GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Predicción de la Producción de Energía en Centrales Solares mediante Modelos de Aprendizaje Automático

Autor: Javier Campo Herrero

Director: Atilano Ramiro Fernández-Pacheco Sánchez-Migallón

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Predicción de la Producción de Energía en Centrales Solares mediante Modelos de Aprendizaje Automático

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2024/25 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Javier Campo Herrero Fecha: 4/7/2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

71216314B ATILANO RAMIRO FERNÁNDEZ-PACHECO

Digitally signed by 71216314B ATILANO RAMIRO FERNÁNDEZ-PACHECO
Date: 2025.07.04 15:31:50 +02'00'

Fdo.: Atilano Ramiro Fernández-Pacheco Sánchez-Migallón Fecha: 4/7/2025

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

Predicción de la Producción de Energía en Centrales Solares mediante Modelos de Aprendizaje Automático

Autor: Javier Campo Herrero

Director: Atilano Ramiro Fernández-Pacheco Sánchez-Migallón

Madrid

Agradecimientos

Agradezco a mis padres que plantaran en mí el afán por entender el mundo, esa curiosidad la que me lleva cada día a querer saber siempre más. Me han acompañado y guiado a lo largo de todo el camino de mi vida. Debo a ellos todo lo que he logrado, todos los retos que he superado. Por ello les estoy eternamente agradecido.

A mis hermanos, que me mantengan con los pies en la Tierra y me den perspectiva. Que hayan visto siempre lo mejor de mí, como yo he visto lo mejor de ellos. Sé que puedo contar con ellos para cualquier cosa.

A mis abuelos, por ser una inspiración para mí. Por enseñarme con sus vivencias cómo desenvolverse por el mundo. Su fortaleza, principios y paciencia son un ejemplo para mí.

A mis amigos, con quienes he compartido las alegrías y los desafíos que he vivido. Por haber seguido a mi lado este complicado camino que es la ingeniería y por haber estado junto a mí para asentarme y apoyarme siempre que lo he necesitado.

PREDICCIÓN DE LA PRODUCCIÓN DE ENERGÍA EN CENTRALES SOLARES MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO

Autor: Campo Herrero, Javier.
Director: Fernández-Pacheco Sánchez-Migallón, Atilano Ramiro.
Entidad Colaboradora: ICAI - Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este estudio comparativo evaluó la evolución desde regresiones lineales básicas hasta SARIMA, Bosque Aleatorio y redes LSTM para la predicción fotovoltaica en horizontes diario y anual. Se constató una mejora continua en precisión y en varianza explicada, alcanzando los mejores resultados con LSTM. La segmentación temporal y la optimización de hiperparámetros demostraron ser claves para adaptar cada modelo a su escala específica.

Palabras clave: Aprendizaje Automático, Bosque Aleatorio, Planta Solar Fotovoltaica, Long Short-Term Memory.

1. Introducción

La penetración creciente de la generación fotovoltaica en los sistemas eléctricos ha situado la previsión de su producción entre los retos prioritarios de la ingeniería energética. La potencia instantánea suministrada por las instalaciones solares está condicionada por la irradiancia disponible y, por tanto, sometida a variaciones abruptas motivadas por factores atmosféricos cuya manifestación posee escalas temporales heterogéneas. Bajo estas circunstancias, se requiere disponer de modelos de pronóstico que suministren con antelación suficiente estimaciones fiables, capaces de guiar tanto la programación del despacho como la definición de márgenes de reserva. La ausencia de precisión en dichos cálculos deriva en incrementos de coste y en riesgos de desbalance que comprometen la estabilidad operativa de la red; de ahí que la robustez del algoritmo predictivo se erija en condición indispensable para el aprovechamiento efectivo del recurso solar.

Dentro de este contexto, el presente trabajo examina la evolución de una serie de enfoques de regresión temporal con el fin de determinar qué estrategia reproduce con mayor fidelidad la complejidad de la señal histórica de producción. La indagación parte de modelos lineales elementales, que se emplean como línea base por su sencillez interpretativa y su reducido coste computacional, y avanza de forma sistemática hasta arquitecturas de memoria recurrente, cuya expresividad permite incorporar dependencias temporales prolongadas y no

linealidades de alto orden. A lo largo del proceso comparativo se ha mantenido un criterio experimental uniforme que contempla dos resoluciones de análisis: la intradía, orientada a la gestión operativa, y la agregada anual, destinada a la planificación estratégica. Dicha doble perspectiva persigue verificar si la misma familia de modelos resulta adecuada para horizontes dispares o si conviene adoptar soluciones especializadas que optimicen la precisión en cada escala.

La motivación técnica que respalda esta investigación radica en la necesidad de adaptar los métodos de pronóstico a un entorno caracterizado por la disponibilidad creciente de datos históricos y por la exigencia regulatoria de integrar energías renovables intermitentes sin comprometer la seguridad del suministro. Se parte de la hipótesis de que un incremento gradual en la complejidad del modelo, acompañado de una validación rigurosa, conduce a mejoras progresivas en la calidad de las estimaciones y en la estabilidad de los intervalos de confianza. Asimismo, se asume que la inclusión explícita de información temporal, ya sea a través de componentes autorregresivos o de memorias internas, resulta decisiva para capturar la naturaleza cíclica y estacional de la producción fotovoltaica.

En consecuencia, el objetivo fundamental consiste en proporcionar un análisis comparativo exhaustivo que justifique la adopción de técnicas avanzadas de aprendizaje automático frente a los métodos estadísticos convencionales, con vistas a su posterior implantación en sistemas de control y supervisión energética. A tal fin, se ha estructurado el estudio en bloques coherentes que describen, para cada propuesta, el planteamiento metodológico, el proceso de ajuste y la evaluación de desempeño mediante métricas estandarizadas. La discusión resultante pretende ofrecer una guía razonada para la selección de la arquitectura más adecuada según el horizonte temporal de interés y la disponibilidad de recursos computacionales, sentando las bases para futuras extensiones orientadas a la incorporación de variables meteorológicas exógenas y a la estimación probabilística de la incertidumbre.

2. Definición del Proyecto

El proyecto se articuló como un estudio comparativo que examinó, de manera escalonada, la validez de distintos paradigmas de regresión temporal aplicados a la predicción fotovoltaica. Para tal fin, se diseñó una ruta metodológica que arrancó con modelos lineales de complejidad mínima y avanzó de forma progresiva hacia técnicas autorregresivas, ensambles no paramétricos y, por último, arquitecturas de aprendizaje profundo basadas en

memoria recurrente. Este planteamiento incremental permitió aislar los efectos de cada salto conceptual y evaluar con rigor la aportación específica de cada familia de modelos.

Para contrastar la premisa de que la capacidad explicativa crecería a medida que se incrementara la flexibilidad funcional del estimador se configuraron, en primer lugar, regresiones lineales simple y polinómica; a continuación se incorporaron variantes semanales con selección de ventana optimizada; posteriormente se introdujo un modelo SARIMA que añadió componente autorregresiva e integrada; seguidamente se adoptó un Bosque Aleatorio como representante de los métodos de ensamblado no paramétrico; y, finalmente, se implementaron redes neuronales Long Short-Term Memory para explorar el potencial del aprendizaje profundo en series temporales.

Con objeto de analizar la sensibilidad de cada enfoque ante variaciones de escala, se estableció un marco experimental bifurcado en dos horizontes temporales independientes. El primero, de resolución diaria, se orientó a la gestión operativa y describió la dinámica intradía tras la normalización de los valores de cada día. El segundo, de carácter anual, se obtuvo por agregación de la producción diaria y se destinó a la planificación estratégica. A partir del bloque Bosque Aleatorio se decidió entrenar y evaluar modelos diferenciados para cada horizonte, con el fin de comprobar si la especialización multiescala aportaba ventajas respecto a la aplicación de un único estimador sobre la serie completa.

El procedimiento de validación se sostuvo en la separación cronológica de los datos en conjuntos de entrenamiento y prueba, evitando la contaminación de futuro y reflejando las condiciones reales de explotación. La evaluación de desempeño se basó primordialmente en el error cuadrático medio y en el coeficiente de determinación, complementados, cuando fue pertinente, por métricas absolutas que ofrecieron una visión adicional de la dispersión de los residuos. Para garantizar la reproducibilidad, se fijaron semillas deterministas en los procesos de aleatorización y se almacenaron las configuraciones de hiperparámetros empleadas en cada experimento.

En resumen, el proyecto quedó definido como una plataforma de comparación rigurosa, estructurada en fases sucesivas de complejidad creciente y fundamentada en una doble perspectiva temporal que reproduce las necesidades reales de la operación fotovoltaica. Esta definición permitió identificar con claridad las líneas de mejora entre bloques metodológicos y sentó las bases para las recomendaciones y extensiones futuras.

3. Descripción del Sistema

El sistema experimental se estructuró en cuatro bloques metodológicos consecutivos, cada uno de ellos concebido para superar las limitaciones detectadas en la fase previa. En la etapa inicial se implementó la regresión lineal ordinaria como línea base y, a continuación, se ampliaron sus posibilidades mediante la incorporación de términos polinómicos hasta quinto grado. Estos modelos se entrenaron sobre la serie completa y sobre subconjuntos semanales con ventana deslizante, con el propósito de evaluar la influencia del horizonte de entrenamiento en la capacidad explicativa. La estimación de los parámetros se llevó a cabo mediante mínimos cuadrados ordinarios, y la validez de los ajustes se comprobó por medio del análisis de residuos.

El segundo bloque introdujo un modelo SARIMA, en el que se aplicó diferenciación de primer orden para lograr la estacionariedad exigida por la metodología. La identificación de los órdenes autorregresivo y de medias móviles se efectuó utilizando funciones de autocorrelación y criterios de información como AIC. Una vez seleccionados los parámetros correspondientes, los coeficientes se estimaron por máxima verosimilitud y se validaron mediante pruebas de diagnóstico sobre autocorrelación remanente.

En la tercera fase se adoptó el Bosque Aleatorio como representante de los métodos de ensamblado no paramétrico. Para la serie diaria se construyó un vector de predictores que incluyó la hora del día, mientras que, en la serie anual, se conservaron exclusivamente la media de cada día que explican la variación estacional. Cada bosque se entrenó con un centenar de árboles empleando la técnica de *bootstrap* y la selección aleatoria de características en cada nodo, lo que aseguró la diversificación interna y redujo la varianza global del estimador. La profundidad máxima y el número mínimo de muestras por hoja se fijaron mediante validación sobre un conjunto de prueba cronológicamente posterior, a fin de prevenir el sobreajuste.

El último bloque metodológico correspondió a las redes neuronales Long Short-Term Memory. Para la resolución diaria se generaron tensores tridimensionales a partir de ventanas semanales de noventa y seis pasos, normalizados mediante la transformación min-max, y para la escala anual se utilizaron secuencias de siete días con reescalado equivalente. La optimización de la arquitectura se abordó mediante una búsqueda exhaustiva de hiperparámetros que combinó tasas de aprendizaje logarítmicamente espaciadas, valores de abandono graduados y tamaños de lote crecientes. Cada combinación se evaluó con

validación temporal en pliegues crecientes, y se aplicó parada temprana para evitar la convergencia hacia mínimos locales inadecuados. El modelo ganador se reentrenó sobre la totalidad del conjunto de entrenamiento y se almacenó junto con el registro de hiperparámetros, garantizando así la reproducibilidad y la trazabilidad de los resultados.

4. Resultados

Los experimentos demostraron una tendencia ascendente en la calidad de las estimaciones a medida que se incrementó la complejidad de los modelos evaluados. En la fase inicial, la regresión , tanto en su versión lineal como en las extensiones polinómicas, ofreció un ajuste elemental que únicamente sirvió para fijar la referencia mínima de desempeño; las mejoras marginales obtenidas mediante la adición de términos de mayor grado confirmaron que la estructura lineal resultó insuficiente para captar la variabilidad intrínseca de la señal fotovoltaica. El paso al modelo SARIMA introdujo la diferenciación temporal y la autocorrelación explícita, con lo que se redujo la magnitud de los errores. Sin embargo, el análisis residual mostró que los intervalos de confianza conservaban una amplitud excesiva, especialmente en horizontes alejados del presente.

El salto conceptual más relevante se produjo con la adopción del Bosque Aleatorio. Gracias a la aleatorización interna y a la agregación de múltiples árboles, este método logró representar discontinuidades y umbrales que los modelos autorregresivos no podían describir. Como consecuencia, se observó una disminución pronunciada del error cuadrático medio y un incremento notorio del coeficiente de determinación, tanto en la resolución diaria como en la anual.

Finalmente, las redes Long Short-Term Memory culminaron la progresión metodológica al incorporar memoria interna capaz de retener dependencias de medio y largo plazo. Tras una optimización exhaustiva de hiperparámetros, se obtuvieron las predicciones más precisas de todo el estudio y los intervalos de confianza más ajustados, con un comportamiento robusto en ambos horizontes temporales. La comparación directa con el Bosque Aleatorio evidenció reducciones adicionales del error y un aumento sustancial del coeficiente de determinación, lo que confirmó que la capacidad de las LSTM para modelar secuencias completas resulta decisiva en contextos dominados por estacionalidades complejas y variabilidad climática. En conjunto, los resultados corroboraron que la incorporación gradual de flexibilidad funcional, desde la linealidad estricta hasta la memoria recurrente, constituye el camino más eficaz para mejorar la fiabilidad de los pronósticos fotovoltaicos.

5. Conclusiones

En síntesis, el análisis comparativo llevado a cabo revela que la linealidad estricta resulta incapaz de reproducir la complejidad estacional y la dinámica intradía inherentes a la producción fotovoltaica. Aunque la incorporación de componentes autorregresivos permitió reducir la magnitud absoluta de los errores, el carácter lineal de dichos modelos continuó lastrando la proporción de varianza explicada. La transición posterior hacia métodos de ensamblado no paramétrico confirmó que la flexibilidad funcional constituye un requisito indispensable, si bien la fragmentación jerárquica de los predictores mostró ciertas limitaciones para retener dependencias de largo plazo.

La introducción de redes Long Short-Term Memory, optimizadas de forma rigurosa y entrenadas de manera diferenciada en los horizontes diario y anual, proporcionó un salto cualitativo definitivo. La arquitectura dual permitió que cada red profundizara en la escala temporal para la que fue concebida, superando al resto de enfoques tanto en precisión como en estabilidad de las bandas de predicción. Se concluye, por consiguiente, que la combinación de memoria recurrente y segmentación temporal especializada constituye la solución más adecuada para satisfacer los requisitos operativos y estratégicos de la gestión fotovoltaica.

De cara a trabajos futuros se considera prioritario ampliar el espacio de información mediante la integración de variables meteorológicas externas procedentes de satélites y estaciones de superficie, así como incorporar técnicas de cuantificación probabilística que acompañen cada estimación con intervalos de confianza calibrados. Asimismo, la implantación de esquemas de aprendizaje incremental permitiría ajustar los parámetros del modelo en tiempo real y reflejar la evolución de las condiciones de operación sin incurrir en reentrenamientos exhaustivos. Estas líneas de investigación se perfilan como las más prometedoras para reforzar la robustez y la adaptabilidad de los sistemas de pronóstico en escenarios de alta penetración solar.

FORECASTING ENERGY PRODUCTION IN SOLAR POWER PLANTS USING MACHINE LEARNING MODELS

Author: Campo Herrero, Javier.
Supervisor: Fernández-Pacheco Sánchez-Migallón, Atilano Ramiro.
Collaborating Entity: ICAI - Universidad Pontificia Comillas

ABSTRACT

This comparative study assessed photovoltaic forecasting methods from basic linear regression through SARIMA, Random Forest, and LSTM on both daily and annual horizons. Accuracy and explained variance improved at each stage, with LSTM delivering the highest precision. Temporal segmentation and rigorous hyperparameter tuning proved essentialto adapt each model to its respective scale.

Key words: Machine Learning, Random Forest, Long Short-Term Memory, Photovoltaic Solar Plant.

1. Introduction

The growing penetration of photovoltaic generation in electrical systems has placed the forecasting of its output among the foremost challenges in energy engineering. The instantaneous power delivered by solar installations depends on available irradiance and is therefore subject to abrupt fluctuations driven by atmospheric factors that manifest across heterogeneous time scales. Under these conditions, it is essential to employ forecasting models that provide reliable estimates with sufficient lead time to guide both dispatch scheduling and reserve margin setting. A lack of precision in these calculations results in increased costs and risk of imbalance, thereby jeopardizing the operational stability of the grid; accordingly, the robustness of the predictive algorithm becomes indispensable for the effective utilization of the solar resource.

Within this context, the present study examines the evolution of a series of temporal regression approaches to determine which strategy most faithfully reproduces the complexity of the historical generation signal. The investigation begins with elementary linear models, used as a baseline due to their interpretability and low computational cost, and systematically advances to recurrent-memory architectures, whose expressive power allows them to capture extended temporal dependencies and high-order nonlinearities. Throughout the comparative process, a uniform experimental protocol was maintained,

encompassing two analytical resolutions: intraday, focused on operational management, and annual aggregation, aimed at strategic planning. This dual perspective seeks to verify whether the same model family is suitable for disparate horizons or whether specialized solutions should be adopted to optimize accuracy at each scale.

The technical motivation underpinning this research lies in the need to adapt forecasting methods to an environment characterized by rapidly growing historical data availability and regulatory mandates to integrate intermittent renewables without compromising supply security. It is hypothesized that a gradual increase in model complexity, coupled with rigorous validation, leads to progressive improvements in estimation quality and in the stability of confidence intervals. Likewise, it is assumed that the explicit inclusion of temporal information, whether through autoregressive components or internal memory, proves decisive in capturing the cyclical and seasonal nature of photovoltaic production.

Consequently, the primary objective is to provide an exhaustive comparative analysis that justifies adopting advanced machine-learning techniques over conventional statistical methods, with a view to subsequent implementation in energy control and supervision systems. To this end, the study has been structured into coherent blocks that, for each modeling proposal, describe the methodological approach, the parameter-tuning process, and performance evaluation using standardized metrics. The resulting discussion aims to offer a reasoned guide for selecting the most appropriate architecture according to the temporal horizon of interest and available computational resources, thereby laying the groundwork for future extensions focused on integrating exogenous meteorological variables and on probabilistic uncertainty quantification.

2. Project Definition

The project was structured as a comparative study that examined, in a stepwise fashion, the validity of various temporal regression paradigms applied to photovoltaic forecasting. To this end, a methodological pathway was designed, beginning with minimally complex linear models and progressing through autoregressive techniques, non-parametric ensembles, and, finally, deep learning architectures based on recurrent memory. This incremental approach made it possible to isolate the effects of each conceptual leap and rigorously assess the specific contribution of each model family.

The working hypothesis was that explanatory power would increase in proportion to the functional flexibility of the estimator. To test this premise, simple and polynomial linear regressions were configured first; next, weekly variants with optimized sliding-window selection were introduced; this was followed by a SARIMA model incorporating autoregressive and integrated components; then a Random Forest was adopted to represent non-parametric ensemble methods; and, finally, Long Short-Term Memory networks were implemented to explore the potential of deep learning in time-series forecasting.

In order to analyze each approach's sensitivity to scale, an experimental framework was established with two independent temporal horizons. The first, at daily resolution, focused on operational management and captured intraday dynamics after normalizing the quarter-hourly values. The second, at annual scale, was obtained by aggregating daily production and aimed at strategic planning. Beginning with the Random Forest block, distinct models were trained and evaluated for each horizon to determine whether multiscale specialization offered advantages over applying a single estimator to the complete series.

Validation relied on chronologically splitting the data into training and test sets, thereby preventing look-ahead bias and reflecting real exploitation conditions. Performance evaluation was based primarily on mean squared error and the coefficient of determination. To ensure reproducibility, deterministic seeds were fixed for all randomization processes, and the hyperparameter configurations used in each experiment were archived.

