TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2024 257

Ana Patricia Talayero Navales

Técnicas de aprendizaje automático en plantas fotovoltaicas

Director/es
Llombart Estopiñán, Andrés
Melero Estela, Julio Javier

http://zaguan.unizar.es/collection/Tesis

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

TÉCNICAS DE APRENDIZAJE AUTOMÁTICO EN PLANTAS FOTOVOLTAICAS

Autor

Ana Patricia Talayero Navales

Director/es

Llombart Estopiñán, Andrés
Melero Estela, Julio Javier

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

Programa de Doctorado en Energías Renovables y Eficiencia Energética

2024

Repositorio de la Universidad de Zaragoza – Zaguan http://zaguan.unizar.es

Universidad Zaragoza
1542

Tesis Doctoral

Técnicas de aprendizaje automático en plantas fotovoltaicas

Autor

Ana Patricia Talayero Navales

Director/es

Julio J. Melero Estela

Andrés Llombart Estopiñan

Instituto Universitario de Investigación Mixto CIRCE

DICIEMBRE 2023

Lista de contribuciones

A continuación, se enumeran las contribuciones científicas relacionadas con la tesis.

[1]*Machine Learning models for the estimation of the production of large utility-scale photovoltaic plants*
Autores: A. Talayero, Julio J. Melero, A. Llombart, Y. Yürüşen
Solar Energy 2023 https://doi.org/10.1016/j.solener.2023.03.007

[2].*Anomaly detection at inverter level via machine learning algorithms under the absence of O&M logbooks*
Autores: A. Talayero, Y. Yürüşen, Julio J. Melero, A. Llombart
Congreso European PV Solar Energy Conference and Exhibition (PVSEC20) 2020
https://doi.org/10.4229/EUPVSEC20202020-5DO.4.2

[3].*Diagnosis of failures in Solar Plants based on Performance monitoring*
Autores: A. Talayero, Julio J. Melero, A. Llombart
International Conference on Renewable Energies and Power Quality (ICREPQ'20) Renewable Energy and Power Quality Journal https://doi.org/10.24084/repqj18.248

[4]. *Operation and Maintenance in Solar Plants: Eight study cases*
Autores: A. Talayero, Julio J. Melero, A. Llombart, A. Casado
International Conference on Renewable Energies and Power Quality (ICREPQ'18)
Renewable Energy and Power Quality Journal https://doi.org/10.24084/repqj16.363

[5] *Apriori and K-Means algorithms of machine learning for spatio-temporal solar generation balancing*
Autores: Nurseda Y. Yürüşen , Bahri Uzunoğlu , Ana P. Talayero, Andrés Llombart Estopiñán
Renewable Energy Elsevier 2022 https://doi.org/10.1016/j.renene.2021.04.098

[6] *Mejora de las estrategias de mantenimiento en plantas de generación renovable a partir de los datos SCADA*
Autores: A. Talayero, Julio J. Melero, Nurseda Y. Yurusen
Congreso VI Smart Grids 2019

Otras contribuciones científicas pueden consultarse en https://orcid.org/0000-0001-9823-4777

Agradecimientos

Deseo expresar mi más profundo agradecimiento a todas las personas e instituciones que han contribuido de manera significativa a la realización de esta tesis doctoral.

En primer lugar, deseo expresar mi gratitud a mis directores, Julio J. Melero y Andrés Llombart. De Julio, valoro enormemente su sabiduría y los debates que han enriquecido el enfoque de esta investigación. De Andrés, aprecio su constante disposición y orientación temática, siempre animándome y ayudándome a superar obstáculos. A ambos, les agradezco su paciencia infinita y apoyo constante durante este proceso. Sus conocimientos y consejos han sido fundamentales para dar forma y desarrollar las ideas de esta investigación.

Mi gratitud también se extiende a mi familia, a quien dedico esta tesis, cuyo amor incondicional y generosidad no han tenido límites. Han sido muchos años de sacrificios, siempre acompañados de una sonrisa y palabras de aliento por su parte. Sin duda, sin ellos, no hubiera tenido la fortaleza para continuar y terminar.

