Echeverría, F. (2004). A comparison of time series forecasting between artificial neural networks and box and jenkins methods. *Revista Técnica de la Facultad de Ingeniería Universidad del Zulia*, *27*(3), 146-160. http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0254-07702004000300002

Contreras, W., Galban, M. G., & Sepúlveda, S. B. (2018). Análisis estadístico de la radiación solar en la ciudad de Cúcuta. *Entre ciencia e ingeniería*, *12*(23), 16-22. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1909-83672018000100016

Cuitiño, F., Ganón, E., Tiscordio, I., & Vicente, L. (2010). Modelos univariados de series de tiempo para predecir la inflación de corto plazo. *XXV Jornadas de Economía del Banco Central del Uruguay*. https://www.bcu.gub.uy/Comunicaciones/Jornadas%20de%20Economa/iees03j3101010.pdf

Enders, W. (2018). Applied econometric time series (4th ed.). Wiley.

Gómez-Mejía, A. (2020). Modelo de Máxima Verosimilitud. *Libre Empresa*, *17*(2), 121-138. https://revistas.unilibre.edu.co/index.php/libreempresa/article/view/8027/7195

Hassani, H. (2018). Time series analysis and forecasting: A practical approach for business forecasting (2nd ed.). Wiley.

Lalaleo Achachi, D. F. (2021). *Diseño de un algoritmo utilizando Machine Learning para la predicción de la radiación solar en el sector de Lasso* (Master's thesis, Ecuador: Latacunga: Universidad Técnica de Cotopaxi: UTC.). http://repositorio.utc.edu.ec/handle/27000/8014

López-García, M. D. R., Martínez-Damián, M. Á., & Arana-Coronado, J. J. (2022). Predictores del precio de maíz blanco en Jalisco y Michoacán. *Revista mexicana de ciencias agrícolas*, *13*(2), 261-272. https://www.scielo.org.mx/scielo.php?pid=S2007-09342022000200261&script=sci_arttext

Ma, T., Gu, W., Shen, L., & Li, M. (2019). An improved and comprehensive mathematical model for solar photovoltaic modules under real operating conditions. *Solar Energy*, *184*, 292-304. https://www.sciencedirect.com/science/article/abs/pii/S0038092X19303159

Mazorra Aguiar, L. (2016). *Modelo predictivo de radiación solar mediante técnicas de machine learning:*

aplicación a la isla de Gran Canaria (Doctoral dissertation). https://accedacris.ulpgc.es/handle/10553/18927

Pérez, D. P., Botto-Tobar, M., & Mora, C. M. (2021). Predicción del Composite Requerido en el Diseño de un Recipiente Toroidal Mediante una Red Neuronal Artificial. *Investigación, Tecnología e Innovación*, *13*(13), 45-53. https://revistas.ug.edu.ec/index.php/iti/article/view/1093

Pinzón, J. E. D. (2020). Precisión del pronóstico de la propagación del COVID-19 en Colombia. *Revista Repertorio de Medicina y Cirugía*. https://revistas.fucsalud.edu.co/index.php/repertorio/article/view/1045

Rangel Heras, E. (2018). Modelo integral para la predicción de la potencia generada por equipos fotovoltaicos de gran escala. http://bibliotecavirtual.dgb.umich.mx:8083/xmlui/handle/DGB_UMICH/317

Rodríguez, A. A. (2019). Análisis de las series temporales a la luz de Deep Learning. Anuario jurídico y económico escurialense, (52), 257-276. https://dialnet.unirioja.es/servlet/articulo?codigo=6883981

Sayed, E. T., Wilberforce, T., Elsaid, K., Rabaia, M. K. H., Abdelkareem, M. A., Chae, K. J., & Olabi, A. G. (2021). A critical review on environmental impacts of renewable energy systems and mitigation strategies: Wind, hydro, biomass and geothermal. *Science of the total environment*, *766*, 144505. https://www.sciencedirect.com/science/article/abs/pii/S0048969720380360?via%3Dihub

Segarra Poma, A. J. (2022). *Ubicación óptima georreferenciada de centrales de generación fotovoltaica considerando restricciones de radiación solar y temperatura* (Bachelor's thesis). https://dspace.ups.edu.ec/handle/123456789/21904

Tafur-Escanta, P., López-Paniagua, I., & Muñoz-Antón, J. (2023). Thermodynamics analysis of the supercritical CO2 binary mixtures for Brayton power cycles. *Energy*, 126838. https://www.sciencedirect.com/science/article/pii/S0360544223002323

Tembhare, S. P., Barai, D. P., & Bhanvase, B. A. (2022). Performance evaluation of nanofluids in solar thermal and solar photovoltaic systems: A comprehensive review. *Renewable and Sustainable Energy Reviews*, *153*, 111738. https://www.sciencedirect.com/science/article/abs/pii/S1364032121010108

Torres, J. (2019). Deep Learning, Introducción práctica con Keras, SEGUNDA PARTE. WATCH THIS SPACE collection – Barcelona: Book 6. https://torres.ai/deep-learning-inteligencia-artificial-keras-2a-parte/#AccesoLibro2aParte

Villarreal Mesa, O. A. (2020). Análisis del recurso solar mediante modelos de predicción de corto plazo en la sabana de Bogotá. https://repositorio.unal.edu.co/handle/unal/77790?locale-attribute=pt_BR