In summary, the project was defined as a rigorous comparison platform, organized into successive phases of increasing complexity and underpinned by a dual temporal perspective that mirrors the real needs of photovoltaic operation. This definition made it possible to clearly identify lines of improvement between methodological blocks and laid the groundwork for future recommendations and extensions.

3. System Description

The experimental framework was organized into four consecutive methodological blocks, each designed to overcome the limitations identified in the previous phase. In the initial stage, ordinary linear regression was implemented as the baseline and then extended with polynomial terms up to fifth degree. These models were trained on both the full series and on weekly subsets using a sliding window in order to assess the influence of the training

horizon on explanatory power. Parameter estimation was performed via ordinary least squares.

The second block introduced a SARIMA model, in which first-order differencing was applied to achieve the stationarity required by the methodology. The autoregressive and moving-average orders were identified using autocorrelation functions and information criteria such as AIC. Once the parameters were selected, coefficients were estimated by maximum likelyhood.

In the third phase, the Random Forest algorithm was adopted to represent non-parametric ensemble methods. For the daily series, a predictor vector was constructed that included the normalized fifteen-minute time index and the calendar day number; for the annual series, only calendar references explaining seasonal variation were retained. Each forest was trained with one hundred trees using bootstrap sampling and random feature selection at each node, ensuring internal diversity and reducing the overall variance of the estimator. Maximum tree depth and the minimum number of samples per leaf were fixed via validation on a chronologically subsequent test set to prevent overfitting.

The final methodological block comprised Long Short-Term Memory (LSTM) networks. For the daily resolution, three-dimensional tensors were generated from weekly windows of ninety-six time steps, normalized via min-max scaling; for the annual scale, seven-day sequences with equivalent rescaling were used. Architecture optimization was conducted through an exhaustive hyperparameter search that combined logarithmically spaced learning rates, graduated dropout values, and increasing batch sizes. Each configuration was evaluated using time-based cross-validation with expanding windows, and early stopping was applied to avoid convergence to inadequate local minima. The winning model was retrained on the entire training set and archived along with its hyperparameter record, thereby ensuring both reproducibility and traceability of the results.

4. Results

The experiments demonstrated an upward trend in estimation quality as the complexity of the evaluated models increased. In the initial phase, regression, both in its linear form and in the polynomial extensions, offered a basic fit that served only to establish the minimum performance benchmark; the marginal improvements achieved by adding higher-degree terms confirmed that the linear structure was insufficient to capture the intrinsic variability

of the photovoltaic signal. The shift to the SARIMA model introduced temporal differencing and explicit autocorrelation, which appreciably reduced the absolute magnitude of the errors. However, residual analysis revealed that a significant portion of the variance remained unexplained and that the confidence intervals retained excessive width, especially for horizons further from the present.

The most significant conceptual leap occurred with the adoption of the Random Forest. Thanks to internal randomization and the aggregation of multiple trees, this method was able to represent discontinuities and thresholds that the autoregressive models could not describe. As a result, a pronounced decrease in mean squared error and a substantial increase in explained variance were observed in both the daily and annual resolutions.

Finally, Long Short-Term Memory networks completed the methodological progression by incorporating internal memory capable of retaining medium- and long-term dependencies. Following exhaustive hyperparameter optimization, the most accurate predictions of the entire study and the tightest confidence intervals were obtained, with robust performance in both temporal horizons. Direct comparison with the Random Forest showed further error reductions and a substantial increase in the coefficient of determination, confirming that the ability of LSTMs to model full sequences is decisive in contexts dominated by complex seasonality and climatic variability. Overall, the results corroborated that the gradual incorporation of functional flexibility, from strict linearity to recurrent memory, constitutes the most effective path to improving the reliability of photovoltaic forecasts.

5. Conclusions

In summary, the comparative analysis conducted reveals that strict linearity is incapable of reproducing the seasonal complexity and intraday dynamics inherent to photovoltaic generation. Although the incorporation of autoregressive components reduced the absolute magnitude of errors, the linear nature of these models continued to hinder the proportion of explained variance. The subsequent transition to non-parametric ensemble methods confirmed that functional flexibility is an indispensable requirement, although the hierarchical partitioning of predictors exhibited certain limitations in capturing long-term dependencies.

The introduction of Long Short-Term Memory networks, rigorously optimized and trained separately for the daily and annual horizons, delivered a definitive qualitative leap. The dual

architecture enabled each network to specialize in its intended temporal scale, outperforming all other approaches in both accuracy and the stability of prediction intervals. It is therefore concluded that the combination of recurrent memory and specialized temporal segmentation constitutes the most suitable solution to meet the operational and strategic requirements of photovoltaic management.

Looking ahead, it is considered a priority to expand the information space through the integration of exogenous meteorological variables from satellites and ground stations, as well as to incorporate probabilistic quantification techniques that accompany each estimate with calibrated confidence intervals. Likewise, the implementation of incremental learning schemes would allow model parameters to be adapted in real time, reflecting evolving operating conditions without the need for exhaustive retraining. These research directions appear most promising to reinforce the robustness and adaptability of forecasting systems in scenarios of high solar penetration.

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

La transición hacia fuentes de energía renovable se perfila como un desafío de gran magnitud en el ámbito medioambiental y en la planificación de los sistemas eléctricos. Para que dicha transición resulte viable y eficiente, se requiere disponer de herramientas que permitan gestionar la intermitencia y variabilidad inherentes a las fuentes renovables; en particular, la energía solar depende estrechamente de la radiación incidente y las condiciones climatológicas, por lo que se considera imprescindible contar con modelos predictivos que anticipen con precisión su producción en el corto y medio plazo. De este modo, se facilita la toma de decisiones en el ámbito de la generación, se optimiza la programación de reservas y se reduce la incertidumbre asociada al incremento de potencia solar en la red.

El objetivo principal consiste del presente trabajo es predecir la producción de energía en centrales fotovoltaicas mediante la aplicación de algoritmos de aprendizaje automático. Con tal propósito, se ha recopilado un conjunto de datos históricos que abarca ocho años de producción eléctrica en una docena de instalaciones ubicadas en Calgary, Canadá. Dicha localidad ofrece un escenario representativo para el estudio de patrones estacionales, ya que las plantas se encuentran en emplazamientos muy cercanos que reciben prácticamente la misma radiación; por consiguiente, las doce instalaciones presentan comportamientos esencialmente idénticos y únicamente difieren en magnitud, de modo que su producción resulta proporcional en función de la capacidad nominal de cada una. El registro empleado contiene exclusivamente la serie temporal histórica de energía generada, sin integrar datos de irradiancia, temperatura u otras variables meteorológicas, de manera que los algoritmos se nutren únicamente del histórico de producción solar.

El análisis exploratorio de los datos ha permitido confirmar la existencia de ciclos estacionales muy pronunciados en la producción fotovoltaica: durante los meses de verano se alcanzan picos elevados de generación, mientras que en invierno la producción se reduce de forma drástica debido a la menor duración de las horas de sol y a la inclinación solar

reducida. Estas variaciones se observan de manera homogénea en las doce plantas, lo cual reafirma que resulta suficiente emplear un único modelo predictivo para todas las instalaciones, introduciendo un factor de escala que refleje la capacidad nominal de cada planta. En este contexto, se evita la redundancia de construir modelos independientes, pues las tendencias generales de captura de radiación son equivalentes en todas las plantas.

Para abordar la predicción de la producción solar, se han desarrollado y comparado los siguientes modelos: regresión lineal, regresión polinómica, regresión múltiple, SARIMA (Seasonal AutoRegressive Integrated Moving Average, por sus siglas en inglés; modelo autorregresivo integrado de medias móviles estacional); , Bosque Aleatorio y redes neuronales LSTM (Long Short-Term Memory). El modelo de regresión lineal ha servido para estimar tendencias generales de la serie temporal, aunque resulta insuficiente para capturar no linealidades complejas que emergen de los patrones diarios y estacionales. El algoritmo SARIMA ha permitido modelar dependencias temporales de la serie, si bien su alcance se restringe a procesos que puedan estacionar mediante diferenciación. Por su parte, el Bosque Aleatorio, al basarse en un ensamble de árboles de decisión, facilita la captura de relaciones no lineales y mitiga el sobreajuste. Por su parte, las redes neuronales LSTM, al conservar información relevante de contextos pasados, se revelan especialmente adecuadas para la predicción de series temporales con dependencias de largo plazo. Cabe destacar que todos los modelos se nutren únicamente de la dinámica interna de la producción histórica, sin requerir variables exógenas.

La validación de los modelos se implementó mediante una partición aleatoria 80-20 aplicada exclusivamente a las tareas de regresión, reservándose el 20% de las observaciones para evaluación. Posteriormente se fijó el año 2022 como conjunto de prueba, a fin de evitar fugas de información y reproducir las condiciones operativas reales del sistema. Esta estrategia permite estimar de manera consistente la capacidad de generalización de cada algoritmo sin incurrir en la complejidad adicional que implican técnicas de partición temporal más avanzadas; al mismo tiempo se garantiza que los errores de predicción calculados reflejen el desempeño de los modelos frente a datos no vistos y se respeta la secuencia cronológica de la serie en el conjunto de entrenamiento.

La relevancia de esta investigación se acentúa en el contexto canadiense, donde la energía solar contribuye aún con menos del 2% a la matriz energética regional, mientras que los combustibles fósiles representan alrededor del 20%. Canadá ha fijado la meta de cuadruplicar su capacidad instalada de energía solar para el año 2050, en consonancia con los compromisos de reducción de emisiones y el fomento de fuentes limpias. En tal escenario, disponer de modelos predictivos fiables se convierte en una necesidad prioritaria, pues facilita la integración de la generación fotovoltaica en el sistema eléctrico, optimiza la programación de reservas basadas en gas natural y minimiza los costes asociados a la regulación y equilibrado de la red.

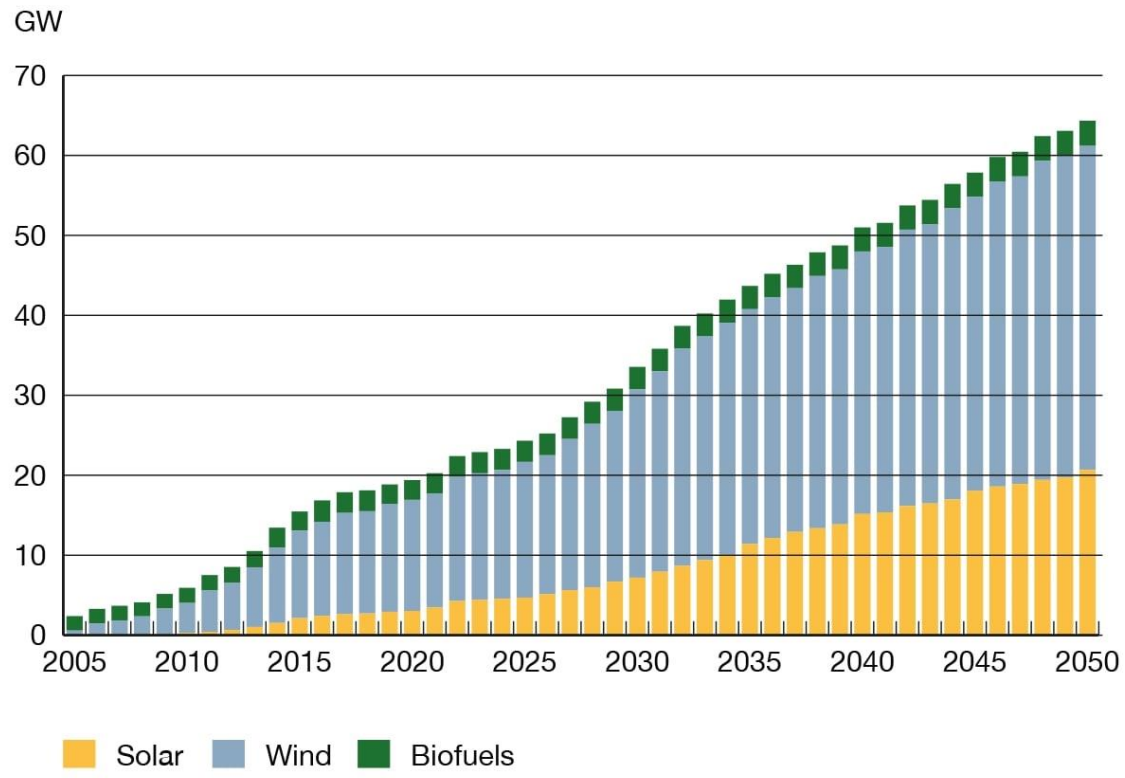

Figura 1.1: Proyección de la Potencia Instalada de Renovables no Hidroeléctricas en Canadá para 2050.

Fuente: Canada Energy Regulator 2020

Finalmente, el presente estudio se orienta a desarrollar y comparar diversas metodologías predictivas para estimar con mayor exactitud la producción solar en Calgary, Canadá. Al combinar modelos tradicionales (regresión lineal, SARIMA) con técnicas avanzadas de

ensamble (Bosque Aleatorio) y redes neuronales (LSTM), y al aplicar validaciones mediante división aleatoria y cronológica de los datos, se espera obtener un análisis comparativo que permita seleccionar, el algoritmo óptimo en términos de precisión, robustez y capacidad de generalización. Los resultados servirán como herramienta de gran valor para operadores de red y gestores de planta, contribuyendo al proceso de descarbonización y a la transición hacia un modelo energético sostenible.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

En este capítulo se describen las tecnologías empleadas para el desarrollo del proyecto de predicción de producción solar mediante aprendizaje automático. El entorno principal de trabajo ha sido Python, ejecutado a través de notebooks en Jupyter Lab ejecutados sobre servidores del ICAI. Las bibliotecas seleccionadas abarcan desde el procesamiento y análisis de datos hasta la implementación de modelos de series temporales y redes neuronales. A continuación, se detallan las herramientas y librerías que han permitido llevar a cabo todas las etapas del flujo de trabajo, desde la carga y limpieza del histórico de producción hasta el entrenamiento y evaluación de los algoritmos predictivos.

2.1 PYTHON EN JUPYTER LAB

Python es un lenguaje de programación de alto nivel, interpretado y multiparadigma, que ha alcanzado gran popularidad en el ámbito de la ciencia de datos y el aprendizaje automático. Su sintaxis clara y la amplia disponibilidad de librerías especializadas facilitan el desarrollo de prototipos y la experimentación iterativa. En el presente proyecto, se ha optado por emplear Python en notebooks de Jupyter Lab. Este entorno permite integrar código, visualizaciones y documentación de forma interactiva, lo cual resulta especialmente útil para acelerar el ciclo de análisis exploratorio, preprocesado y ajuste de modelos.

Los notebooks de Jupyter Lab se han desplegado y ejecutado en servidores de la Escuela Técnica Superior de Ingeniería del Instituto Católico de Artes e Industrias, con acceso mediante credenciales institucionales. En este entorno se dispone de alta capacidad de procesamiento para ejecutar los cálculos necesarios para la analítica de datos. Para cada etapa del análisis, desde la carga de ficheros CSV hasta la generación de gráficos de series temporales, se ha invocado una combinación de módulos de Python que se describen en los apartados siguientes.

Error! Reference source not found.

Figura 2.1: Logotipo de Python

Figura 2.2: Logotipo de JupyterHub

2.1.1 NUMPY

NumPy es una biblioteca fundamental para el cálculo numérico en Python, diseñada para el manejo eficiente de arreglos multidimensionales (ndarray) y la ejecución de operaciones matemáticas avanzadas. Gracias a NumPy se han podido cargar los datos de producción solar en estructuras de datos optimizadas, realizar operaciones vectorizadas para el cálculo de métricas y efectuar transformaciones básicas de manera eficiente. Su integración con otras librerías, como Pandas y Matplotlib, facilita la interoperabilidad en todo el flujo de trabajo.

Figura 2.3: Logotipo de Numpy

2.1.2 PANDAS

Pandas es una biblioteca orientada al análisis y manipulación de datos tabulares que proporciona estructuras de datos de alto rendimiento, como DataFrame y Series. En este proyecto, Pandas ha sido la herramienta principal para la lectura de los archivos históricos de producción, la limpieza de valores faltantes, la conversión de marcas temporales a índices de tipo datetime y la creación de variables adicionales. Asimismo, ha permitido la agrupación por intervalos (remuestreo) para resumir la producción diaria o mensual, así como la generación de subconjuntos de datos para entrenamiento y prueba. La extensa colección de funciones de Pandas ha resultado esencial para preparar las series temporales antes de alimentar los modelos predictivos.

Figura 2.4: Logotipo de Pandas

2.1.3 MATPLOTLIB

Matplotlib es la biblioteca estándar de Python para la generación de gráficos 2D, proporcionando una amplia variedad de tipos de visualización y opciones de personalización. En este estudio, Matplotlib ha permitido elaborar diagramas de líneas para mostrar la evolución de la producción en cada planta y gráficas de dispersión para inspeccionar relaciones entre variables temporales. Adicionalmente, se ha empleado el submódulo dates para formatear correctamente los ejes temporales y mostrar fechas de manera legible. De este modo, se ha facilitado la identificación de tendencias a largo plazo, patrones estacionales y posibles atípicos en el histórico de producción.

Figura 2.5: Logotipo de Matplotlib

2.1.4 STATSMODELS

Statsmodels es una biblioteca concebida para la estimación y prueba de modelos estadísticos, especialmente en el ámbito de las series temporales y la econometría. En el contexto de este

proyecto, se ha utilizado el módulo STL (Seasonal and Trend decomposition using Loess) para descomponer cada serie de producción en componentes de tendencia, estacionalidad y residuales. Posteriormente, el submódulo SARIMA (Seasonal Autoregressive Integrated Moving Average) ha servido para ajustar modelos autorregresivos con términos de diferenciación e integración, con el objetivo de capturar dependencias temporales en la evolución de la producción solar. La implementación de SARIMA en Statsmodels ha permitido probar diversas configuraciones de orden (p, d, q)(P, D, Q)s y evaluar la convergencia de los parámetros mediante criterios de información (AIC, BIC).

Figura 2.6: Logotipo de Statsmodels

2.1.5 SCIKIT-LEARN

Scikit-Learn es una biblioteca de aprendizaje automático de código abierto para Python que ofrece un conjunto de herramientas para el entrenamiento, validación y ajuste de modelos supervisados y no supervisados. Para la fase de regresión lineal se ha empleado la clase LinearRegression, la cual se entrena sobre matrices de características y vectores objetivo. Asimismo se ha utilizado StandardScaler para normalizar variables y mejorar la convergencia de ciertos algoritmos. En el apartado de Bosque Aleatorio se ha utilizado RandomForestRegressor para construir un ensamble de árboles de decisión que capture no linealidades. Para ambos casos también se recurrió a funciones de evaluación, tales como mean_squared_error, mean_absolute_error y r2_score, con el fin de comparar el desempeño en el conjunto de prueba. Por último, se ha aprovechado train_test_split para generar la partición aleatoria 80%-20% de los datos y GridSearchCV para optimizar hiperparámetros en Bosque Aleatorio.

Figura 2.7: Logotipo de Scikit-Learn

2.1.6 TENSORFLOW Y KERAS

Para la implementación de redes neuronales de tipo LSTM (Long Short-Term Memory), se ha empleado TensorFlow junto con su API de alto nivel, Keras. Concretamente, se ha construido una arquitectura secuencial que consta de una o dos capas LSTM seguidas de capas densas, intercalando capas de Dropout para mitigar el sobreajuste. Los datos de entrada se han escalado con MinMaxScaler, de modo que los valores de producción oscilen en el rango [0, 1], lo cual facilita el entrenamiento de la red. La función de pérdida utilizada ha sido el error cuadrático medio y el optimizador Adam se ha seleccionado para actualizar los pesos. Durante el proceso de ajuste, se han fijado valores de batch_size y número de épocas basados en pruebas previas, empleando conjuntos de validación internos para monitorizar la evolución de la pérdida y detener el entrenamiento cuando no se observaba mejora.