Agradezco al Instituto Universitario de Investigación Mixto CIRCE, pionero en ofrecer una línea de doctorado centrada en las energías renovables, esencial para permitir la evolución y los avances científicos tan necesarios en este campo, el aceptar desarrollar esta tesis. También agradezco a la Fundación CIRCE, centro tecnológico donde trabajo, su compromiso con el desarrollo científico y el fomento de las tesis doctorales. En particular, agradezco a Enrique mi superior más directo, su apoyo durante todos estos años y el permitirme compaginar el trabajo diario con la realización de la tesis.

Quisiera también mencionar en mis agradecimientos a mis compañeros de trabajo, quienes han estado presentes durante estos años. Todos ellos han soportado mis nervios, tensiones y quejas en los momentos críticos. En especial a Roberto, con quien comencé la tesis al mismo tiempo, durante todo este tiempo hemos compartido ideas y me ha escuchado mis preocupaciones.

Finalmente, dedico un agradecimiento especial a todas las fuentes bibliográficas y académicas que consulté durante este trabajo. Sus investigaciones allanaron el camino para mi estudio y ampliaron mi comprensión del campo.

En resumen, esta tesis no habría sido posible sin el apoyo y contribución de tantas personas e instituciones. Estoy sinceramente agradecido por todas sus contribuciones.

Resumen

En esta tesis se ha analizado el comportamiento de modelos de aprendizaje automático para estimar la producción y la irradiancia en instalaciones fotovoltaicas de gran tamaño a partir de variables principalmente meteorológicas medidas en el propio emplazamiento. La necesidad de conocer estas variables con una mayor precisión se sustenta en el desarrollo que ha alcanzado la energía fotovoltaica en los últimos años y en el tamaño de sus plantas. El conocimiento de estas variables servirá para mejorar la eficiencia operativa y facilitar el cálculo del performance ratio de las plantas fotovoltaicas.

En el estudio, se lleva a cabo una revisión bibliográfica de los métodos que actualmente están dando mejores resultados en la modelización de estas variables y se seleccionan como métodos de trabajo "Random Forest", "Gradient Boosting" y las redes neuronales “Multi Layer Perceptrons”. Adicionalmente, se trabaja con la regresión lineal múltiple, que se utilizará como modelo de referencia debido a su simplicidad.

Se describen los principios matemáticos de los algoritmos seleccionados, así como de sus variantes probabilísticas utilizadas para calcular los intervalos de confianza de los modelos y las técnicas de búsqueda que permiten ajustar los hiperparámetros de los modelos.

El trabajo consiste en aplicar las metodologías a tres inversores diferentes de plantas distintas para predecir la producción y a una estación meteorológica para predecir la irradiancia. Los modelos en el caso de la producción presentan una buena precisión y baja incertidumbre, incluso la regresión lineal múltiple. En el caso de la irradiancia, la regresión lineal no es óptima y son necesarios modelos más complejos para captar las relaciones entre las variables.

Analizando los resultados, el modelo “Gradient Boosting” resulta la mejor opción presentando los errores más bajos con una amplitud de intervalo pequeña, además de ser el modelo más ágil y con menos demanda computacional.

El estudio revela que los modelos de aprendizaje automático representan una mejora significativa en la predicción de la producción e irradiancia con respecto a otras alternativas y que cada inversor o estación meteorológica necesita su propio modelo de comportamiento, no pudiéndose aplicar los modelos estudiados a otras plantas fotovoltaicas. Sin embargo, el proceso que se ha llevado a cabo y las metodologías definidas, sí que puede ser extrapolable a otras plantas ubicadas en otros lugares.

Índice

Lista de Figuras

Lista de Tablas

Lista de Abreviaturas y símbolos

Símbolo	Descripción
AC	Corriente Alterna
ACO	“Ant Colony Optimisation”
AHPA	“Adaptative Hyper Paramter Adjustment”
ANFIS	“Adaptive Network-based Fuzzy Inference Systems”
ANN	Redes Neruonales
AR	“Auto-Regressive”
ARIMA	“Auto-Regressive Integrated Moving-Average”
ARIMAX	“Auto-Regressive Integrated Moving-Average with eXogenous variables”
ARMA	“Auto-Regressive Moving-Average”
ARX	“Auto-Regressive with eXogenous variables”
BSRN	“Baseline Surface Research Network”
B	Presión
CC	Corriente Continua
CMA-ES	“Covariance Matrix Adaptation-Evolutionary Estrategy”
Dir	Dirección del viento
DL	Aprendizaje Profundo
DT	Árboles de decisión
ELM	“Extreme Learning Machines”
ESRA	“European Solar Radiation Atlas”
GB	“Gradient Boosting”
GOA	“Grass Hopper Optimization Algorithm”
h	Hora
HPO	Optimización de Hiperparámetros
HPO	Humedad
IA	Inteligencia Artificial
IEA	Agencia Internacional de la Energía
IR	Irradiación
$IR_{teorica}$	Irradiancia Teórica
IRENA	Agencia Internacional de las Energías Renovables
KF	“Dynamic Kalman Filters”

KNN	“K-Nearest Neighbours”
LCOE	“Levelized Costof Energy”
LSTM	“Long Short Term Memory neural network”
m	Mes
MA	Media Movil
ML	Aprendizaje Automático
MLFFNN	“MultiLayer FeedForward Neural Network”
MLR	Regresión Lineal Multivariable
MPNN	“Multilayer Perceptron Neural Networks”
NARXNN	“Nonlinear AutoRegressive eXogenous Neural Network”
nBias	Sesgo normalizado
nMAE	Error Absoluto Medio Normalizado
Nº Árboles	Número de árboles
Nº Capas	Número de Capas
Nº Neuronas	Número de Neuronas
Nº Variables	Número de variables
nRMSE	Raíz del Error Cuadrático Medio Normalizado
NWP	“Numerical Weather Prediction”
Pac	Potencia de salida del inversor
Pdc	Potencia de entrada al inversor
PNIEC	Plan Nacional Integrado de Energía y Clima
PPAs	“Power Purchase Agreements”
PPI	“Parametric Prediction Interval”
PR	“Performance Ratio”
Profundidad	Profundidad del Árbol
PV	Planta Fotovoltaica
QR	“Quantile Regression”
QRF	“Quantile Regression Forest”
ReLU	“Rectified Linear Unit”
RF	“Random Forest”
RNN	“Recurrent Neural Network”
RSS	Suma Residual de Cuadradod
RT	“Regression Trees”
SARIMA	“Seasonal Auto-Regressive Integrated Moving-Average”
SCADA	“Supervisory Control And Data Acquisition”
SS	“Skill Score”
SVM	“Support Vector Machine”
SVR	“Support Vector Regression”
T. Aprendizaje	Tasa de aprendizaje
T. Descarte	Tasa de descarte
T. Descarte inicial	Tasa de descarte inicial
T. Mauestreo	Tasa de muestreo
Tamb	Temperatura ambiente

Tmod	Temperatura módulo
TPE	“Tree Structured Parzen Estimators”
Vel	Velocidad del viento
WANN	“Wavelets and Artificial Neural Networks”
WARIMAX	“Wavelet Auto-Regressive Integrated Moving-Average with eXogenous variables”
μ	Media de una distribución lineal
σ	Desviación típica de una distribución normal
τ	Cuantil

1. Introducción

1.1. Contexto y motivación

Evolución de la energía fotovoltaica

La energía fotovoltaica se basa en el efecto fotovoltaico, descubierto en 1839 por Becquerel. Mediante experimentos se observó que cuando un fotón impacta en un electrón, éste sale despedido liberando un hueco que es reemplazado por otro electrón libre, lo que genera un flujo de cargas eléctricas que se convierte en corriente eléctrica. Este efecto fue observado en el Selenio, pero su bajo rendimiento y altos costes limitaron su aplicación práctica, hasta que un siglo después se descubrió que el Silicio tenía la misma capacidad y un mejor rendimiento.

A mediados del siglo XX, el Silicio se convirtió en la solución energética para los proyectos espaciales, la inversión económica que se llevó a cabo en ellos le permitió a la fotovoltaica avanzar tecnológicamente y, desde entonces, esta energía ha experimentado un gran desarrollo, convirtiéndose en una de las energías renovables más competitiva en términos de coste, fiabilidad, predictibilidad y adecuación de la generación a la demanda. Este desarrollo permitió que al final del 2022 se alcanzaran a nivel mundial los 1050 GW de potencia instalada de energía fotovoltaica, según la Agencia Internacional de las Energías Renovables (IRENA) [7]. Además, la capacidad de potencia nueva instalada cada año está acelerándose, con un ritmo de crecimiento del sector del 20% anual [7,8].