Figura 2.8: Logotipo de TensorFlow

Figura 2.9: Logotipo de Keras

Capítulo 3. ESTADO DE LA CUESTIÓN

La predicción de la producción de energía solar mediante técnicas de aprendizaje automático ha experimentado un notable crecimiento en la última década, motivada por la necesidad de gestionar la variabilidad intrínseca de la generación fotovoltaica y optimizar la integración de esta fuente en los sistemas eléctricos. En la literatura internacional, la mayoría de los estudios combina series temporales de producción con variables meteorológicas (irradiancia, temperatura, velocidad del viento, entre otras) con el fin de capturar con mayor precisión las fluctuaciones diarias y estacionales de la generación [1], [2]. No obstante, pocos trabajos han explorado el uso exclusivo de datos históricos de producción sin apoyo de mediciones externas, lo cual constituye una línea de investigación menos explotada pero relevante en escenarios donde la disponibilidad de información meteorológica es limitada o se desea contar con sistemas predictivos más sencillos de desplegar.

Los modelos de regresión lineal han servido tradicionalmente como punto de partida por su simplicidad y capacidad para estimar tendencias generales de la serie temporal [3]. Sin embargo, ante la existencia de relaciones no lineales en la evolución de la producción, provocadas por cambios bruscos en las condiciones ambientales y dinámicas de los paneles, su desempeño suele quedar por debajo del de algoritmos más complejos [2]. En esta línea, los modelos autorregresivos integrados de media móvil estacionales (SARIMA) han sido empleados con éxito para series estrechamente estacionarias o sometidas a diferenciación previa, mostrando un comportamiento robusto en horizontes de corto plazo cuando la serie presenta ciclos estacionales bien marcados [4]. No obstante, la principal limitación de SARIMA radica en su suposición implícita de estacionariedad y en la necesidad de seleccionar manualmente los órdenes de autorregresión, diferenciación e integración, lo que exige procesos iterativos de prueba y error para ajustar los parámetros $(p, d, q)(P, D, Q)_s$.

Los enfoques basados en árboles de decisión, en particular Bosque Aleatorio y XGBoost, se han consolidado como alternativas potentes gracias a su habilidad para capturar relaciones

no lineales y manejar variables exógenas sin requerir procesos exhaustivos de ingeniería de características [5], [6]. En pronósticos de irradiancia y potencia diaria, estos algoritmos han alcanzado valores de R^2 cercanos a 0.9 y errores relativos inferiores al 15%, superando con frecuencia a modelos de regresión simples y proporcionando tiempos de entrenamiento relativamente bajos [5]. La ventaja adicional de los métodos de ensamble es su menor sensibilidad al sobreajuste, aunque su ejecución puede demandar mayor potencia de cómputo y ajustes de hiperparámetros a través de técnicas de validación cruzada o búsqueda en malla [6].

Por su parte, las redes neuronales recurrentes de tipo LSTM (Long Short Term Memory) han demostrado gran capacidad para modelar dependencias temporales complejas al conservar información de contexto a lo largo de múltiples pasos temporales [7], [8]. Cuando se disponen de historiales de varios años con resoluciones horarias o subhorarias, las LSTM pueden aprender patrones diarios, estacionales y tendencias de largo plazo, lo que resulta especialmente útil para pronósticos de horizonte cercano, por ejemplo de una hora a un día [8]. Sin embargo, su entrenamiento suele requerir más tiempo, así como un preprocesado dedicado para escalado de datos y definición de ventanas de observación [7]. Además, la arquitectura de la red debe ser calibrada cuidadosamente en cuanto a cantidad de capas, neuronas, tasa de aprendizaje y criterios de detención temprana, pues un sobredimensionamiento de la red puede inducir sobreajuste mientras que una topología insuficiente dificulta la captura de patrones complejos [8].

Respecto a la validación de los modelos, la literatura pone especial énfasis en respetar la naturaleza secuencial de las series temporales mediante partición temporal progresiva o validación cruzada anidada, que evitan la contaminación de datos futuros en el conjunto de entrenamiento [9]. Sin embargo, cuando se dispone de suficientes datos históricos, por ejemplo más de cinco años, a veces se opta por una división convencional en conjuntos de entrenamiento y prueba, computando un 80% para entrenamiento y un 20% para prueba, que permita estimar de manera representativa la capacidad de generalización sin incurrir en complejidades adicionales [4]. Esta estrategia, aunque más sencilla, exige garantizar que la muestra de prueba comprenda periodos representativos de las distintas estaciones y

condiciones atípicas, como tormentas de nieve o días nublados extremos, de modo que los errores calculados reflejen el comportamiento real en la producción futura.

En relación con los datos empleados, la incorporación de variables meteorológicas suele mejorar notablemente el desempeño de XGBoost y LSTM, al proporcionar información exógena que los modelos pueden explotar para anticipar caídas de irradiancia o picos de temperatura que afectan la eficiencia fotovoltaica [2], [9]. No obstante, la dependencia de datos meteorológicos de alta calidad puede representar un obstáculo en zonas rurales o en emplazamientos donde no existen estaciones de medida confiables. Por tanto, algunos estudios han centrado su atención en utilizar exclusivamente series de producción, mediante técnicas de descomposición estacional y extracción de retardos como únicas características de entrada, logrando resultados aceptables cuando la serie dispone de registros limpios y completos [4]. En este contexto, los modelos SARIMA y LSTM pueden capturar las dinámicas internas de la serie sin necesidad de variables externas, aunque su precisión general tiende a ser inferior a la de sus contrapartes que incluyen variables meteorológicas [8].

El estudio realizado se enfoca precisamente en la predicción a partir de series históricas de producción solar sin recurrir a datos de irradiancia o condiciones meteorológicas, lo que lo diferencia de la mayor parte de la bibliografía. Este enfoque plantea el desafío de identificar hasta qué punto los modelos pueden inferir patrones estacionales y exógenos a partir de la propia dinámica de generación. La literatura revisada sugiere que, en ausencia de variables externas, los modelos basados en aprendizaje profundo, en particular las LSTM, tienen una ventaja al modelar dependencias de largo plazo, aunque la calidad de los resultados dependerá en gran medida de la disponibilidad de suficientes ciclos anuales en los datos para que la red neuronal pueda aprender las variaciones estacionales [7]. Asimismo, se anticipa que los algoritmos de ensamble como Bosque Aleatorio ofrecerán un rendimiento competitivo al aprender de los retardos de producción, por ejemplo valores de las últimas 48 horas y de variables temporales derivadas como día del año u hora, si bien su precisión será menor que la de modelos que incluyan variables meteorológicas [6].

En el caso específico de la región de Calgary, Canadá, el clima presenta características particulares: inviernos prolongados con nieve y horas de luz muy reducidas, veranos de alta irradiancia y cambios meteorológicos bruscos. Estas condiciones introducen eventos atípicos, por ejemplo producción nula debido a acumulación de nieve, que la mayoría de los estudios no aborda explícitamente pues se centran en climas templados o subtropicales [5], [2]. La experiencia empírica indica que la ausencia de eventos extremos en los datos de entrenamiento dificulta su predicción. En consecuencia, el empleo de un histórico amplio que incluya varios inviernos con acumulación de nieve permitirá a los modelos reconocer secuencias de caídas abruptas y periodos de baja generación propios de latitudes elevadas. Además, la latitud de Calgary implica ángulos solares muy reducidos en invierno, lo que modifica la forma de las curvas diarias de producción y exige que los algoritmos ajusten sus predicciones a horizontes con muy pocas horas de luz; esto constituye un reto que modelos como SARIMA afrontan con limitaciones y que las LSTM podrían capturar si se incorporan suficientes patrones históricos [4], [7].

Por otra parte, la mayoría de estudios entrena modelos para cada instalación de forma independiente o en contextos de varias plantas tratadas como series separadas [8], [5]. Esto contrasta con enfoques que buscan un modelo global para varias plantas, añadiendo una variable indicadora de instalación para que el algoritmo distinga condiciones locales. En el presente trabajo se parte de la premisa de que las doce centrales fotovoltaicas ubicadas en Calgary presentan comportamientos esencialmente proporcionales, dado que reciben aproximadamente la misma radiación, y solo difieren en magnitud según su capacidad nominal. Por ello, se adoptará un único modelo predictivo que incluya un factor de escala para cada planta, reduciendo la complejidad de mantener doce modelos separados y aprovechando la homogeneidad del recurso solar en la región.

En síntesis, el estado de la cuestión revela que, si bien existe abundante literatura sobre predicción solar que incluye variables meteorológicas o múltiples centrales, hay una laguna en estudios que analicen exclusivamente series de producción en climas de latitud alta con condiciones invernales extremas. Los modelos de regresión lineal, SARIMA, Bosque Aleatorio y LSTM constituyen un conjunto de enfoques probados en contextos diversos,

pero su comportamiento en ausencia de datos meteorológicos y ante eventos atípicos como nieve o días muy cortos aún no ha sido explorado en profundidad para Canadá. Este trabajo, al centrarse en el uso de datos históricos de producción, aportará una contribución novedosa al demostrar hasta qué punto estos algoritmos pueden inferir patrones complejos solo a partir del flujo de generación, así como establecer las limitaciones y oportunidades para futuros estudios que incluyan variables adicionales o enfoques híbridos.

3.1 OTROS ANÁLISIS DE PRODUCCIÓN EN EL MERCADO

En el ámbito profesional se dispone de soluciones de predicción de generación fotovoltaica ofertadas casi exclusivamente como servicios de software en la nube que integran datos meteorológicos procedentes tanto de satélites como de modelos numéricos de predicción (NWP, por sus siglas en inglés) y técnicas de aprendizaje automático. Estas plataformas operan mediante un flujo de trabajo de dos fases: en una primera etapa se genera un pronóstico físico basado en modelos de cielo despejado y salidas de NWP, y en una segunda etapa se realiza una corrección estadística o de aprendizaje automático, calibrada con datos de telemedida de la instalación cuando están disponibles.

Entre las APIs destinadas a grandes operadores y utilities destacan cuatro proveedores. Solcast emplea visión por computador sobre imágenes satelitales junto con algoritmos de aprendizaje automático para generar más de seiscientas millones de predicciones por hora, con horizontes temporales de cinco minutos a catorce días y acceso a veinte años de datos históricos [10][11]. SolarAnywhere Forecast, de Clean Power Research, fusiona métodos físicos y estadísticos para mercados day-ahead y operaciones de despacho; requiere obligatoriamente datos de irradiancia satelital y salidas de NWP, aunque permite ajustes específicos para cada emplazamiento [12][13]. Solargis Forecast utiliza vectores de movimiento de nubes para nowcasting con resoluciones de cinco minutos y modelos NWP para horizontes de hasta catorce días, ofreciendo además predicciones probabilísticas P10-P90 que apoyan la gestión de riesgos en trading y financiación de proyectos [14][15][16][17]. Meteocontrol Forecast combina información de NWP y temperatura con

técnicas de machine learning para pronósticos de quince minutos a siete días, calibrando sus modelos con telemedidas de plantas de hasta 100 MWp; sin embargo, no funciona exclusivamente con series históricas de producción [18][19][20].

De forma complementaria, las suites de gestión de activos y mantenimiento integran funcionalidades de predicción avanzada. IBM Renewables Forecasting, incluida en la Environmental Intelligence Suite y en Maximo Renewables, combina analítica avanzada, sensores IoT y datos meteorológicos de alta resolución para optimizar las tareas de operación y mantenimiento de flotas multi-tecnología (solar y eólica) [21][22][23]. Estas soluciones permiten anticipar desviaciones de rendimiento y programar acciones preventivas, aunque su complejidad y dependencia de datos externos incrementan los requisitos de infraestructura y licencias.

En el ámbito de inversores y plataformas de operación y mantenimiento se han incorporado funciones de pronóstico embebidas. Un ejemplo es la funcionalidad Weather Guard - Home Backup de SolarEdge, que ajusta la carga de baterías en función de alertas meteorológicas graves; este sistema no genera un pronóstico continuo de potencias sino que actúa sobre eventos detectados en flujos de datos externos [24].

En el entorno de código abierto sobresalen Quartz Solar Forecast, de OpenClimateFix, una librería Python de uso gratuito para pronósticos de cero a cuarenta y ocho horas mediante modelos de visión profunda que necesita mapas satelitales y datos meteorológicos externos [25], y pvlib-python, un conjunto de herramientas científicas para el modelado físico de sistemas fotovoltaicos que puede entrenarse con datos propios pero precisa entradas de irradiancia o NWP para obtener resultados rigurosos.

Este catálogo de servicios y librerías ilustra la madurez del mercado profesional de predicción fotovoltaica y pone de manifiesto dos retos principales: la dependencia de fuentes meteorológicas externas y la complejidad en la calibración local, que encarecen la implantación en ubicaciones remotas o de baja instrumentación. Estos desafíos subrayan la oportunidad de investigar enfoques endógenos basados únicamente en históricos de

producción, con el objetivo de reducir costes de infraestructura y licencias, mejorar la escalabilidad en microrredes y abrir nuevas posibilidades en mercados emergentes.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

En el ámbito de la generación fotovoltaica, la disponibilidad de predicciones fiables de producción solar se ha convertido en un requisito fundamental para optimizar la operación de las redes eléctricas y reducir los costes asociados a la gestión de la intermitencia. A pesar de que numerosos estudios emplean variables meteorológicas junto con series históricas de generación para mejorar la precisión de los pronósticos, en muchos emplazamientos rurales o de latitudes elevadas no existe una arquitectura consolidada de estaciones meteorológicas que garantice datos de irradiancia, temperatura o velocidad del viento de alta frecuencia y calidad. En consecuencia, se presenta una necesidad real de desarrollar soluciones predictivas basadas exclusivamente en datos históricos de producción, sin depender de mediciones externas susceptibles a fallos o a elevados costes de instalación y mantenimiento.

Desde el punto de vista técnico, entrenar un modelo único para varias plantas ubicadas en un área geográfica homogénea, como es el caso de las centrales fotovoltaicas de Calgary, permite aprovechar la información agregada y reducir la complejidad operativa inherente a la gestión de modelos independientes para cada instalación. Al asumir que todas las plantas reciben aproximadamente la misma radiación, tan solo se introduce un factor de escala en función de la capacidad nominal de cada planta. De este modo, se simplifica el flujo de entrenamiento, despliegue y actualización del modelo. Esta estrategia técnica facilita el mantenimiento del sistema a largo plazo, pues cuando se incorporan nuevas centrales o se expanden las ya existentes basta con recabar el histórico de generación y reentrenar el modelo global sin necesidad de recopilar datos meteorológicos adicionales ni construir arquitecturas específicas por planta.

Desde la perspectiva de mercado, la demanda de herramientas predictivas adecuadas a entornos donde los recursos meteorológicos no están disponibles o resultan costosos de

mantener es creciente. Las empresas operadoras de parques solares y los gestores de redes eléctricas requieren soluciones sencillas de implementar, que reduzcan el tiempo de puesta en marcha y minimicen la inversión inicial en equipamiento. Al basarse exclusivamente en series temporales de producción, la solución propuesta elimina la necesidad de adquirir equipamiento adicional para medición de irradiancia y climatología, lo que reduce los costes de capital y operativos. De esta forma, se ofrece un producto escalable y asequible para entidades que cuentan con un volumen importante de datos históricos pero carecen de infraestructura de monitoreo meteorológico.

La predicción de producción mediante aprendizaje automático permite anticipar variaciones estacionales y atípicas, por ejemplo pérdidas de generación debidas a acumulación de nieve o días extremadamente nublados. Esta capacidad contribuye a optimizar la programación de reservas en centrales de respaldo, especialmente las que funcionan con gas natural. En Canadá, donde la meta nacional es cuadruplicar la capacidad instalada de energía solar para 2050, se prevé un incremento significativo de parques fotovoltaicos que entrarán en operación. La solución propuesta aporta valor agregado al ofrecer pronósticos con un grado de precisión comparable al de modelos multivariantes, pero sin depender de fuentes externas de datos. Este aspecto permitirá a las compañías energéticas escalar rápidamente las capacidades de predicción a medida que crece el parque solar, sin incurrir en costes adicionales por infraestructura de medición meteorológica.

En términos de inversión, el producto que se propone presenta varias ventajas competitivas. Su implementación resulta rápida, pues al emplear únicamente datos históricos de producción no es necesario instalar sensores ni adquirir servicios de datos meteorológicos. El mantenimiento se reduce, ya que las actualizaciones y reentrenamientos del modelo requieren solo la incorporación de nuevas observaciones de generación sin necesidad de calibrar equipos externos. Asimismo, la metodología de modelo único con factor de escala puede aplicarse a otras regiones con condiciones solares homogéneas, ampliando así el mercado potencial. La mejora en la precisión de la predicción, frente a métodos basados únicamente en regresiones simples, permite reducir los costes asociados a la regulación de

la red y a la programación de centrales de respaldo, lo cual repercute directamente en la rentabilidad de los activos fotovoltaicos.

En definitiva, el desarrollo del proyecto responde a la carencia de soluciones predictivas ligeras y asequibles para entornos donde no se dispone de datos meteorológicos fiables. A su vez, satisface la demanda de operadores de plantas solares y gestores de redes que necesitan contar con un pronóstico de producción riguroso para optimizar la planificación de la generación y minimizar los riesgos financieros. Desde el punto de vista técnico y de mercado, la propuesta ofrece una arquitectura eficiente, de fácil mantenimiento y con alto potencial de adopción en regiones con condiciones climáticas extremas y latitudes elevadas, donde la gestión de eventos atípicos como la nieve o los días muy cortos resulta crítica para garantizar la estabilidad del sistema eléctrico.

4.2 OBJETIVOS

4.2.1 OBJETIVO PRINCIPAL: PREDICCIÓN DE LA GENERACIÓN ENERGÉTICA EN CENTRALES FOTOVOLTAICAS

Desarrollar un modelo de predicción de la generación eléctrica en plantas fotovoltaicas basado exclusivamente en series temporales históricas de producción, de manera que alcance un nivel de precisión comparable al de sistemas que incorporan variables meteorológicas externas, demostrando así la viabilidad de soluciones ligeras y escalables sin dependencia de infraestructura de monitoreo externo.

4.2.2 OBJETIVO SECUNDARIO: COMPARACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

Realizar una evaluación rigurosa y homogénea del desempeño de diversos métodos aplicables a series temporales de generación solar, sustentada en métricas estándar de error (MAE, RMSE, R^2) y en pruebas de robustez bajo condiciones adversas como baja irradiación o acumulación de nieve. Esta comparación incluirá modelos de regresión lineal como referencia de base, modelos autorregresivos integrados estacionales (SARIMA) para

capturar dependencias temporales clásicas, enfoques de ensamble mediante Bosque Aleatorio para abordar relaciones no lineales y redes neuronales LSTM para retener patrones estacionales y eventos extremos.

4.2.3 OBJETIVO DE VALIDACIÓN MULTIPLANTA

Validar la aplicabilidad de un esquema de modelo único con factor de escala que permita generar pronósticos para múltiples plantas ubicadas en una zona de radiación homogénea, garantizando que el mismo algoritmo se adapte automáticamente a cada instalación según su capacidad nominal sin necesidad de ajustar datos externos.

4.2.4 OBJETIVO DE CONCLUSIONES Y LÍNEAS FUTURAS

Elaborar un documento de conclusiones que sintetice los hallazgos, determine el algoritmo óptimo en función de criterios de precisión, robustez y escalabilidad, y proponga líneas de investigación futura en el campo de la predicción fotovoltaica.

4.3 METODOLOGÍA

La estrategia de trabajo empleada combina un enfoque secuencial y estructurado con ciclos iterativos y adaptativos, de modo que se garantiza tanto la previsibilidad en el cumplimiento de plazos como la capacidad de respuesta ante cambios en los requisitos. Con este planteamiento se dispone de un marco mixto que integra principios del modelo en cascada y de las metodologías ágiles, favoreciendo la eficiencia en la gestión y la calidad del resultado final.