Se espera que este crecimiento se mantenga en el futuro, ya que las plantas fotovoltaicas cada vez son de mayor tamaño y se están posicionando como una posible solución al problema mundial de abastecimiento de agua y energía [9]. Esto se debe, en parte, a la reducción de costes, ya que la energía fotovoltaica es más barata que la energía generada por plantas de combustibles fósiles, en términos de coste nivelado de la energía "Levelized Cost of Energy" (LCOE), con un valor de 3 c€/kWh en muchas partes del mundo y una perspectiva de mantenerse entre 4 y 5 c€/kWh en el largo plazo [10,11]. Además, el marco jurídico político, que incluye contratos de compraventa de energía a largo plazo "Power Purchase Agreements" (PPAs) y subastas de energía y capacidad, garantiza la viabilidad económica de los proyectos [12,13].

El país líder del sector es China, con un tercio de la potencia mundial instalada. Le sigue el continente europeo con el 19% de la potencia total instalada y una capacidad de crecimiento anual superior al 30%. Dentro de los países europeos, es Alemania quien lidera la lista de países con mayor instalación acumulada, seguida por Italia y España [14].

Debido al marco retributivo, España estuvo prácticamente estancada durante una década en los 4,5 GW de potencia fotovoltaica instalada. Esta tendencia comenzó a cambiar en 2018, y se consiguió superar los 25 GW a finales del 2022 [15]. Además, se espera que para 2030, España tenga instalados casi 40 GW de energía fotovoltaica, ya que esta tecnología, además de las ventajas mencionadas, está siendo impulsada por el Plan Nacional Integrado de Energía y Clima 2021-2030 (PNIEC) [16], que está incluido en el marco del Reglamento de Gobernanza de la Unión Europea para la Energía y la Acción por el Clima [17]. El PNIEC establece un objetivo ambicioso de alcanzar el 74% de la generación de electricidad a partir de fuentes renovables para 2030, lo que implica una importante expansión de la capacidad fotovoltaica instalada en España.

En resumen, la energía fotovoltaica está experimentando un crecimiento significativo en todo el mundo y se espera que siga creciendo en el futuro, impulsada por la disminución de los costes, la viabilidad económica de los proyectos y la adopción de políticas favorables. De entre los tipos de instalaciones que se están promoviendo destacan, por volumen, las grandes instalaciones conectadas a red que llegan a decenas e incluso centenas de megavatios en una única planta.

Introducción a las plantas fotovoltaicas

Una planta fotovoltaica es la instalación que genera energía eléctrica a partir de la radiación solar. Está compuesta por varios elementos [18], siendo el módulo fotovoltaico el elemento generador que transforma la energía de los fotones en energía eléctrica a través de las células de Silicio de las que está formado. Los módulos fotovoltaicos pueden ser monocristalinos, policristalinos o de capa fina ("thin film"), dependiendo de la morfología de la célula [18]. Los módulos generan electricidad en forma de corriente continua (CC) que hay que convertir en corriente alterna (AC) para su transporte, distribución y utilización.

Los módulos fotovoltaicos se montan generalmente sobre estructuras móviles con uno o dos ejes para maximizar la irradiación captada. Estas estructuras pueden ser polares, acimutales u horizontales. También pueden situarse sobre una estructura fija con cierta inclinación para optimizar la producción en una ubicación específica.

Varias decenas de módulos se conectan en serie para formar una cadena conocida como “string”. Los “strings” se conectan a un inversor para convertir la corriente continua en corriente alterna. Existen dos modalidades de conexión: con un inversor centralizado o con un “string inverter”.

Cuando se utiliza un inversor centralizado, unos cuantos “strings” se agrupan en paralelo en las cajas “string boxes”, que son agregadores de corriente. Varias cajas se unen a un mismo inversor como se muestra en la Figura 1, donde puede verse la distribución de los módulos e inversores, Figura 1 (a), y un detalle de un inversor durante la fase de montaje de una planta Figura 1 (b). En este caso, dos inversores comparten transformador formando una estación de potencia, “Power Station”. Este sistema permite ahorrar los kilómetros de cable que serían necesarios para conectar individualmente los “string” al inversor.

Un “string inverter” es un inversor de pequeña potencia que agrega y transforma en corriente alterna unos pocos “strings”, sin necesidad de disponer de una caja, como se muestra en la Figura 2. En esta imagen, se puede ver cómo varias filas de módulos, que forman cada una de ellas un “string”, se conectan a un inversor. El cable de salida de este equipo llega hasta el transformador. El transformador no suele ser único para cada inversor, sino que suele centralizar varios inversores.

Una representación esquemática de la disposición de estos equipos se presenta en la Figura 3, en la cual, la parte (a) corresponde a una planta con inversor centralizado, donde se puede observar cómo se va agrupando la corriente continua en dos pasos a través de las cajas. La Figura 3 (b) corresponde a una planta con “string inverter”, la agrupación de corriente se hace en continua y en alterna. El área de todos los módulos que vierten a un mismo inversor se le denomina campo solar del inversor, y en la modalidad de conexión mediante “string invertir” el campo solar es de

menor tamaño y potencia lo que facilita su control.

(a) (b)

Figura 1. Planta fotovoltaica con inversor centralizado

Figura 2. Planta fotovoltaica con "string inverter"

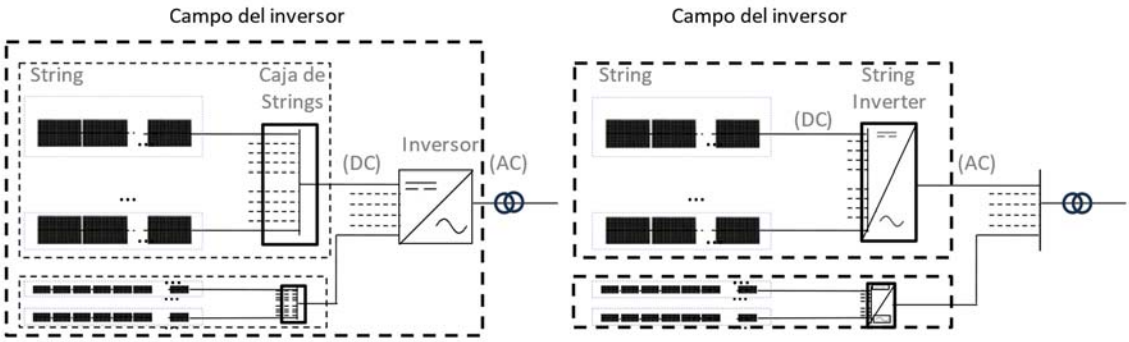


Figura 3. Diagrama distribución de equipos en una planta fotovoltaica (a) Inversor centralizado, (b) "String invertir"

Las plantas fotovoltaicas pueden estar dotadas de un sistema de monitorización y control. Este sistema recoge datos de distintas variables: medidas de corriente, tensión o potencia registradas en diferentes equipos de la planta y medidas meteorológicas como la irradiancia, la temperatura y la velocidad del viento recogidas en las estaciones de medición [19]. Al ser la instalación de proyectos fotovoltaicos un mercado cada vez más competitivo, el ahorro de inversión es fundamental, y se tiende a minimizar el número de sensores instalados. Por este motivo, la información disponible en las plantas suele encontrarse solo a nivel de inversor. En cuanto a las estaciones meteorológicas instaladas en las plantas, también tienden a reducirse, ubicándolas a grandes distancias unas de otras, lo que genera diferencias entre la irradiancia que reciben los módulos y la que captan los sensores de las estaciones, sobre todo para aquellos módulos que están alejados de las estaciones meteorológicas, y, que son la mayoría.

Las plantas fotovoltaicas en su funcionamiento están sujetas a degradaciones y fallos que ocasionan pérdidas de rendimiento e incluso paradas parciales o totales. Las incidencias más frecuentes que afectan a la producción de una planta se describen en la sección siguiente.