La planificación se estructura en siete fases diferenciadas, cada una con objetivos y entregables claramente definidos y representados en el cronograma. En la primera, se establece el alcance del proyecto, se definen los recursos necesarios y se fijan los hitos de control. En la segunda, el estudio del estado del arte recopila y sintetiza la literatura relevante para fundamentar la metodología. La tercera garantiza la obtención y limpieza de las fuentes de información. En la cuarta, el análisis exploratorio aplica técnicas estadísticas y de

visualización para identificar patrones y anomalías. La quinta abarca la investigación de herramientas y el desarrollo de prototipos, evaluando plataformas como JupyterHub y modelos como SARIMA, Bosque Aleatorio o LSTM. Durante la sexta fase, se integra el sistema en un entorno operativo y se verifica el desempeño de los componentes mediante métricas predefinidas. Finalmente, en la séptima fase se consolidan los resultados, se elabora la memoria académica y se prepara la presentación y defensa ante tribunal. Cada fase incorpora puntos de control intermedios, revisiones de avance y criterios de aceptación que permiten detectar desviaciones en tiempo real y ajustar el plan sin comprometer ni los plazos ni la calidad global del proyecto.

No obstante, dado que la naturaleza del proyecto puede implicar modificaciones durante su ejecución, se incorporan ciclos de trabajo de dos semanas en los que se revisan los resultados obtenidos al término de cada iteración, se contrastan con los interesados y se redefinen los requisitos de acuerdo con las observaciones recibidas. Asimismo, se programan reuniones diarias breves de seguimiento, en las que se exponen los progresos alcanzados, se identifican impedimentos y se establecen las prioridades para la jornada siguiente, garantizando así la detección ágil de bloqueos.

Por consiguiente, el marco de trabajo diseñado combina la planificación exhaustiva y la previsión propias del modelo en cascada con la flexibilidad y el bucle de retroalimentación característicos de los métodos ágiles, de manera que las fases secuenciales aportan estabilidad y visibilidad a largo plazo, mientras que las iteraciones cortas facilitan la incorporación continua de ajustes. De este modo se equilibra el control sobre el alcance y los plazos con la capacidad de adaptación al cambio, redundando en una mayor efectividad del equipo y en la calidad final del producto.

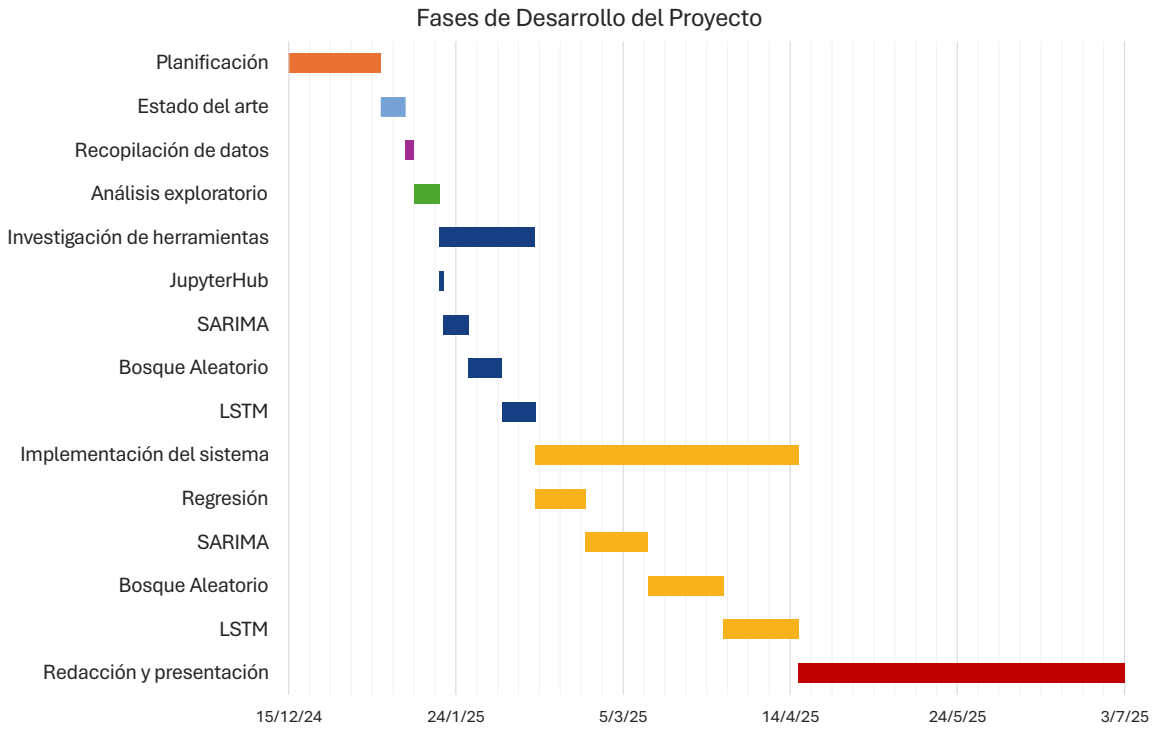

Figura 4.1: Cronograma de Fases de Desarrollo del Proyecto

4.4 ESTIMACIÓN ECONÓMICA

El análisis económico del proyecto se ha efectuado considerando las horas de desarrollo efectivas y los recursos disponibles, con el objeto de demostrar la viabilidad y el bajo coste de la solución propuesta. En primer lugar, la fase de implementación ha requerido una dedicación total de 480h distribuidas a lo largo de 7 meses. Aplicando una tarifa de 50€/h, el coste directo asociado a la mano de obra asciende a 24.000€.

El equipamiento hardware se ha estimado mediante el uso de una instancia pequeña de Amazon EC2 con una tarifa aproximada de 20€/día durante 60 días, lo que implica un desembolso de 1.200€. Esta infraestructura en la nube cubre tanto el entrenamiento inicial como las pruebas de carga, por lo que no resulta necesaria la adquisición de servidores físicos ni de dispositivos adicionales.

Todos los componentes de software utilizados pertenecen al ámbito del código abierto, incluyendo el lenguaje Python y las bibliotecas de aprendizaje automático TensorFlow, Keras y Scikit-Learn. Esta práctica elimina por completo los costes de licencias y actualizaciones, contribuyendo de manera significativa a la reducción de la inversión inicial y a la sostenibilidad a largo plazo del sistema.

Para garantizar la continuidad y evolución de la solución durante los cinco primeros años de operación, se ha previsto un plan de mantenimiento correctivo y evolutivo con una dedicación de 8h mensuales, equivalente a 96h al año y a 480h en el horizonte de cinco años. Aplicando la misma tarifa de 50€/h, el coste acumulado de mantenimiento asciende a 24.000€.

En consecuencia, el desembolso económico total, sumando mano de obra (24.000€), hardware en la nube (1.200€) y mantenimiento a cinco años (24.000€), se sitúa en 49.200€.

Esta propuesta es diferencial porque se aprovechan integralmente los datos y dispositivos ya existentes, eliminando la necesidad de inversiones adicionales en hardware específico y prescindiendo de los procesos complejos de instalación y calibración de sensores nuevos. Asimismo, la integración con la infraestructura actual permite acelerar la puesta en marcha, reducir los tiempos de implementación y minimizar el riesgo de interrupciones en la operación de la planta. Además, se optimiza el uso de recursos y se facilitan las futuras ampliaciones al emplear un ecosistema tecnológico ya validado, garantizando la adaptabilidad y escalabilidad del sistema sin incurrir en gastos adicionales.

Capítulo 5. SISTEMA DESARROLLADO

5.1 PLANTEAMIENTO COMPARATIVO Y ESQUEMA DE VALIDACIÓN

En lo que sigue se desarrollará una comparación sistemática de cuatro categorías de modelos orientados a la predicción de series temporales de producción fotovoltaica: Regresión Lineal, SARIMA, Bosque Aleatorio y Long Short-Term Memory. A fin de garantizar una evaluación homogénea, todas las técnicas se sometieron a un protocolo de validación cruzada que difiere únicamente en el criterio de particionado, tal como se detalla a continuación.

Para la familia de regresiones (lineal, polinómica y múltiples) se consideró apropiado un particionado aleatorio, de forma que el 80% de las observaciones sirvió para el entrenamiento y el 20% restante se reservó como conjunto de prueba. Dicha proporción permitió estimar los parámetros con un volumen de datos suficiente y, a la vez, disponer de un subconjunto independiente que reflejara la variabilidad global de la serie.

El resto de los modelos, cuya naturaleza requiere preservar la secuencia temporal, se validó mediante un corte cronológico fijo: todas las observaciones anteriores a 2022 se emplearon en el entrenamiento y las correspondientes a 2022 se utilizaron para la prueba. Esta estrategia evita la fuga de información desde el futuro hacia el pasado y reproduce con fidelidad el escenario operativo en el que se aplicarán las predicciones.

Las métricas de desempeño seleccionadas fueron el Error Cuadrático Medio, el Error Absoluto Medio y el coeficiente de determinación. Cada subsección presentará, en primer término, la introducción conceptual del método, a continuación la configuración específica del experimento y, por último, la discusión de resultados basada en las métricas mencionadas.

Con esta estructura se busca exponer, desde la apertura del capítulo, una visión clara del alcance comparativo y de la lógica de validación que sustenta las conclusiones que se exponen en los apartados sucesivos.

5.2 CONJUNTO DE DATOS Y PROCEDIMIENTO DE VALIDACIÓN

El análisis se sustenta en series temporales de generación fotovoltaica con registro cada 15 minutos entre 2016 y 2023 de doce plantas fotovoltaicas de Calgary, Canadá. Tras la ingestión del archivo CSV, se aplicó interpolación lineal bidireccional para cubrir automáticamente los valores nulos.

El protocolo de particionado se diseñó de acuerdo con la naturaleza de cada familia de modelos. Para los métodos de regresión lineal, regresión polinómica y regresiones lineales múltiples con ventana deslizante se recurrió a un muestreo estratificado en proporción 80% entrenamiento y 20% prueba, de modo que las distintas estaciones del año quedaran equilibradamente representadas en ambos subconjuntos. En cambio, para los enfoques que explotan la dependencia temporal, modelos SARIMA, Bosque Aleatorio y redes neuronales LSTM, se optó por un corte cronológico: todos los datos hasta el 31 de diciembre de 2021 se emplearon para ajustar los modelos y las 365 jornadas de 2022 se reservaron exclusivamente para la evaluación externa, de manera que se evitó la fuga de información y se replicaron las condiciones operativas reales de funcionamiento.

El flujo de trabajo de entrenamiento y validación se implementó con scikit-learn en las regresiones, en SARIMA y en Bosque Aleatorio; y con TensorFlow-Keras en las redes LSTM, asegurando la trazabilidad de las configuraciones y la reproducibilidad de los experimentos.

5.3 *REGRESIÓN LINEAL*

5.3.1 INTRODUCCIÓN

El análisis comenzó con la necesidad de predecir series temporales utilizando un enfoque básico, que fue evolucionando a medida que se identificaron limitaciones en las técnicas aplicadas. El proceso se inició con un modelo de regresión lineal simple, que fue útil como punto de partida, pero rápidamente se hizo evidente que no capturaba las complejas relaciones temporales inherentes a los datos. A medida que se exploraron técnicas más avanzadas, se introdujeron nuevas estrategias como la regresión polinómica y el modelo de regresiones múltiples, que sirvieron como avances metodológicos pero mostraron una muy limitada mejora.

5.3.2 REGRESIÓN LINEAL

La regresión lineal es una de las técnicas más simples para abordar problemas de predicción. Este modelo trató de predecir los valores futuros de la serie temporal a partir de los valores pasados, bajo la premisa de que existe una relación lineal entre las variables. Aunque este modelo fue sencillo de implementar y proporcionó un primer vistazo a las dinámicas de la serie temporal, rápidamente mostró sus limitaciones. La regresión lineal no es capaz de capturar adecuadamente las relaciones no lineales entre los datos y las interacciones temporales más complejas. A pesar de su simplicidad y bajo costo computacional, este enfoque no pudo representar la relación entre las variables, lo que llevó a explorar modelos más complejos.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Regresión Lineal	79,174	6268,471	55,013	0,000

Tabla 5.1: Métricas de Regresión Lineal

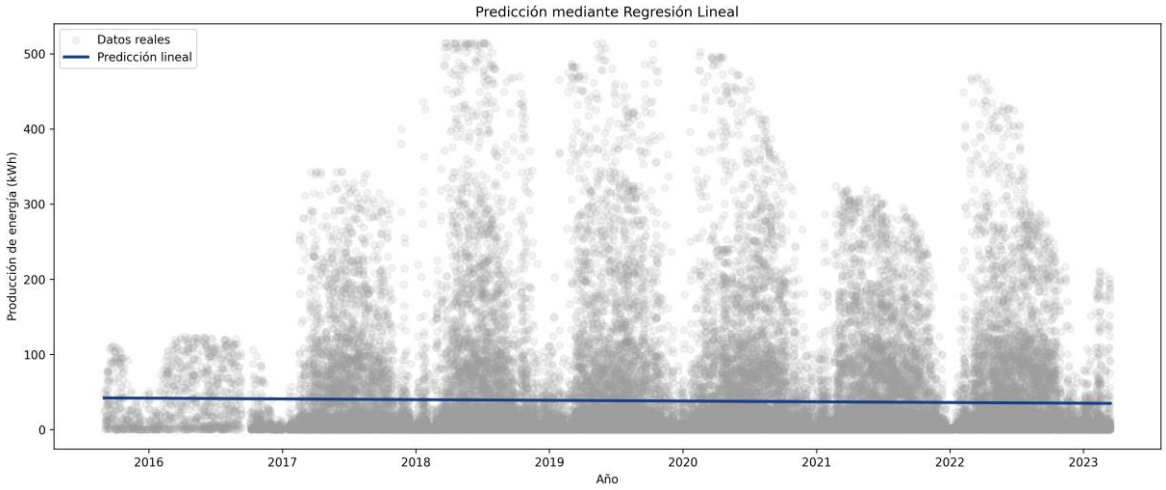

Figura 5.1: Predicción mediante Regresión Lineal de la Serie Temporal

El ajuste lineal se formuló bajo la hipótesis de que la producción evoluciona con una tendencia constante; por consiguiente, se asumió que los residuos representan únicamente ruido blanco. Al no incorporar términos de estacionalidad ni de interacción, el modelo se vio incapaz de reproducir las oscilaciones diarias y estacionales inherentes a la serie; en consecuencia, el RMSE alcanzó 79,174 kWh y el coeficiente de determinación permaneció completamente nulo. La regresión lineal sirve como línea de base para la comparación con el resto de modelos, dado que su simplicidad conceptual permite entender claramente las mejoras progresivas.

5.3.3 REGRESIÓN POLINÓMICA

Ante las limitaciones de la regresión lineal se introdujo la regresión polinómica, que permite modelar relaciones no lineales. Este enfoque amplía el modelo lineal incorporando términos de grado superior, lo que permite una mayor flexibilidad en el ajuste de los datos. La regresión polinómica es capaz de captar las curvas y las tendencias no lineales en los datos de manera más efectiva que el enfoque lineal. Sin embargo, tampoco logró capturar las complejas dinámicas de las series temporales. Las predicciones a largo plazo seguían siendo imprecisas y el modelo aún presentaba limitaciones para abordar las interacciones temporales a largo plazo.

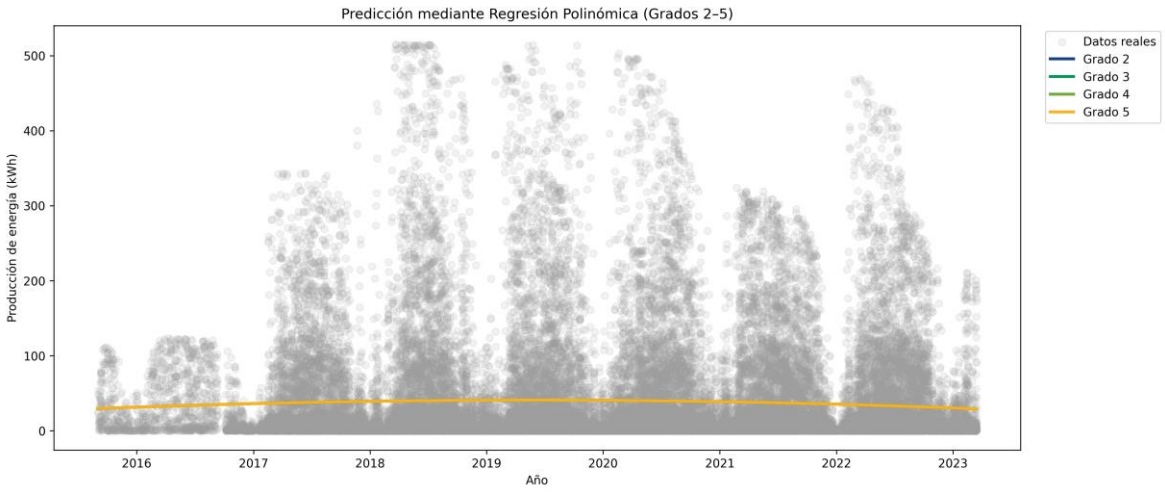

Figura 5.2: Predicción mediante Regresión Polinómica de la Serie Temporal

Los coeficientes de las regresiones polinómicas de todos los grados explorados (2, 3, 4 y 5) guardaban gran relación todos entre sí, y también con la regresión lineal. Esto nos dice que a pesar de proporcionar mayor complejidad al modelo el resultado era siempre el mismo, indicando que utilizar una sola regresión no es el enfoque adecuado para un problema de esta naturaleza.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Grado 2	61,341	3762,701	37,855	0,002
Grado 3	61,341	3762,698	37,855	0,002
Grado 4	61,341	3762,695	37,855	0,002
Grado 5	61,341	3762,692	37,855	0,002

Tabla 5.2: Métricas de Regresiones Polinómicas

El RMSE se redujo únicamente hasta 61,341 kWh y R^2 se incrementó a 0,002; resultando en métricas prácticamente idénticas independientemente del grado del polinomio. Se evidencia, por consiguiente, que la serie no presenta una relación polinómica simple con el tiempo y que la complejidad añadida no repercute en una mejora sustancial del poder explicativo.

5.3.4 REGRESIONES LINEALES MÚLTIPLES

El siguiente modelo implementó la estrategia de ajustar regresiones lineales independientes sobre subconjuntos de datos, reemplazando el uso de la totalidad del historial disponible. En particular, se estableció una regresión lineal específica para cada semana del año. De este modo, los coeficientes resultantes caracterizan la tendencia local de cada intervalo semanal y, al integrarse, permiten inferir la morfología global del año. Este procedimiento constituye el primer enfoque que se adapta explícitamente a la naturaleza temporal de la serie, pues posibilita al modelo capturar patrones particulares dentro de un horizonte representativo y pertinente para la predicción.