Incidencias en una planta fotovoltaica

En una planta fotovoltaica las pérdidas de energía son causadas principalmente por roturas, deterioro o desgastes de los módulos, sombras, fallos en las conexiones y en los cables o fallos en el inversor. La probabilidad de fallo de los distintos equipos de una planta fotovoltaica es baja, considerada de manera individualizada, pero al tener una planta fotovoltaica numerosos equipos conectados, la probabilidad de que alguno de los equipos tenga un fallo es muy alta. Por eso la probabilidad de fallo en una planta pequeña es muy inferior a la probabilidad de fallo en las plantas grandes [20].

Los fallos en las plantas fotovoltaicas de pequeño tamaño, han sido extensamente estudiados atendiendo a diferentes criterios, como el componente al que hacen referencia, su afección sobre la potencia, o el riesgo que suponen para la planta [18,21–25]. El criterio que se considera más interesante para diagnosticar una planta fotovoltaica es el basado en la probabilidad de ocurrencia en función de su vida útil [18,26], que discrimina los tipos de fallo en tres grandes grupos: fallos prematuros, fallos de vida media y fallos de final de vida.

Fallos prematuros (año 1):

Los fallos prematuros son los que se dan durante los primeros meses de explotación y suponen entre el 0.5% y el 5% de los fallos totales durante la vida útil de planta [18]. Están asociados generalmente al deterioro prematuro del material, a los procesos de fabricación [27–29], al transporte y montaje de los módulos [18,28]. Su número puede incrementarse con la ayuda de otros agentes como la humedad, la radiación ultravioleta o la temperatura [21,30–34].

Fallos durante la vida media (1 a 15 años) de la planta:

Los fallos de decoloración y degradación en los módulos representan el 10% del total de fallos que suceden en una planta y tienen mayor afección sobre la producción [35]. Los problemas de delaminación o rotura de las celdas corresponden a otro 10% de los fallos de la planta. Finalmente, la corrosión, que en mayor medida suele estar asociada a interconexión de las celdas, representa el 20% de la totalidad de los fallos [18].

La suciedad de los módulos, también debe analizarse en este periodo, ya que influye directamente en la producción de la planta llegando a afectar hasta en un 3% al rendimiento de la misma [18,28,36] y, además, puede ser detonante de otros fallos como puntos calientes, “mismatching”, u oxidaciones [37,38]. Un tipo concreto de suciedad es la que corresponde a sustancias orgánicas o biológicas como excrementos de pájaros, hojas, resina de árboles y musgos que se adhieren a las placas, produciendo no solo la disminución de producción sino también, fenómenos de oxidación [28,39].

La rotura parcial de un módulo, provoca también efectos sobre la producción que, en algunos casos, puede llegar a significar hasta el 8% del total del módulo [40] y puede además

desencadenar fallos mayores en la planta [41,42]. Las roturas pueden identificarse mediante técnicas de electroluminiscencia [40].

Los inversores son equipos que requieren especial atención por la repercusión económica y energética que supone su reparación o sustitución. Los fallos en estos equipos suelen ser causados por: aumento de temperatura, cortes de red de la planta, defectos o fallos del cableado, errores en el control del punto de máxima potencia (MPP), caída de rayos, o rotura del filtro [2,20]. Su vida útil es de 10 años [4] frente a los 20-25 años del resto de los componentes de las plantas, por lo que tienen que ser remplazados más de una vez durante la operación de la planta.

Los soportes y estructura sobre los que se disponen los módulos también son susceptibles de fallos. En el caso de que la planta tenga estructuras con seguidores solares, se pueden encontrar fallos mecánicos en los mismos que afecten a la orientación de los módulos, lo que puede disminuir drásticamente la producción [4,43].

Fallos de final de vida:

Pasados 15 años de vida media, el proyecto entra en sus últimos 5 -10 años de vida útil. En este periodo cobra una mayor importancia la degradación [44], que puede seguir la misma tendencia que durante la vida media de la planta o acelerarse ligeramente.

Durante este periodo son mucho más críticos los fallos asociados a roturas de celdas y delaminación, que se triplican, y la corrosión, que aparece de forma drástica pudiendo llegar a afectar al 50% de la planta [18]. También los fallos de “mismatching” se incrementan con el envejecimiento [45].

Todos estos fallos dan lugar a una pérdida de energía que puede ser sustancial si el fallo afecta a una parte importante de la planta o si perdura en el tiempo. Un estudio que cuantifica la energía perdida por tres incidencias diferentes en plantas fotovoltaicas, reveló que las sombras creadas por las hierbas que crecieron al retrasar el desbroce un mes, provocaron una pérdida equivalente a un 1.5 % de la producción de un año de la planta, mientras que el desalineamiento de los seguidores supuso un 6.7% de la energía de un año [4].

Las estaciones meteorológicas también están sujetas a fallos. En concreto, el sensor más delicado y sujeto a fallo es el piranómetro [46] que mide la radiación global que llega a un emplazamiento. Este sensor es también una parte sensible de la planta ya que sus medidas son importantes en el cumplimiento de contratos, cálculos de rendimiento y verificación de la disponibilidad de la planta.

En resumen, la energía fotovoltaica se ha convertido en una fuente de energía significativa en la actualidad y se espera que continúe creciendo en los próximos años. La instalación de grandes plantas solares contribuye a este crecimiento y consolida la energía fotovoltaica como un pilar importante en el panorama energético mundial.

Los sistemas fotovoltaicos están compuestos por numerosos componentes, lo que los hace susceptibles a fallos o errores. Estos fallos pueden dar lugar a interrupciones en la generación de energía provocando pérdidas e indisponibilidades en la planta.

1.2. Justificación y objetivo de la tesis

Para minimizar las pérdidas de energía e indisponibilidad de la planta solar, es esencial una gestión eficaz del mantenimiento y la implementación de tecnologías de supervisión. Estas tecnologías se orientan a la identificación rápida de las causas de los fallos y al cálculo de las pérdidas de energía asociadas a cada uno para mantener la productividad de las plantas fotovoltaicas y reducir las interrupciones. Esto se puede realizar si se dispone de un modelo que proporcione la producción de la planta bajo las condiciones de trabajo. De esta manera se compara la producción teórica con la real detectando así disminuciones de producción anómalas que permiten identificar y corregir las faltas en un periodo de tiempo corto reduciendo el OPEX ("Operating Expenses") de la planta [47] y aumentando así la producción anual, lo que repercute en el LCOE ("Levelized Costof Energy") del proyecto [48] y disminuye los riesgos financieros.

Los modelos se nutren de los datos de producción, medidos en los equipos de las plantas, y de las variables meteorológicas (principalmente la irradiancia solar). Una de las dificultades de estos modelos es la escasez de estaciones meteorológicas por lo que la mayoría de las veces la producción se debe asociar con irradiancias medidas en zonas alejadas de donde se está produciendo la transformación energética, originando mala correlación entre las magnitudes producción e irradiancia. La segunda dificultad principal es garantizar la calidad de las medidas de irradiancia.

La justificación de esta tesis se basa en la necesidad de aplicar técnicas de inteligencia artificial, específicamente modelos de aprendizaje automático y profundo, para abordar dos aspectos críticos en la producción de plantas fotovoltaicas: la caracterización de la producción y la garantía de calidad de los datos de irradiancia.

En primer lugar, mediante el uso de modelos de aprendizaje automático y profundo, se busca caracterizar la producción de las plantas fotovoltaicas de manera precisa. Esto implica que los modelos sean capaces de analizar relaciones complejas entre múltiples variables, como la irradiancia solar, la temperatura ambiente, y la temperatura de los módulos para predecir la energía generada por la planta en un momento dado. Una caracterización precisa de la producción permitirá a los operadores de la planta comprender mejor su funcionamiento y tomar decisiones informadas para optimizar su funcionamiento.

En segundo lugar, la aplicación de modelos de inteligencia artificial permitirá asegurar la calidad de los datos de irradiancia. La irradiancia solar es un factor crítico en la generación de energía fotovoltaica, y cualquier imprecisión o error en su medición puede afectar significativamente los cálculos de producción. Los modelos de aprendizaje automático y profundo pueden analizar los datos de irradiancia recopilados y detectar anomalías, corrigiendo o descartando mediciones erróneas.

Además, estos modelos ML también pueden utilizarse para simular la producción esperada de la unidad mínima monitorizada en potencia de la planta, como el "string" o la caja. De esta manera, se pueden detectar disminuciones en la energía generada a menor escala, lo que permite la detección temprana de problemas y mejorar significativamente el rendimiento y la disponibilidad

de las plantas fotovoltaicas.