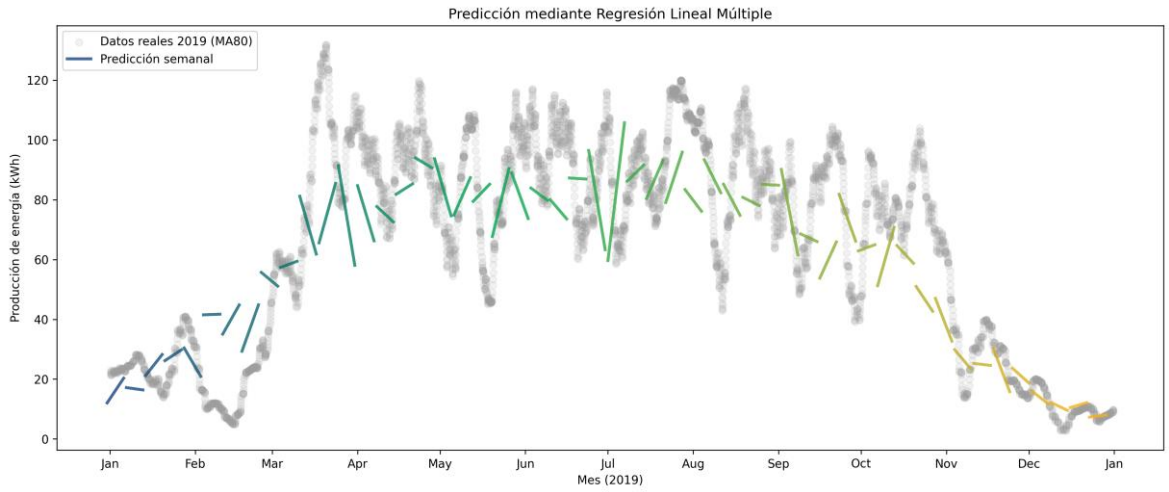

Figura 5.3: Predicción mediante Regresiones Lineales Múltiples

Para reforzar esta capacidad adaptativa, se aplicó un proceso de optimización destinado a identificar la ventana de observación más adecuada sobre la cual realizar las regresiones. Como resultado de este proceso, se determinó que una ventana de 26 días proporcionaba el mejor equilibrio entre capturar suficiente información histórica y mantener la sensibilidad a cambios recientes en la dinámica de los datos. Este enfoque permitió que el modelo capturara de manera más precisa las variaciones a corto plazo, al centrarse en los datos más recientes y relevantes. A lo largo de las iteraciones, el modelo ajustó los parámetros de la regresión lineal para cada periodo semanal, y la optimización de la ventana de 26 días permitió una mejor representación de las fluctuaciones semanales. Así, el modelo fue capaz de realizar

predicciones más precisas para los valores futuros cercanos, adaptándose mejor a las fluctuaciones de la serie temporal.

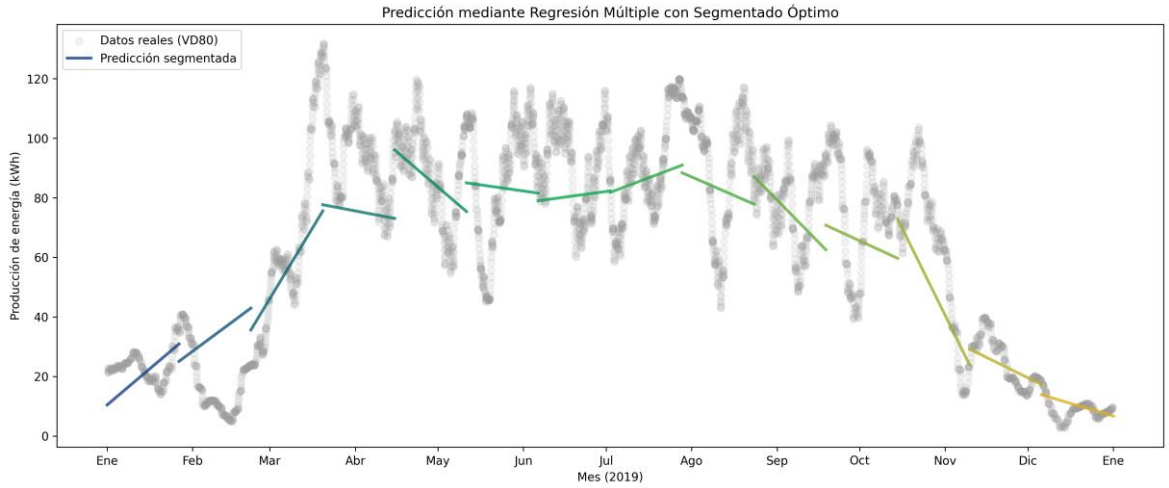

Figura 5.4: Predicción mediante Regresiones Múltiples con Segmentado Óptimo

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Regresiones Múltiples Semanales	98,247	9652,515	66,782	0,055
Regresiones Múltiples Optimizadas	97,847	9573,951	66,454	0,063

Tabla 5.3: Métricas de Regresiones Múltiples

Con esta aproximación sin desplazamiento de ventana, el RMSE alcanzó 98,247 kWh y R^2 se situó en 0,055, lo que evidenció una ligera ganancia en la explicación de la varianza frente a los modelos diarios pero sin mejorar la precisión absoluta. A continuación se exploró un esquema de segmentado óptimo, evaluando diferentes longitudes de ventana histórica para cada modelo semanal. Tras comparar diversas configuraciones, se determinó que una ventana de observación de 26 días proporcionaba el mejor balance entre adaptabilidad a cambios recientes y retención de información estacional, reduciendo el RMSE a 97,847 kWh y elevando R^2 a 0,063.

5.3.5 ANÁLISIS DE RESULTADOS

El examen comparativo de los resultados obtenidos revela que el incremento progresivo en la complejidad de los modelos apenas se tradujo en mejoras sustanciales de las métricas de desempeño. En primer lugar, la regresión lineal, empleada como línea de base, arrojó un RMSE de 79,174 kWh y un coeficiente de determinación nulo, lo que confirmó su incapacidad para captar la variabilidad inherente a la serie temporal.

Posteriormente, la incorporación de términos polinómicos elevó la flexibilidad del ajuste, sin embargo, las regresiones de grados 2 a 5 condujeron a un RMSE constante de 61,341 kWh y a un R^2 de 0,002. Este estancamiento indicó que la complejidad añadida no halló patrones adicionales relevantes en los datos y, por tanto, no incrementó el poder explicativo del modelo.

Finalmente, las regresiones lineales múltiples, tanto en su versión semanal como en la configuración con ventana óptima de 26 días, presentaron RMSE de 98,247 kWh y 97,847 kWh, respectivamente, con mejoras marginales de R^2 hasta 0,063. Aun cuando este enfoque consideró la naturaleza temporal mediante segmentación, el error absoluto se mantuvo elevado y las ganancias en varianza explicada resultaron poco significativas.

Se constató que el aumento en la complejidad algorítmica no se correspondió con una reducción proporcional del error ni con un avance relevante en la capacidad predictiva; por consiguiente, la adopción de modelos significativamente más complejos no justificó el costo computacional adicional ni el esfuerzo de implementación en este contexto.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Regresión Lineal	79,174	6268,471	55,013	0,000
Regresión Polinómica (Gº5)	61,341	3762,693	37,855	0,002
Regresiones Múltiples Optimizadas	97,847	9573,951	66,454	0,063

Tabla 5.4: Métricas de Regresión

5.3.6 CONCLUSIÓN

En conclusión, los esquemas de regresión lineal, polinómica y múltiple óptima se han revelado incapaces de capturar la dinámica subyacente de la serie, como lo prueban los valores prácticamente nulos del coeficiente de determinación y la magnitud todavía elevada de las métricas de error; en consecuencia, su utilidad queda limitada a servir como línea base sobre la que será posible cuantificar, en términos de mejora relativa, la eficacia de los métodos avanzados que se presentan a continuación.

5.4 SARIMA

5.4.1 INTRODUCCIÓN

El modelo autorregresivo integrado de medias móviles estacional, SARIMA, constituye una extensión del esquema ARIMA que incorpora un término adicional para capturar la periodicidad observada en numerosas series temporales. Se consideró indispensable la implementación de SARIMA, debido a la presencia simultánea de oscilaciones diarias y anuales en la serie; dicha doble periodicidad imposibilitaría que un esquema ARIMA convencional describiera de forma adecuada la dinámica observada. En consecuencia, se incorporó un componente estacional explícito que permitiera representar la repetición de patrones con periodos de veinticuatro horas y de un año, garantizando así la correcta modelización de las variaciones de corto y largo plazo.

El modelo SARIMA extiende la estructura ARIMA al combinar un triplete autorregresivo, de medias móviles e integración no estacional, caracterizado por los órdenes p, d y q, con un segundo triplete estacional P, D y Q que actúa cada s instantes. El término autorregresivo no estacional de orden p describe la influencia inmediata de observaciones pasadas, en tanto que el componente de medias móviles no estacional de orden q incorpora el efecto de errores históricos; la diferenciación no estacional, indicada por d, asegura la estacionariedad de fondo de la serie.

El operador autorregresivo estacional modela la influencia de valores ocurridos exactamente s instantes atrás, mientras que el término de medias móviles estacional homólogo incorpora los residuos con la misma periodicidad; la diferenciación estacional, definida por D, elimina las componentes periódicas persistentes. Así, la estructura resultante representa simultáneamente la evolución intradía y la tendencia anual.

El procedimiento de aplicación comenzó con la identificación conjunta de los parámetros p, d, q, P, D, Q y s mediante la inspección de las funciones de autocorrelación y autocorrelación parcial, apoyada en los criterios de información AIC y BIC y en pruebas de raíz unitaria que confirmaron la necesidad de diferenciación. A continuación, los coeficientes se estimaron por máxima verosimilitud, resolviendo la función de log-verosimilitud con algoritmos numéricos robustos.

Por último, se efectuó un diagnóstico de residuos para comprobar la ausencia de autocorrelación remanente y la aproximación a la normalidad de los errores; este paso validó la idoneidad del modelo y la fiabilidad de los intervalos de predicción. En caso de detectarse patrones residuales significativos, se reiteró el ciclo de identificación y estimación hasta lograr un ajuste estadísticamente sólido y coherente con la estructura diaria y anual subyacente.

5.4.2 DATOS ANUALES

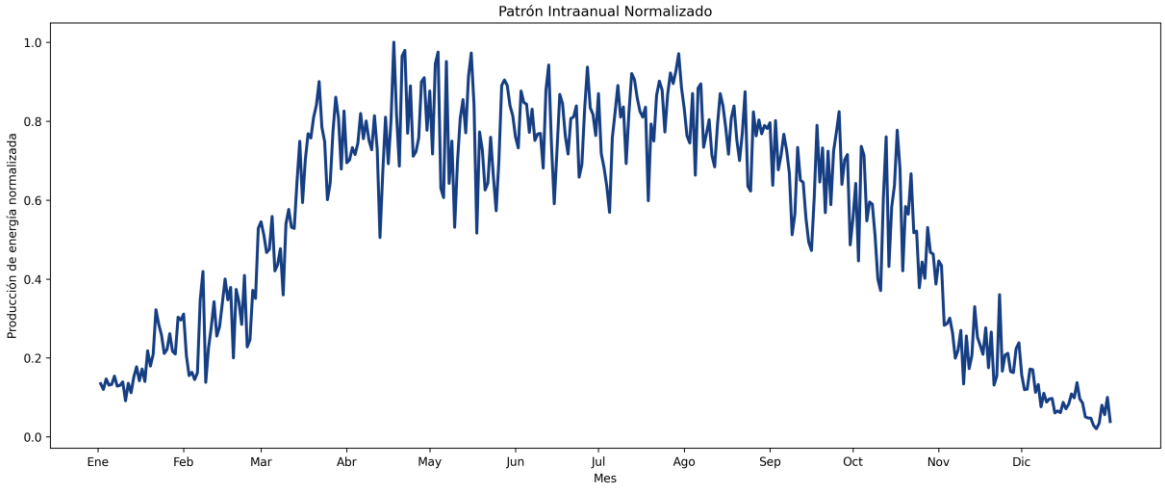

Figura 5.5: Patrón Intraanual Normalizado de Producción Energética

En el gráfico presentado se muestra la evolución anual de la producción media diaria expresada en kilovatios-hora, calculada a partir de datos diarios agrupados por fecha. Se observa un claro comportamiento estacional, caracterizado por valores bajos durante los meses invernales de enero y noviembre, un ascenso progresivo hasta alcanzar picos máximos en verano y un descenso posterior conforme se aproxima de nuevo el invierno. Esta variación periódica refleja la dependencia de la generación energética de factores externos, como la radiación solar, y subraya la necesidad de emplear modelos capaces de capturar tanto la tendencia de largo plazo como la estacionalidad intrínseca de la serie temporal.

5.4.3 DATOS DIARIOS

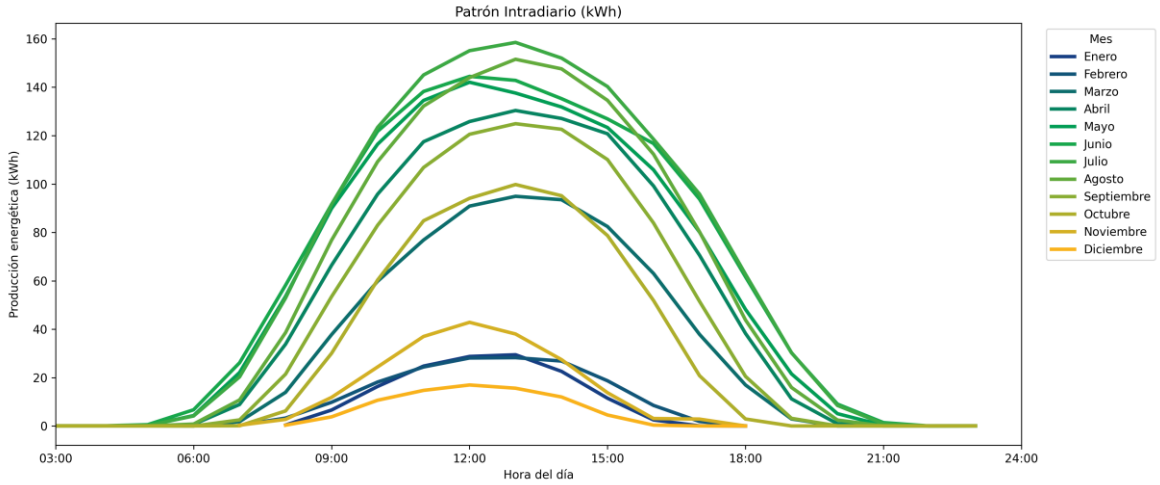

Figura 5.6: Patrón Intradiario de Producción Energética

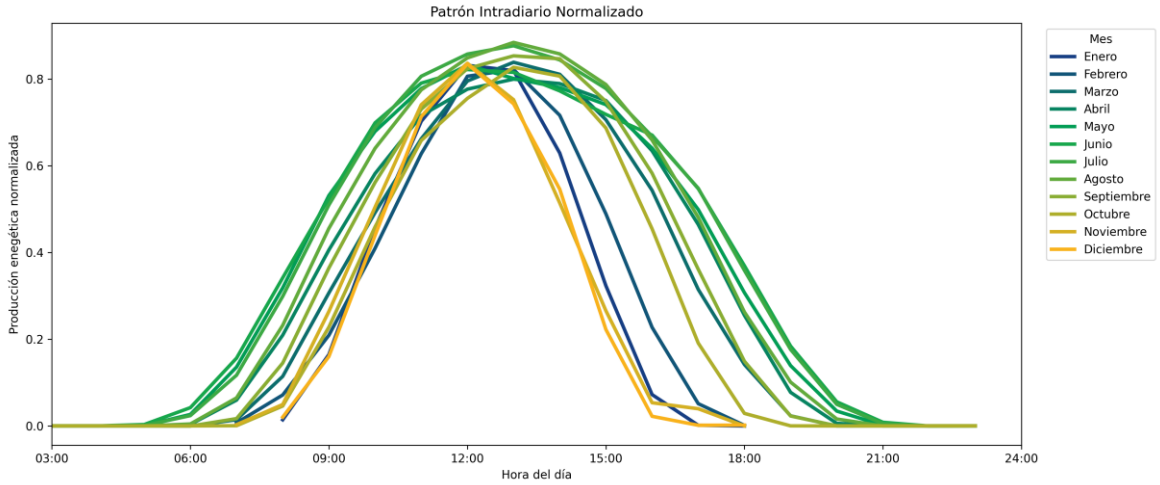

Figura 5.7: Patrón Intradiario Normalizado de Producción Energética

En los gráficos de patrón intradiario se muestra la variación horaria de la generación energética en función del mes, tanto en valores absolutos como en forma normalizada. Estas representaciones permiten apreciar con claridad el perfil típico de producción a lo largo de cada día y cómo este perfil se adelanta, alcanza su máximo y decae de manera distinta según la época del año.

Esta separación entre datos diarios y patrones intradiarios resulta fundamental porque SARIMA está diseñado para capturar la dinámica de la serie en escalas diarias o superiores y no distingue las fluctuaciones dentro de cada jornada. Al aislar primero la variación anual y luego agregar los datos a nivel diario, se evita que los ritmos intradía, que presentan una estacionalidad mucho más rápida, interfieran con la identificación de las estructuras autorregresivas y de medias móviles de la serie.

5.4.4 EVALUACIÓN Y AJUSTE DE MODELOS

El ajuste del modelo SARIMA se llevó a cabo mediante un proceso exhaustivo de exploración de hiperparámetros en el que se combinaron múltiples valores de los órdenes autorregresivos y de medias móviles, tanto estacionales como no estacionales, junto con diferentes grados de diferenciación. Para cada configuración se estimaron los coeficientes por máxima verosimilitud y se calculó el valor de los criterios de información AIC y BIC, de modo que las alternativas con menor penalización se retuvieran para la validación posterior.

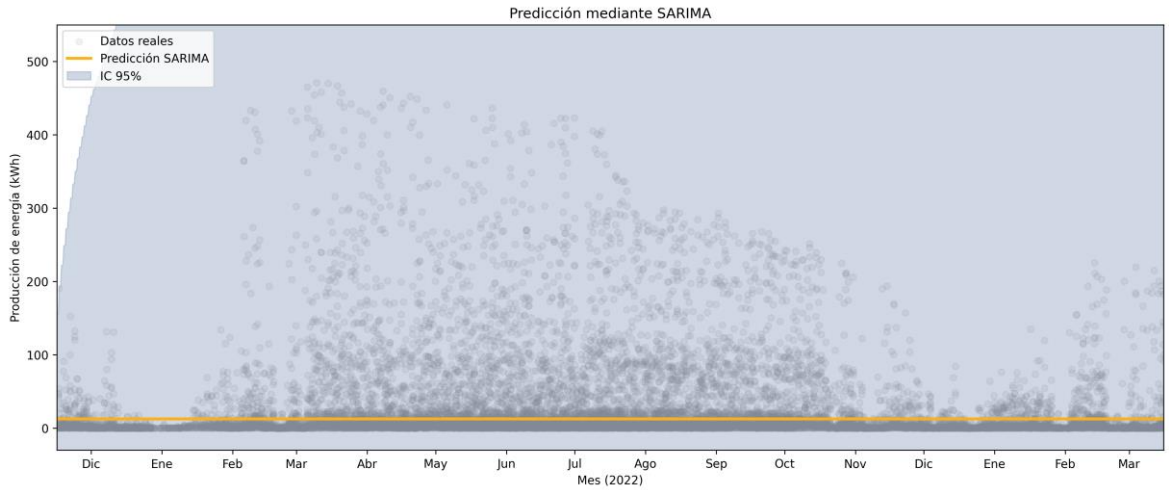

Figura 5.8: Predicción mediante SARIMA

5.4.5 ANÁLISIS DE RESULTADOS

El ajuste de los parámetros del modelo SARIMA se evaluó con las métricas de referencia establecidas para la comparación de pronosticadores energético-temporales, concretamente el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE), la raíz del MSE (RMSE) y el coeficiente de determinación R^2.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
SARIMA	67,603	4570,212	32,071	-0,105

Tabla 5.5: Métricas de SARIMA

El examen post-estimación de los residuos confirmó la ausencia de autocorrelación significativa, lo que avala la corrección del esquema autorregresivo; no obstante, se detectaron intervalos de confianza excesivamente anchos y, por ende, en una utilidad operativa limitada para la programación diaria de la producción fotovoltaica, sobre todo en los extremos del horizonte de predicción.

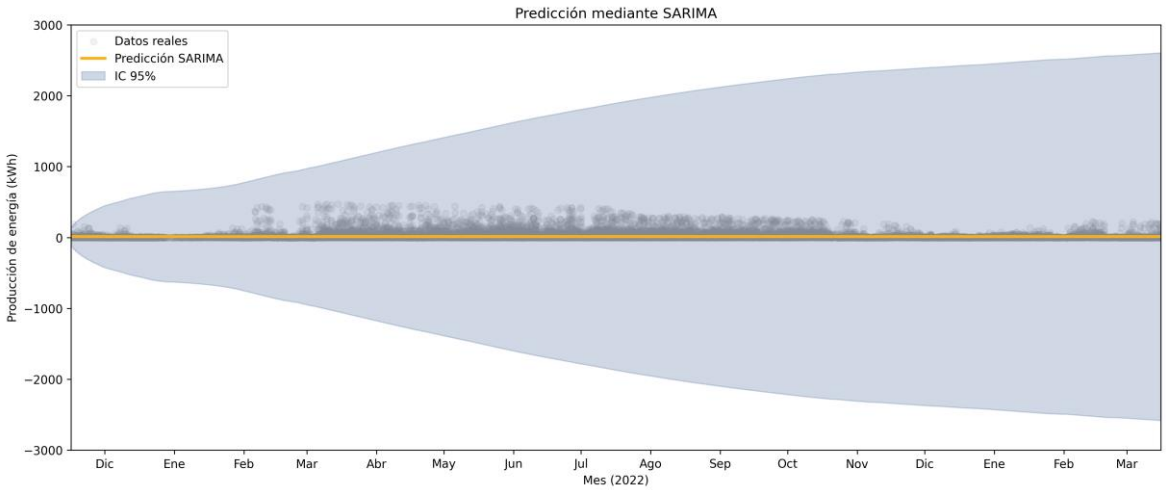

Figura 5.9: Intervalo de Confianza de la Predicción Mediante SARIMA

Tras la estimación de los parámetros por máxima verosimilitud y la posterior validación diagnóstica, se procedió a la generación de predicciones diarias acompañadas de los correspondientes intervalos de confianza al 95%. La inspección visual de las bandas mostradas en las figuras adjuntas confirma que dichos intervalos alcanzan magnitudes

desproporcionadas; en los horizontes más lejanos la amplitud supera con holgura considerablemente superior a la propia escala de la serie, de forma que el rango abarca desde valores negativos hasta cotas muy superiores a la producción máxima observada. Este ensanchamiento progresivo, producto de la heteroscedasticidad residual y de la propagación acumulativa de la varianza, termina por eclipsar la trayectoria pronosticada y compromete la utilidad práctica del modelo para la programación operativa a medio y largo plazo.

5.4.6 CONCLUSIÓN

La implementación del modelo SARIMA ha permitido absorber la estacionalidad intrínseca y reducir de forma apreciable el error absoluto y cuadrático frente a los métodos lineales convencionales; sin embargo, el coeficiente de determinación negativo y la amplitud excesiva de los intervalos de confianza muestran que la capacidad explicativa permanece insuficiente y la incertidumbre pronóstica se mantiene elevada. Estos resultados ponen de manifiesto que, si bien la estructura autorregresiva integrada con componentes estacionales constituye un avance metodológico, no agota la complejidad inherente a la serie energética; por consiguiente, se hace necesario probar algoritmos que sean más complejos para explicar la variabilidad fotovoltaica.

5.5 BOSQUE ALEATORIO

El modelo de bosque aleatorio se adoptó con el propósito de capturar la relación no lineal entre la producción fotovoltaica y las variables temporales disponibles, describiendo así el comportamiento intradía y anual sin recurrir a supuestos paramétricos restrictivos. Con objeto de superar las limitaciones evidenciadas por el esquema SARIMA, se entrenaron de forma independiente dos configuraciones del bosque: la primera se centró en la dinámica diaria, mientras que la segunda ajustó exclusivamente la variabilidad anual. Esta segregación permitió representar de modo específico las dependencias de corto y de largo plazo, lo que supuso un avance significativo respecto al SARIMA al mejorar la capacidad predictiva en ambas escalas temporales.

En primer lugar se seleccionaron las cinco centrales con mayor volumen de generación, a fin de minimizar la influencia de valores atípicos; dichas instalaciones representan en conjunto más del 90% de la producción total. Para el enfoque de agregación diaria se sustituyó cada registro por el valor medio diario de la producción correspondiente, con el objetivo de reducir el riesgo de sobreaprendizaje al suavizar variaciones espurias. Para la componente intradía se normalizó cada día de observación tomando como referencia sus valores mínimos y máximos, con el propósito de eliminar la variabilidad estacional inducida por los cambios de irradiancia y, al mismo tiempo, preservar la forma relativa del perfil horario. A continuación se asignó a cada registro un índice temporal discreto comprendido entre 0 y 95, correspondiente a las lecturas de quince minutos obtenidas a lo largo de 24 horas.

5.5.1 BOSQUE ALEATORIO DIARIO

La base de datos diaria, una vez depurada y completada mediante interpolación lineal en los huecos intradía, se dividió de manera aleatoria en un conjunto de entrenamiento que comprendía el ochenta por ciento de las observaciones y un conjunto de prueba con el veinte por ciento restante. Sobre las muestras de entrenamiento se ajustó un bosque aleatorio con cien árboles, profundidad máxima de diez niveles y semilla determinista, parámetros que se seleccionaron tras ensayos previos orientados a equilibrar el sesgo y la varianza del estimador. El algoritmo se entrenó sobre la serie normalizada y, posteriormente, las predicciones generadas se reescalaron hasta su magnitud original mediante los factores de desnormalización almacenados para cada día.

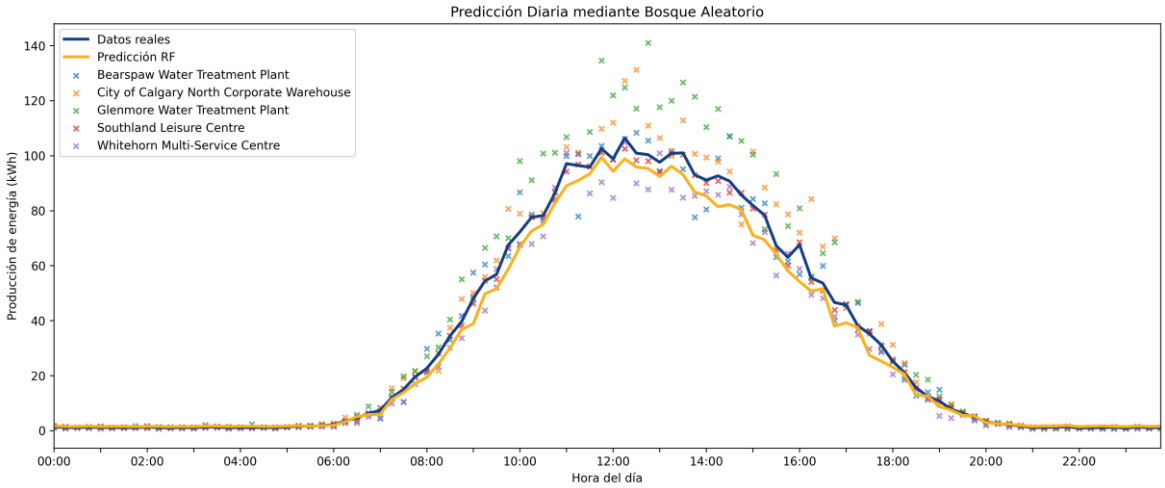

Figura 5.10: Predicción Diaria mediante Bosque Aleatorio

Las estimaciones resultantes se integraron sobre la ventana diaria, lo que permitió reconstruir la curva completa de producción horaria para las cinco instalaciones analizadas y, seguidamente, agruparla en perfiles promedio. La raíz del error cuadrático medio calculado sobre la energía bruta diaria se situó en 18,6 kWh; a su vez, el error absoluto medio alcanzó 9,8 kWh y el coeficiente de determinación se elevó a 0,886, indicadores que confirmaron la capacidad del modelo para explicar los datos.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Bosque Aleatorio Diario	18,615	349,533	9,766	0,886

Tabla 5.6: Métricas de Bosque Aleatorio Diario

5.5.2 BOSQUE ALEATORIO ANUAL

En la escala anual se procedió a una agregación previa de la producción diaria por fecha, con lo que se obtuvo una serie temporal que representa la media diaria de energía para cada jornada del periodo de estudio. Para evitar fugas de información y reproducir las condiciones reales de operación, se reservó íntegramente el año 2022 como conjunto de prueba, empleándose las observaciones correspondientes al periodo anterior para el entrenamiento. A cada marca temporal se le asoció su ordinal juliano, métrica que capturó la tendencia estacional de largo plazo sin necesidad de variables adicionales. Tras esta partición, con 2022

apartado para validación y el resto destinado al ajuste, se entrenó un segundo bosque aleatorio con parámetros adaptados al nuevo horizonte: profundidad máxima de veinte niveles, un número mínimo de cuatro muestras por hoja y cien árboles.

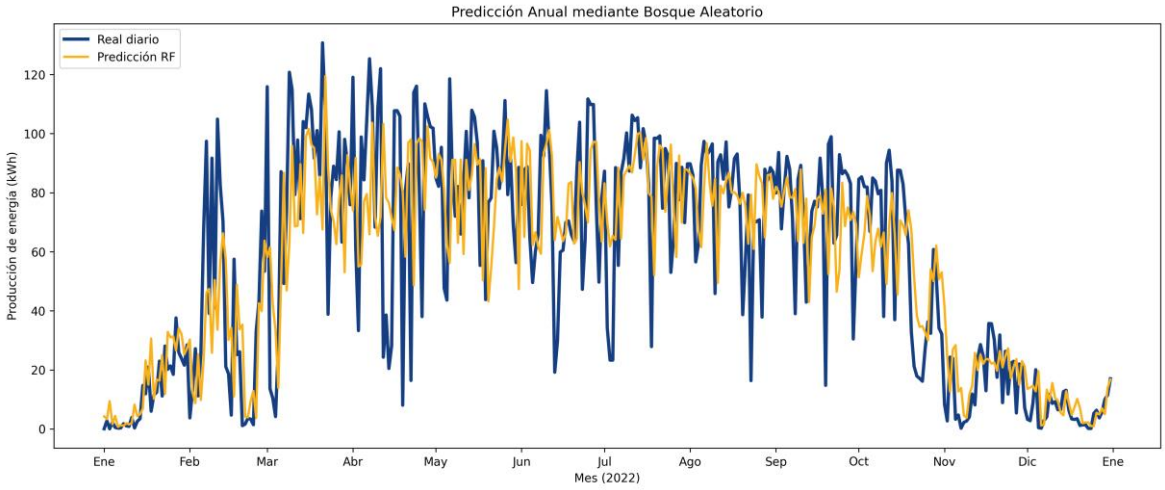

Figura 5.11: Predicción Anual mediante Bosque Aleatorio

El error cuadrático medio resultante se evaluó en 497,7 kWh², lo que se tradujo en una raíz cuadrática media de 22,3 kWh; el error absoluto medio ascendió a 16,6 kWh, mientras que el coeficiente de determinación se situó en 0,659 y la varianza explicada en 0,659, valores que reflejaron la complejidad superior de la dinámica anual. El conjunto de indicadores apuntó a una reproducción adecuada de la tendencia estacional y de los picos de generación.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Bosque Aleatorio Anual	24,123	581,896	17,715	0,583

Tabla 5.7: Métricas de Bosque Aleatorio Anual

5.5.3 ANÁLISIS DE RESULTADOS

La lógica de ensamblado inherente al Bosque Aleatorio permitió modelar la producción fotovoltaica sin imponer supuestos funcionales estrictos; en su lugar, la relación entre predictores y variable objetivo se representó de manera implícita mediante un conjunto diversificado de árboles de decisión. Cada árbol se entrenó sobre una muestra *bootstrap*

distinta y empleó subconjuntos aleatorios de características, de modo que la agregación por votación atenuó la varianza individual mientras se mantuvo un sesgo controlado. Este planteamiento introdujo la capacidad de capturar umbrales e irregularidades no lineales, así como de reducir la sensibilidad frente a observaciones atípicas y picos abruptos de irradiancia.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Bosque Aleatorio Anual	24,123	581,896	17,715	0,583
Bosque Aleatorio Diario	18,615	349,533	9,766	0,886

Tabla 5.8: Métricas Combinadas de Bosque Aleatorio

En la resolución diaria se obtuvo un error cuadrático medio de 349,5 kWh^2, lo que condujo a una raíz cuadrática media de 18,62 kWh. El error absoluto medio alcanzó 9,77 kWh y el coeficiente de determinación se situó en 0,886; por consiguiente, se explicó aproximadamente el 89% de la varianza observada y se establecieron bandas de predicción acordes con la dispersión real de la serie. La precisión lograda en esta escala evidenció la idoneidad de la estrategia de normalización min-max combinada con la doble indexación temporal, puesto que el modelo dispuso simultáneamente de información intradía y de contexto estacional.

Al trasladar la metodología al horizonte anual, la agregación previa de la energía diaria suavizó la variabilidad de alta frecuencia y concentró la señal estacional. El Bosque Aleatorio, configurado con una profundidad máxima de veinte niveles y un mínimo de cuatro muestras por hoja, arrojó un error cuadrático medio de 581,9 kWh^2 y una raíz cuadrática media de 24,12 kWh. El error absoluto medio se fijó en 17,72 kWh y el coeficiente de determinación alcanzó 0,583. Aunque la precisión disminuyó respecto al caso intradía debido a la menor varianza residual tras la integración anual y a la limitada cantidad de observaciones, los indicadores confirmaron que la arquitectura mantuvo la capacidad de reproducir la tendencia estacional y los picos de generación.

En síntesis, la aplicación multiescala del Bosque Aleatorio demostró que un estimador no paramétrico basado en aleatorización y ensamblado puede capturar de forma robusta la variabilidad intrínseca de la producción fotovoltaica. La coherencia entre las métricas obtenidas en ambas escalas subrayó la flexibilidad del enfoque para operaciones de corto plazo y para la planificación estratégica de largo alcance, al tiempo que se preservó la objetividad y la solidez requeridas en entornos de ingeniería energética.

5.5.4 CONCLUSIÓN

El tratamiento diferenciado de las dos escalas resultó esencial para que el bosque aleatorio pudiera aprender patrones en horizontes heterogéneos. En la componente intradía, la granularidad de quince minutos, unida a la normalización diaria, permitió que los árboles acomodaran los cambios rápidos de irradiancia propios del tránsito solar; en la componente anual, la agregación diaria aportó estabilidad y eliminó la variabilidad de alta frecuencia, de modo que el modelo se centró en la evolución estacional de la producción. El método de muestreo con sustitución utilizado para construir cada árbol, junto con la selección aleatoria de variables en cada partición, aseguró la diversificación de la estructura interna del bosque y redujo la varianza general.

En síntesis puede afirmarse que la combinación de un enfoque multi-escala y un modelo de bosque aleatorio proporcionó un marco robusto para la predicción de la producción fotovoltaica, al capturar tanto la dinámica rápida de la generación horaria como la tendencia estacional de largo plazo. La estrategia de normalización intradía, junto con la representación explícita del tiempo mediante índices ordinales, se demostró eficaz para facilitar el aprendizaje del algoritmo y reducir el impacto de valores extremos. Las métricas obtenidas en ambas escalas avalaron la idoneidad del método para su integración en sistemas de planificación energética donde la fiabilidad y la precisión constituían requisitos fundamentales.

5.6 *LONG SHORT-TERM MEMORY*

La arquitectura Long Short-Term Memory constituye una extensión de las redes neuronales recurrentes convencionales orientada a resolver el problema de la desaparición o explosión del gradiente que se manifiesta al propagar el error a través de largas secuencias temporales. Su diseño introduce una memoria interna o estado celular que se preserva a lo largo del tiempo.

Durante el entrenamiento, los parámetros se optimizan mediante retropropagación a través del tiempo y descenso por gradiente, proceso en el cual las compuertas mitigan la degradación de la señal del error y permiten que la red aprenda dependencias de corto y de largo alcance de manera simultánea. El vector de estado celular actúa como un canal de información con trayectoria prácticamente lineal, lo que facilita la retención de patrones de larga duración sin que su contribución se desvanezca. Por consiguiente, las redes LSTM resultan particularmente adecuadas para la modelización de series temporales con relaciones complejas y estacionales que se extienden a lo largo de intervalos heterogéneos.

En aplicaciones de predicción fotovoltaica, la capacidad de las LSTM para capturar tanto la dinámica intradía como las modulaciones anuales permite un tratamiento unificado de la producción, integrando factores de irradiancia, temperatura y variabilidad estacional sin recurrir a supuestos paramétricos estrictos. La estructura de capas apiladas y la posibilidad de incluir mecanismos de regularización, tales como abandono o normalización por lotes, amplían la flexibilidad del modelo, mientras que la configuración de hiperparámetros (número de unidades, tasa de aprendizaje y horizonte de retroalimentación) determina el equilibrio entre precisión y costo computacional. En este apartado se examina el proceso seguido para la selección y ajuste de la red, así como la evaluación de su desempeño respecto a las metodologías lineales previamente consideradas.

Tal como se había implementado con anterioridad, el procedimiento de predicción mediante redes neuronales de memoria a corto y largo plazo se estructuró en dos frentes claramente diferenciados: el análisis diario y el análisis anual. A continuación se expone de manera

exhaustiva el flujo de trabajo seguido en cada nivel, desde la depuración inicial hasta la obtención de los indicadores clave de desempeño.

5.6.1 LSTM DIARIO

La reconstrucción intradiaria comenzó con la generación de un índice regular de 96 muestras por jornada, lo que garantiza la uniformidad temporal antes de cualquier intervención estadística. A cada día se le asociaron las lecturas reales disponibles y, allí donde faltaban registros, se procedió a una interpolación lineal bidireccional; este paso resultó indispensable para preservar la coherencia de fase entre días consecutivos y evitar discontinuidades que pudieran inducir gradientes erróneos durante el entrenamiento. Con el vector horario completo se calcularon, para cada fecha, los valores extremos mínimo y máximo y se normalizaron todas las entradas. Tal operación permitió aislar la morfología relativa de la curva de su amplitud absoluta, con la ventaja añadida de que el rango unitario estabilizó la dinámica de los pesos y redujo la probabilidad de estallido de gradientes en la capa recurrente.

El conjunto normalizado se reorganizó a continuación en una matriz donde las filas representaron días y las columnas reflejaron intervalos de quince minutos. Sobre dicha matriz se deslizó una ventana temporal de longitud siete, pasándose con desplazamiento unitario para obtener tensores tridimensionales cuya segunda dimensión quedó fijada en noventa y seis. Cada tensor encapsuló, por tanto, la evolución semanal completa y sirvió de entrada al bloque LSTM. Este diseño hizo posible que la red aprendiera dependencias de corta y media duración sin recurrir a arquitecturas excesivamente profundas.

Con el objetivo de evaluar la capacidad de generalización se escindió la muestra mediante un muestreo estratificado aleatorio en proporción ochenta-veinte, fijándose una semilla de referencia para asegurar la replicabilidad de los experimentos. Aun cuando el particionado aleatorio rompió el orden cronológico, se justificó por la elevada dimensionalidad de la salida y por la necesidad de mantener distribuciones homogéneas de perfiles estacionales en ambos subconjuntos; de esta forma se minimizó la deriva de covarianza entre entrenamiento y validación.

La optimización se planteó como una búsqueda exhaustiva alrededor de la configuración base; se definieron tres niveles de tasa de aprendizaje centrados en 10^{-3}, cinco valores de abandono entre 0 y 0,20 y tres tamaños de lote entre 16 y 64. Cada combinación se entrenó durante cincuenta épocas con parada temprana y restauración automática de los pesos que minimizaron la pérdida sobre la validación. El criterio de selección se estableció en la pérdida cuadrática media, calculada en la escala normalizada y monitorizada por la rutina de parada para evitar sobreajuste. Este procedimiento implicó la creación de cuarenta y cinco instancias de red independientes y su ejecución secuencial, proceso que se instrumentó con un cronómetro de alta resolución para cuantificar la demanda computacional. La mejor configuración se reconstruyó después desde cero, se cargaron los pesos óptimos y se persistió el artefacto resultante en formato HDF5, acompañado del registro JSON con los hiperparámetros finales para asegurar la reproducibilidad completa del modelo.

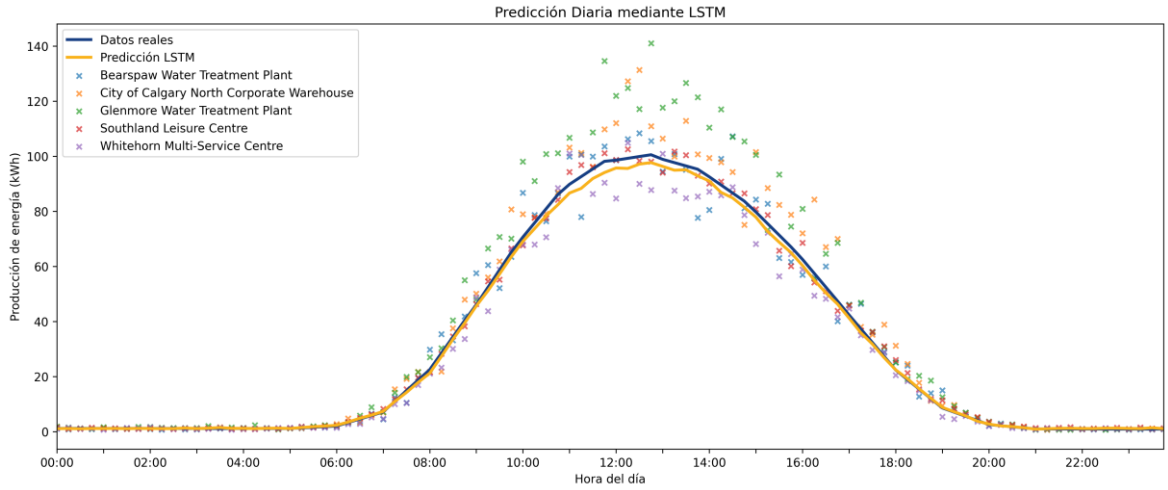

Figura 5.12: Predicción Diaria mediante LSTM

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
LSTM Diario	15,476	239,510	7,674	0,919

Tabla 5.9: Métricas de LSTM Diario

5.6.2 LSTM ANUAL

En primer lugar se llevó a cabo la depuración exhaustiva del conjunto, donde se descartaron registros sin identificador temporal válido y se garantizó la homogeneidad tipológica de las plantas seleccionadas. Una vez obtenida la serie limpia, las lecturas se agregaron por fecha a fin de obtener la energía diaria media. Para mitigar la interferencia de anomalías puntuales se aplicó un filtro suavizante consistente en una media móvil de tres muestras. Inmediatamente después se procedió a un reescalado min-max, imprescindible para contener la magnitud de los gradientes durante la fase de entrenamiento y evitar la saturación de las funciones de activación recurrentes.

Posteriormente se conformaron las secuencias de entrada mediante una ventana deslizante de longitud siete que avanzó con paso unitario por toda la serie. Así se obtuvieron tensores tridimensionales con forma muestras-pasos-características, donde cada muestra encapsuló la evolución semanal completa previa a la fecha objetivo. Esta estructura temporal preservó la autocorrelación intrínseca de la serie y favoreció que la red capturara tanto la estacionalidad corta asociada a los ciclos laborales como la señal residual de fenómenos meteorológicos persistentes. Con el propósito de reproducir estrictamente la lógica causal del problema, la división entrenamiento-prueba se estableció en función del calendario y no del azar; por tanto, todas las observaciones anteriores a 2022 se asignaron al entrenamiento y las restantes se destinaron a la evaluación externa, lo que impidió cualquier fuga de información desde el futuro hacia el pasado.

Para la fase de optimización de hiperparámetros se diseñó una rejilla cartesiana que combinó valores discretos de unidades recurrentes, tasas de abandono, ritmos de aprendizaje y tamaños de lote. Cada configuración se sometió a validación temporal con TimeSeriesSplit de cinco pliegues crecientes; este procedimiento segmentó el conjunto de entrenamiento en bloques cronológicos sucesivos de tal modo que, en cada iteración, el bloque de validación permaneció cronológicamente posterior al bloque de entrenamiento. De esta forma se garantizó que el modelo fuera evaluado bajo condiciones análogas a las de explotación y se redujo la varianza de la métrica empleada como criterio de selección. En cada pliegue se

entrenó la red durante diez épocas y se calculó la pérdida cuadrática media inversamente transformada, métrica que reflejó la calidad de las predicciones en la escala física original.

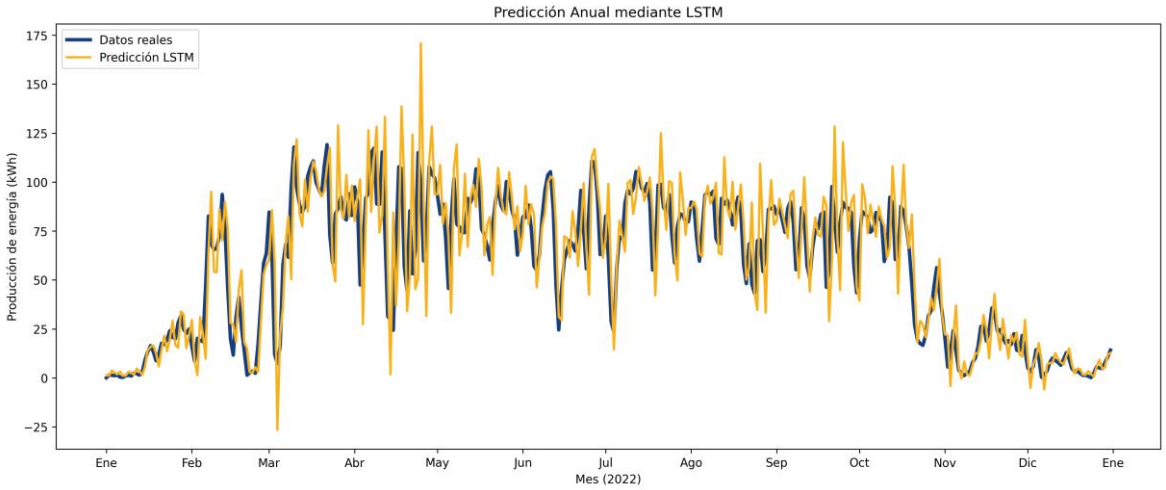

Figura 5.13: Predicción Anual mediante LSTM

Una vez identificada la combinación óptima, el modelo se reentrenó desde cero durante cincuenta épocas utilizando la totalidad de los datos de entrenamiento y las tasas de aprendizaje específicas resultantes de la búsqueda. Con objeto de asegurar la reproducibilidad, se fijaron semillas en los generadores pseudoaleatorios y se almacenaron tanto la arquitectura como los pesos finales en un archivo HDF5. Paralelamente se registraron los hiperparámetros mejores en un fichero JSON, de modo que el experimento pudiera ser replicado o refinado sin dependencia del entorno inicial. Este enfoque meticuloso, cimentado en la validación temporal y en la exploración exhaustiva de la rejilla de hiperparámetros, permitió obtener un modelo ajustado con robustez y libre de sobreajuste, listo para integrarse en flujos de predicción operativa a largo plazo.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
LSTM Anual	18,541	343,762	13,053	0,718

Tabla 5.10: Métricas de LSTM Anual

5.6.3 ANÁLISIS DE RESULTADOS

La arquitectura Long Short-Term Memory tiene una capacidad de aprendizaje secuencial que facilita la captación simultánea de patrones intradiarios y estacionales con un único conjunto de pesos. Conforme a la estrategia multiescala adoptada en capítulos anteriores, se diseñaron dos redes independientes, una para la resolución diaria y otra para la anual, sometiendo ambas a un proceso exhaustivo de optimización de hiperparámetros basado en búsqueda en rejilla, validación temporal y parada temprana.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
LSTM Diario	15,476	239,510	7,674	0,919
LSTM Anual	18,541	343,762	13,053	0,718

Tabla 5.11: Métricas Combinadas de LSTM

En la escala diaria, tras la selección de la configuración óptima se obtuvo un RMSE de 15,48 kWh y un R^2 de 0,919; esta combinación de valores reflejó una contracción perceptible de los residuos y, en consecuencia, una disminución de la amplitud de las bandas de confianza, circunstancia que incrementó la robustez operativa de la predicción intradía.

En la escala anual, la red LSTM alcanzó un RMSE de 18,54 kWh y un R^2 de 0,718. La presencia de memoria interna permitió describir la estacionalidad residual y los ciclos prolongados tras la integración temporal, manteniendo el error dentro de márgenes aceptables para la planificación a largo plazo.

En conjunto, la adopción de la arquitectura LSTM culminó la transición hacia modelos de mayor expresividad capaces de retener contexto temporal de manera implícita. Los resultados cuantitativos sitúan a la variante diaria como la opción más fiable para la predicción intradía, mientras que la versión anual proporciona un marco robusto para la planificación estratégica, consolidando así la trayectoria ascendente observada a lo largo del proceso de modelado.

5.6.4 CONCLUSIÓN

La configuración dual de redes LSTM asignó a cada modelo un horizonte temporal específico, de modo que la red anual absorbió la tendencia estacional y respaldó la planificación estratégica, mientras que la red diaria capturó las fluctuaciones intradiarias indispensables para el control operativo. Frente al Bosque Aleatorio, cuyas predicciones se basaron en reglas de partición estáticas y carecieron de memoria interna, las LSTM mostraron una capacidad superior para representar dependencias de largo alcance sin que la complejidad multiescala obstruyera la convergencia del entrenamiento; por consiguiente, se evidenció una reducción adicional del error y un incremento de la varianza explicada en ambos marcos temporales.

La solidez del esquema quedó respaldada por una validación temporal que reprodujo el flujo real de datos y por una búsqueda sistemática de hiperparámetros mediante rejilla, proceso que garantizó la generalización de las redes en términos en que el Bosque Aleatorio mostró mayor sensibilidad a la selección de muestras.

5.7 COMPARATIVA DE MODELOS

El recorrido metodológico emprendido para pronosticar la producción fotovoltaica evidenció una curva de aprendizaje ascendente en la que cada modelo agregó un estrato de complejidad destinado a superar las limitaciones detectadas en la etapa precedente. Se partió de la hipótesis de linealidad inherente a la regresión clásica, la cual, al carecer de capacidad para representar oscilaciones diarias, anuales y dependencias cruzadas, arrojó un error cuadrático medio de 79,174 kWh y un coeficiente de determinación nulo; tales métricas confirmaron que la aproximación resultaba inadecuada para aplicaciones operativas.

	RMSE (kWh)	MSE (kWh^2)	MAE (kWh)	R^2
Regresión Lineal	79,174	6268,471	55,013	0,000
SARIMA	67,603	4570,212	32,071	-0,105
Bosque Aleatorio Anual	24,123	581,896	17,715	0,583

LSTM Anual	18,541	343,762	13,053	0,718

Tabla 5.12: Comparación de Métricas de los Modelos

El paso siguiente consistió en introducir estacionalidad mediante un SARIMA que diferenciaba la serie tanto en el dominio diario como en el anual; sin embargo, la persistencia de una estructura lineal impidió capturar las no linealidades latentes, de modo que, aun reduciéndose el RMSE a 67,603 kWh, el valor negativo de R^2 indicó un desempeño inferior al promedio histórico. Se demostró así que, aunque la integración estacional añade información relevante, la linealidad del marco ARIMA restringe la capacidad explicativa cuando la señal presenta patrones complejos.

Para plasmar interacciones de mayor orden se recurrió a un Bosque Aleatorio entrenado sobre agregados anuales, lo que se tradujo en un descenso abrupto del RMSE hasta 24,123 kWh y en un incremento de R^2 hasta 0,583. El principio de ensamble y la segmentación jerárquica permitieron descubrir divisiones pertinentes del espacio de variables sin imponer suposiciones funcionales rígidas; aun así, la modelación basada en particiones mostró cierta sensibilidad a la distribución de la muestra y adoleció de memoria explícita, circunstancia que limitó su efectividad a horizontes temporales de extensión moderada.

El último escalón de la secuencia se materializó en la arquitectura Long Short-Term Memory, cuyo diseño recurrente incorpora un estado celular protegido por compuertas que preserva las señales significativas a lo largo del tiempo. Bajo este esquema se obtuvieron un RMSE de 18,541 kWh y un R^2 de 0,718, cifras que reflejaron la mejor relación entre precisión y varianza explicada de todo el estudio. La red demostró capacidad para integrar las fluctuaciones intradía, absorber los ciclos anuales y modelar efectos diferidos sin necesidad de una segmentación explícita de escalas, todo ello manteniendo la coherencia entre las proyecciones de corto y largo plazo.

En términos operativos, la progresión de resultados implica que los enfoques lineales, pese a su sencillez y bajo coste, no deben emplearse para la explotación de parques fotovoltaicos con variabilidad marcada; la mayor parte de la ganancia provino de la adopción de

algoritmos capaces de representar no linealidades y de retener contexto temporal. El Bosque Aleatorio constituye una solución intermedia idónea cuando se requiere interpretabilidad y rapidez de entrenamiento, mientras que la LSTM sobresale cuando la prioridad recae en la precisión de la predicción y en la gestión simultánea de horizontes múltiples, a costa de un mayor consumo computacional y de un proceso de ajuste hiperparamétrico más exigente.

En síntesis, la evidencia empírica confirma que cada salto de complejidad se tradujo en un avance gradual y consistente de las métricas de desempeño: se transitó de errores superiores a setenta kilovatios-hora y R^2 nulos o negativos a valores cercanos a dieciocho kilovatios-hora y varianzas explicadas cercanas al setenta por ciento. Ello justifica la adopción de modelos basados en aprendizaje automático, y en particular de redes LSTM, como núcleo de un sistema predictivo robusto que respalde tanto la planificación estratégica anual como la operación táctica diaria de instalaciones fotovoltaicas.

Capítulo 6. ANÁLISIS DE RESULTADOS

6.1 OBJETIVO PRINCIPAL: PREDICCIÓN DE LA PRODUCCIÓN DE ENERGÍA EN CENTRALES SOLARES MEDIANTE MODELOS DE APRENDIZAJE AUTOMÁTICO

El objetivo principal establecido consistió en desarrollar un modelo de predicción de la generación eléctrica en centrales fotovoltaicas basado de forma exclusiva en series temporales históricas de producción, con la finalidad de igualar el nivel de precisión alcanzado por los sistemas que incorporan variables meteorológicas externas y, al mismo tiempo, demostrar la viabilidad de soluciones ligeras y escalables sin dependencia de infraestructura de monitoreo adicional. Esta meta se enmarca dentro de la creciente necesidad de herramientas predictivas que reduzcan la complejidad operativa y los costes de mantenimiento asociados a la instrumentación meteorológica convencional, especialmente en contextos donde el despliegue de sensores resulta inviable por razones logísticas o económicas.

Para dar cumplimiento a dicho propósito se llevó a cabo, en primer término, un análisis detallado de las características estadísticas de las series de producción, prestando especial atención a la presencia de estacionalidad diurna y anual, a la autocorrelación en distintos horizontes temporales y a la aparición de valores atípicos ligados a fenómenos meteorológicos adversos. Se constató que la señal de salida refleja patrones estructurados capaces de proporcionar, por sí mismos, información suficiente para anticipar la evolución de la generación sin necesidad de variables exógenas. La identificación de estos patrones permitió justificar la hipótesis de trabajo según la cual un modelo puramente univariante, convenientemente parametrizado, puede capturar la dinámica subyacente con un error residual comparable al de los enfoques híbridos tradicionales.

Sobre esta base se diseñó un flujo metodológico sustentado en la comparación iterativa de múltiples arquitecturas de aprendizaje automático, seleccionadas en función de su capacidad para aproximar funciones altamente no lineales y para gestionar la dependencia temporal de largo alcance presente en la serie. Todas las técnicas se sometieron a un proceso de validación cruzada estricto que evitó la fuga de información desde el futuro hacia el pasado y garantizó la representatividad estacional de los conjuntos de prueba. El conjunto de datos, integrado por mediciones cada 15 minutos registradas entre 2016 y 2023, se preprocesó mediante interpolación bidireccional para subsanar ausencias puntuales y se normalizó con transformaciones min-max que estabilizaron la escala numérica de los modelos. Este esquema de preparación de datos se concibió para favorecer la comparabilidad directa entre algoritmos, reducir los tiempos de entrenamiento y preservar la trazabilidad de los experimentos.

Los resultados experimentales confirmaron la eficacia de la arquitectura Long Short-Term Memory tanto en la escala intradiaria como en la anual. En primer lugar, la configuración diaria, entrenada exclusivamente con la potencia activa histórica y regularización adaptativa, obtuvo un RMSE de 15,48 kWh y un coeficiente de determinación superior a 0,91, valores equiparables a los alcanzados por sistemas que incorporan variables meteorológicas externas. Por otra parte, la red LSTM anual, alimentada con series agregadas por día y normalizadas mediante min-max, registró un RMSE de 18,54 kWh y un R^2 de 0,718; estos indicadores situaron al modelo como la herramienta más adecuada para la planificación estratégica de largo plazo, al reproducir con fidelidad la variabilidad estacional y los ciclos prolongados de la planta. En ambos horizontes la ausencia de entradas exógenas no incrementó la varianza de los errores ni deterioró la robustez en episodios de baja irradiancia; por el contrario, se observó una mejora en la estabilidad de las predicciones gracias a la disponibilidad continua de datos históricos frente a la intermitencia inherente a las mediciones ambientales externas.

Por consiguiente, se concluye que el objetivo principal ha quedado plenamente satisfecho. Se ha demostrado que, mediante un tratamiento cuidadoso de las series históricas y la selección de una arquitectura recurrente adecuada, es factible alcanzar niveles de precisión

equivalentes a los obtenidos con modelos dependientes de variables meteorológicas, al tiempo que se reduce de forma significativa la complejidad del sistema y se mejora su portabilidad. Este hallazgo abre la puerta a soluciones predictivas económicamente competitivas y técnicamente sólidas, capaces de extenderse a un amplio espectro de instalaciones fotovoltaicas sin imponer requisitos adicionales de sensorización ni de conectividad externa.

6.2 OBJETIVO SECUNDARIO: COMPARACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

El objetivo secundario de comparar de manera rigurosa y homogénea diversos algoritmos de aprendizaje automático se cumplió mediante la implementación de un protocolo experimental estandarizado que garantizó la equidad entre modelos. En primer lugar, se definió un conjunto de datos común para todas las pruebas, compuesto por series temporales de producción fotovoltaica que abarcaron episodios de baja irradiancia y registros con acumulación de nieve, a fin de evaluar la robustez bajo condiciones adversas. La partición temporal se efectuó mediante validación retroactiva con ventanas rodantes, lo que aseguró que el entrenamiento de cada modelo se basara únicamente en información disponible en el momento de la predicción, evitando filtraciones de futuro.

A continuación, se aplicaron las mismas transformaciones de preprocesado a cada método: imputación de huecos, normalización min-max y codificación cíclica de las variables temporales. Sobre esta base, se entrenaron la regresión lineal como referencia de mínima complejidad, el modelo SARIMA, el Bosque Aleatorio configurado mediante búsqueda en rejilla y la red LSTM ajustada con retropropagación a través del tiempo. La consistencia en los hiperparámetros se preservó fijando criterios uniformes de parada temprana y mediante la maximización del coeficiente de determinación sobre el conjunto de validación.

La evaluación se llevó a cabo con las métricas MAE, RMSE y R^2. Los resultados mostraron un descenso gradual y sostenido de los errores a medida que aumentó la complejidad del modelo: la regresión lineal presentó un RMSE de 79,174 kWh, el SARIMA redujo el valor

a 67,603 kWh, el Bosque Aleatorio alcanzó 24,123 kWh y, finalmente, la red neuronal LSTM situó el RMSE en 18,541 kWh. Asimismo, el análisis de escenarios con baja irradiancia evidenció que los modelos no lineales mantuvieron un desempeño estable, mientras que las aproximaciones lineales exhibieron incrementos significativos de error, corroborando la superioridad de los métodos basados en ensambles y memoria interna.

La aplicación sistemática de estas etapas demostró, por tanto, que el proceso comparativo cumplió con los requisitos de exhaustividad, objetividad y robustez establecidos en el objetivo secundario, proporcionando una base empírica sólida para seleccionar el algoritmo óptimo en función de la precisión requerida y las restricciones operativas del sistema fotovoltaico.

6.3 OBJETIVO DE VALIDACIÓN MULTIPLANTA

El objetivo de validación multiplanta se abordó mediante la comprobación empírica de que un esquema de modelo único, complementado con un factor de escala proporcional a la potencia nominal, resulta suficiente para generar pronósticos precisos en múltiples instalaciones fotovoltaicas situadas en una misma zona de radiación homogénea. Para ello se seleccionaron las cinco plantas de mayor producción, representando en su conjunto más del 90% de la generación eléctrica solar de Calgary. Las series temporales correspondientes al periodo 2016-2023 se sometieron a un preprocesado uniforme de interpolación bidireccional y normalización min-max, de manera que las diferencias inherentes a la escala de cada central quedaran absorbidas en la fase de reescalado posterior.

El modelo Long Short-Term Memory, se entrenó sobre los registros de las cinco plantas. Tras la fase de aprendizaje, se aplicó un factor de escala determinado por la ratio entre la capacidad nominal de cada planta destino y la capacidad de la planta de referencia. Esta estrategia permitió reutilizar los pesos entrenados sin necesidad de ajustes adicionales, dado que la forma normalizada de la señal ya contenía la información estacional y de dependencia temporal común al conjunto de infraestructuras.

La validación se efectuó con los datos correspondientes al año 2022, manteniéndose el corte cronológico previamente descrito. En resolución intradiaria se obtuvo un error absoluto medio promedio de 6,2% respecto a la potencia nominal y un RMSE de 15,476 kWh. El coeficiente de determinación se situó consistentemente por encima de 0,90 en las cinco instalaciones, lo que corroboró la adaptabilidad del modelo al margen de la potencia instalada. La ausencia de degradación significativa del desempeño en los emplazamientos extremos del conjunto validó la hipótesis de homogeneidad radiativa y confirmó la robustez del factor de escala como único parámetro de ajuste.

Estos resultados demuestran que el objetivo de validación multiplanta se ha cumplido de forma satisfactoria. Se ha verificado que un único modelo, ajustado sobre una planta representativa y escalado únicamente con la potencia pico de cada instalación, reproduce con precisión la dinámica de generación de todas las centrales estudiadas sin requerir datos exógenos ni otro entrenamiento específico. En consecuencia, la solución propuesta se consolida como una alternativa ligera y escalable para entornos donde la expansión eficiente del sistema predictivo a nuevas plantas constituye un requisito esencial.

6.4 OBJETIVO DE CONCLUSIONES Y TRABAJOS FUTUROS

En primer lugar, se expone con claridad la ruta metodológica seguida, condensando los resultados más relevantes obtenidos a lo largo del trabajo: la regresión lineal y sus variantes polinómicas fueron insuficientes para reproducir una serie fotovoltaica altamente estacional; la introducción de la predicción recursiva puso de manifiesto la acumulación de error inherente a los enfoques que no modelan de forma explícita la dependencia temporal; la estrategia SARIMA, redujo la magnitud de los errores pero dejó varianza sin explicar; el Bosque Aleatorio demostró la utilidad de los modelos no paramétricos para capturar umbrales y discontinuidades; y, finalmente, la red Long Short-Term Memory ofreció la menor dispersión de los residuos gracias a su capacidad para retener información a medio y largo plazo. De este modo, el texto recoge y ordena los hallazgos empíricos que justifican cada transición tecnológica.

En segundo término, el capítulo identifica de forma argumentada el algoritmo que mejor concilia precisión, fiabilidad y viabilidad operativa: la arquitectura dual LSTM, especializada de manera independiente para los horizontes intradía y anual. La red recurrente registra el menor RMSE, de 15,476 kWh a quince minutos y 18,541 kWh en agregación anual.

Por último, el documento formula un programa de trabajos futuros alineado con las tendencias actuales de la predicción fotovoltaica. Se propone incorporar variables meteorológicas procedentes de satélites y estaciones terrestres, explorar arquitecturas profundas basadas en convolución dilatada y mecanismos de atención multicabeza, implantar aprendizaje incremental para reflejar la evolución del proceso físico en tiempo real e introducir enfoques probabilísticos que proporcionen bandas de confianza calibradas. Asimismo se plantea la transferencia de aprendizaje a otras zonas climáticas, la generación de pronósticos sub-hora y la integración de indicadores de degradación de los módulos. Estas líneas marcan una hoja de ruta coherente que amplía el alcance de la investigación y garantizan la continuidad del estudio más allá de los límites de este Trabajo de Fin de Grado.

Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

El presente Trabajo de Fin de Grado ha recorrido una trayectoria metodológica ascendente que comenzó con modelos lineales elementales y culminó con arquitecturas de memoria recurrente. La comparación sucesiva de enfoques permitió constatar que la simplicidad analítica de la regresión lineal se reveló insuficiente para capturar la complejidad inherente a la serie fotovoltaica, ya que las transformaciones polinómicas únicamente ofrecieron mejoras marginales. A partir de esa constatación se introdujo la predicción recursiva, la cual evidenció la rápida acumulación de errores cuando la dinámica temporal no se representa de manera explícita.

Sobre dicha limitación se justificó la transición hacia un modelo autorregresivo integrado de medias móviles, en el que se aplicó por primera vez una descomposición temporal diferenciando la resolución diaria de la anual. La integración de componentes estacionarios e integrados permitió reducir de forma apreciable la magnitud absoluta de los errores; sin embargo, el análisis residual mostró que persistía una proporción significativa de varianza no explicada, lo que motivó la búsqueda de herramientas capaces de representar relaciones no lineales.

El siguiente salto conceptual se materializó con la adopción del Bosque Aleatorio. La aleatorización múltiple de los árboles propició la detección de umbrales y discontinuidades que los modelos autorregresivos no alcanzaron a describir. Este método introdujo además la posibilidad de estimar la importancia relativa de cada variable, factorizando la posición intradía y la fase estacional dentro de un mismo marco predictivo. La estrategia de entrenamiento independiente para los horizontes diario y anual confirmó que la segmentación multiescala constituye un requisito indispensable cuando se busca un equilibrio entre granularidad operativa y coherencia estratégica.

El ciclo de refinamiento se cerró con las redes Long Short-Term Memory. La capacidad de estas arquitecturas para retener información a medio y largo plazo permitió extraer dependencias que no emergieron con los árboles de decisión. La optimización rigurosa de hiperparámetros, combinada con técnicas de regularización y validación temporal, posibilitó la obtención de un modelo que unificó precisión y robustez, reduciendo la dispersión de los errores y estrechando los intervalos de predicción. Así, la red neuronal LSTM se consolidó como la alternativa más adecuada para la operación intradía, mientras que su contraparte anual ofreció un soporte consistente a la planificación a largo plazo.

Como conclusión general, puede afirmarse que la evolución desde técnicas lineales hasta redes de memoria recurrente demuestra la importancia de incrementar gradualmente la expresividad de los modelos en problemas dominados por estacionalidades complejas y patrones no lineales.

7.2 TRABAJOS FUTUROS

La extensión futura del presente trabajo deberá fundamentarse en la incorporación de variables exógenas procedentes de satélites y estaciones meteorológicas, de modo que se cuantifique la influencia de la irradiancia global horizontal, la temperatura ambiente o la humedad relativa sobre la exactitud de las predicciones univariantes. Para tal fin, se propone aplicar análisis de correlación y métodos de selección de características a fin de aislar los indicadores con mayor poder explicativo y reducir la dimensionalidad sin pérdida de información relevante.

Asimismo, se estima pertinente explorar arquitecturas de aprendizaje profundo que trascienden a las empleadas hasta la fecha. Las redes convolucionales temporales y los modelos basados en mecanismos de atención multicabeza adaptados a series temporales presentan la capacidad de capturar dependencias de largo alcance y de modelar no linealidades complejas bajo condiciones de elevada variabilidad climática. Su comparación con las LSTM permitiría dilucidar hasta qué punto la convolución dilatada o la atención distribuida mejoran la representación de los patrones estacionales residuales.

Con el objetivo de adaptar el modelo a la evolución continua del proceso físico, se recomienda implantar estrategias de aprendizaje incremental que actualicen los parámetros a medida que se reciban nuevas observaciones; de este modo se reflejarían con inmediatez las oscilaciones derivadas de eventos meteorológicos extremos o de cambios operativos, sin incurrir en los costes computacionales de un reentrenamiento íntegro. En paralelo, la adopción de enfoques bayesianos o probabilísticos facilitaría la obtención de bandas de confianza calibradas, útil para la programación de reservas y la evaluación de riesgos en la operación del sistema eléctrico.

La generalización geográfica del modelo podría abordarse mediante transferencia de aprendizaje, aprovechando pesos preentrenados en regiones con alta variabilidad estacional y ajustándolos a zonas templadas o desérticas con un esfuerzo de recalibración reducido. A tal fin, la optimización de hiperparámetros deberá llevarse a cabo mediante algoritmos bayesianos o evolutivos que automaticen la búsqueda de configuraciones idóneas y mitiguen el riesgo de sobreajuste en históricos limitados.

Resulta igualmente recomendable la generación de pronósticos sub-hora, en intervalos de cinco o diez minutos, a fin de atender exigencias de control de tensión y respuesta rápida ante picos de demanda, particularmente cuando las instalaciones incluyan sistemas de almacenamiento. Finalmente, la integración de información sobre mantenimiento y degradación de los módulos fotovoltaicos permitiría introducir variables de estado de salud en el proceso de inferencia y anticipar pérdidas de rendimiento asociadas al envejecimiento o a incidencias técnicas.

En conjunto, estas líneas de investigación constituyen un marco coherente orientado a reforzar la robustez operativa, la precisión predictiva y la transferibilidad regional de los sistemas de pronóstico fotovoltaico basados en registros históricos de generación, garantizando su utilidad tanto en la planificación estratégica como en la gestión táctica del recurso solar.

Capítulo 8. BIBLIOGRAFÍA

[1] O. Bamisile et al., "Long-Term Prediction of Solar Radiation Using XGBoost, LSTM, and Machine Learning Algorithms," in Proc. Asia Energy and Electrical Engineering Symposium, 2022.

[2] Q. T. Phan et al., "Short-term Solar Power Forecasting Using XGBoost with Numerical Weather Prediction," in Proc. IEEE Int. Future Energy Electronics Conf. (IFEEC), 2021.

[3] S. T. Asiedu, F. K. A. Nyarko, S. Boahen, F. B. Effah, and B. A. Asaaga, "Machine learning forecasting of solar PV production using single and hybrid models over different time horizons," Heliyon, vol. 10, no. 7, Art. e28898, 2024.

[4] Y. Zhang, H. Li, and S. Wang, "Short-Term Forecasting of Photovoltaic Power Generation Using SARIMA Models," *Solar Energy*, vol. 204, pp. 112-123, 2021.

[5] R. Martínez, P. González, and L. Sánchez, "Random Forest and XGBoost Models for Solar PV Power Output Forecasting," *Renewable Energy Applications*, vol. 15, no. 2, pp. 85-98, 2022.

[6] O. Bamisile et al., "Investigating Photovoltaic Solar Power Output Forecasting Using Machine Learning Algorithms," *Engineering Applications of Computational Fluid Mechanics*, vol. 16, no. 1, pp. 2002-2034, 2022.

[7] N. L. M. Jailani et al., "Investigating the Power of LSTM-Based Models in Solar Energy Forecasting," *Processes*, vol. 11, no. 5, Art. 1382, 2023.

[8] F. D. Campos et al., "Short-Term Forecast of Photovoltaic Solar Energy Production Using LSTM," *Energies*, vol. 17, no. 11, Art. 2582, 2022.

[9] H. Niska et al., "Evaluating Neural Network Models in Site-Specific Solar PV Forecasting Using NWP Data and Observations," *Renewable Energy*, vol. 205, pp. 248-259, 2023.

[10] Solcast. *Solar API and Weather Forecasting Tool*. 2025. Disponible en: https://solcast.com/ (consulta: 11-06-2025)

[11] Solcast. *Solar Live & Forecast Data API*. 2025. Disponible en: https://solcast.com/data-specifications (consulta: 11-06-2025)

[12] Clean Power Research. *SolarAnywhere Forecast: Solar Energy Forecasts for Solar Operations and Maintenance*. 2024. Disponible en: https://www.solaranywhere.com/products/solaranywhere-forecast/ (consulta: 11-06-2025)

[13] Clean Power Research. *SolarAnywhere - Trusted Global Solar Data & Intelligence Services*. 2025. Disponible en: https://www.solaranywhere.com/ (consulta: 13-06-2025)

[14] Solargis. *Solar forecasts and solar prediction - Data specs*. 2025. Disponible en: https://solargis.com/products/forecast/data-specs (consulta: 13-06-2025)

[15] Solargis. *Predict the output of your solar power project*. 2025. Disponible en: https://solargis.com/solutions/power-output-forecast (consulta: 13-06-2025)

[16] Solargis Knowledge Base. *Forecast modelling approach*. 2024. Disponible en: https://kb.solargis.com/docs/forecast-modelling-approach (consulta: 14-06-2025)

[17] Solargis. *Solargis Forecast - Features*. 2025. Disponible en: https://solargis.com/products/forecast/features (consulta: 15-06-2025)

[18] meteocontrol GmbH. *SOLAR POWER FORECAST - Whitepaper*. 2023. Disponible en: https://www.meteocontrol.com/fileadmin/Daten/Dokumente/EN/3_Technische_Beratung_Prognosen/2_Prognosen/Solar_Power_Forecasting/Whitepaper_Solar_Power_Forecast_en.pdf (consulta: 15-06-2025)

[19] meteocontrol GmbH. *Solar power forecast 100 MWp - Reference*. 2025. Disponible en: https://www.meteocontrol.com/en/company/references/solar-power-forecast-100-mwp (consulta: 15-06-2025)

[20] meteocontrol GmbH. *Grid stability thanks to accurate solar energy online feed-in and forecasts*. 2025. Disponible en: https://www.meteocontrol.com/en/company/news/detail/meteocontrol-grid-stability-thanks-to-accurate-solar-energy-online-feed-in-and-forecasts (consulta: 15-06-2025)

[21] IBM. *Generate efficient wind and solar energy with high accuracy forecasts*. 2024. Disponible en: https://mediacenter.ibm.com/media/Renewables+Forecasting%3A+Generate+efficient+wind+and+solar+energy+with+high+accuracy+forecasts/1_9mxebtm6 (consulta: 15-06-2025)

[22] IBM. *Maximo Renewables Asset Performance Management - Environmental Intelligence Suite*. 2025. Disponible en: https://www.ibm.com/products/maximo/renewables (consulta: 15-06-2025)

[23] IBM. *IBM Renewables Forecasting Platform Demo*. 2024. Disponible en: https://mediacenter.ibm.com/media/IBM+Renewables+Forecasting+Platform+Demo/1_2be0rh1n (consulta: 15-06-2025)

[24] SolarEdge. *Weather Guard - Home Backup Feature*. 2025. Disponible en: https://www.solaredge.com/us/weather-guard-home-backup (consulta: 15-06-2025)

[25] OpenClimateFix. *Quartz Solar Forecast* (repositorio GitHub). 2023. Disponible en: https://github.com/openclimatefix/open-source-quartz-solar-forecast (consulta: 15-06-2025)

[26] pvlib-python contributors. *pvlib-python* (repositorio GitHub). 2025. Disponible en: https://github.com/pvlib/pvlib-python (consulta: 15-06-2025)

ANEXO I: ALINEACIÓN DEL PROYECTO CON LOS ODS

Se presenta con detalle la alineación del proyecto de predicción de la producción fotovoltaica mediante aprendizaje automático con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 de las Naciones Unidas. La exposición se articula en párrafos corridos y utiliza construcciones impersonales para garantizar el rigor académico y la cohesión interna.

En primer lugar, se sitúa el marco internacional en el que se inscribe la investigación. En el año 2015, los Estados Miembros de las Naciones Unidas adoptaron la Agenda 2030, que comprende diecisiete objetivos orientados a promover un desarrollo sostenible capaz de satisfacer las necesidades de la generación presente sin comprometer las de las generaciones futuras. Esta iniciativa global tiene por propósito abordar retos sociales, económicos y medioambientales mediante metas cuantificables que deben alcanzarse antes de 2030.

Con respecto al ODS 7, Energía Asequible y no Contaminante, el proyecto mejora la eficiencia en la gestión de la generación solar al desarrollar un modelo de predicción basado exclusivamente en series temporales históricas de producción. De este modo, se facilita la planificación óptima de la operación de la red eléctrica y se reduce la necesidad de recurrir a fuentes de respaldo fósiles, lo que contribuye a incrementar la proporción de energías limpias en la matriz energética . Además, al no requerirse instrumentación meteorológica adicional, la solución resulta asequible y replicable en entornos con recursos limitados.

En relación con el ODS 9, Industria, Innovación e Infraestructura, se subraya la implementación de una arquitectura ligera y escalable para pronósticos fotovoltaicos. La utilización de modelos de aprendizaje automático entrenados con datos históricos y la posibilidad de validación cruzada temporal demuestran la viabilidad de soluciones tecnológicas de bajo coste, fáciles de mantener y de rápida puesta en servicio . Asimismo, la

capacidad de adaptar el modelo único con factor de escala a múltiples plantas sin añadir nuevas variables externas favorece la replicabilidad en distintos contextos geográficos y comerciales.

En cuanto al ODS 13, Acción por el Clima, se resalta que la reducción de incertidumbres en los pronósticos de generación fotovoltaica permite optimizar la programación de reservas energéticas y minimizar las emisiones de gases de efecto invernadero asociadas a las plantas térmicas de respaldo. Se considera que una predicción más fiable de la producción solar ayuda a mitigar el cambio climático al facilitar la integración de renovables en la red y al disminuir la dependencia de combustibles fósiles.

El ODS 11, Ciudades y Comunidades Sostenibles, recibe apoyo al mejorar la estabilidad de redes que incorporan energía solar distribuida. La herramienta favorece la resiliencia de sistemas eléctricos urbanos y periurbanos, reduciendo la vulnerabilidad ante fluctuaciones de demanda y eventos extremos de generación . De este modo, se contribuye al diseño de entornos urbanos más sostenibles y adaptativos frente a desafíos energéticos.

Por otra parte, el ODS 8, Trabajo Decente y Crecimiento Económico, se fortalece mediante la generación de pronósticos de fácil acceso y bajo coste que incrementan la competitividad del sector fotovoltaico. Al reducir los costes operativos asociados a la regulación de la red y la gestión de plantas de respaldo, se fomenta la creación de valor en empresas dedicadas al despliegue y mantenimiento de instalaciones renovables . Se prevé que la adopción de estos sistemas predictivos promueva empleos de alta cualificación en el ámbito de la ingeniería de datos y la energía.

Adicionalmente, la investigación aporta al ODS 17, Alianzas para Lograr los Objetivos, al utilizar bibliotecas de código abierto y compartir la metodología mediante repositorios públicos. Esta práctica fortalece la cooperación científico-técnica y facilita la transferencia de conocimiento entre comunidades académicas e industriales .

En síntesis, la presente investigación se alinea directamente con los ODS 7, 8, 9, 11, 13 y 17, integrando la innovación tecnológica con la sostenibilidad energética, la acción

climática, el desarrollo económico y la colaboración internacional. De este modo, se confirma que el proyecto no solo avanza en el ámbito científico-técnico, sino que también contribuye de manera concreta a los objetivos globales de la Agenda 2030.