1.3. Estado del arte de los modelos de aprendizaje automático aplicados a plantas fotovoltaicas

Introducción

Un modelo es una representación matemática de la realidad física de un problema. En los casos estudiados en esta tesis, caracterización de la producción y de la irradiancia, los modelos, trabajan con variables independientes, que son medidas en la planta fotovoltaica, y dan como resultado la producción a nivel de inversor o la irradiancia que son las variable objetivo o dependientes de las variables de entrada [49,50]. En general, los modelos se pueden clasificar en dos grandes grupos, paramétricos y no paramétricos.

Los modelos paramétricos están basados en las propiedades del sistema de estudio, tienen unos parámetros fijos y finitos y describen todo el proceso o comportamiento del sistema mediante ecuaciones analíticas [51,52] que se desarrollan en plataformas de cálculo como “Matlab” [53], y PSIM [54]. Un ejemplo de modelo paramétrico consistiría en obtener la potencia del módulo fotovoltaico como el producto de la corriente del módulo y su diferencia de potencial. Los modelos paramétricos encontrados se utilizan para caracterizar partes de la planta, como el sistema de seguimiento solar, [55] pero la parametrización completa de plantas de gran tamaño, es un trabajo muy costoso.

Los modelos no paramétricos consideran el sistema de estudio como una caja negra y relacionan las variables de entrada con la variable objetivo, en base a una metodología propia de cada algoritmo, es decir modelizan el comportamiento sin conocer propiamente el sistema [49,56].

Las técnicas utilizadas en los modelos no paramétricos, van desde las estadísticas [57,58], pasando por los algoritmos de control gráficos, “Control Chart” [59], hasta los algoritmos más desarrollados de Inteligencia Artificial (IA), capaces de aprender el comportamiento del sistema e incluso emular el cerebro humano. Los algoritmos de IA han mostrado, como veremos más adelante, una mejor adaptación a las necesidades de caracterización de las plantas fotovoltaicas.

La IA engloba algoritmos que resuelven problemas lógicos mediante unas reglas definidas por el programador [60–62], lo que se conoce como automatización de sistemas, pero también incluye a la familia de algoritmos de aprendizaje automático “Machine Learning” (ML) [63,64] objeto de este estudio.

El elemento diferencial del ML respecto de los modelos paramétricos, es la capacidad de averiguar las reglas implícitas existentes entre las variables del problema mediante un entrenamiento o aprendizaje que permiten a su vez volver a ser aplicadas a nuevos datos resolviendo problemas complejos [65]. Estos modelos trabajan con una serie de datos histórica que utilizan para llevar a cabo el aprendizaje.

Los modelos de ML incluyen un subgrupo que trabaja con relaciones más avanzadas, pudiendo

mejorar en muchos de los problemas los resultados, los denominados algoritmos de aprendizaje profundo “Deep Learning” (DL). Los modelos incluidos en este campo son las redes neuronales, “Artificial neural networks” (ANN) por su similitud a emular el funcionamiento del cerebro humano [66].

A la hora de aplicar los modelos de ML, es necesario identificar el tipo de problema que se quiere abordar, de clasificación o regresión. Se denominan problemas de clasificación aquellos cuya respuesta es la definición del grupo al que pertenecen las observaciones, es decir los datos son clasificados en diferentes grupos. Por ejemplo, estudiando la producción de una planta fotovoltaica podemos clasificar los datos en dos categorías, los correspondientes a una producción normal y los correspondientes a una producción anómala. En los modelos de regresión el resultado es un valor numérico dentro de un conjunto infinito de posibilidades. Por ejemplo la determinación de la producción de una planta fotovoltaica cuando trabaja bajo unas determinadas condiciones de radiación y temperatura [60].

Los modelos de aprendizaje automático también pueden ser supervisados o no supervisados, atendiendo a como realizan el aprendizaje. En el aprendizaje no supervisado el algoritmo estudia un conjunto de datos no etiquetados e identifica patrones que traduce en relaciones entre las variables [67].

En función de cómo se relacionan las variables independientes con la respuesta, los modelos de aprendizaje automático de regresión pueden ser lineales o no lineales. Los modelos lineales, tal y como su nombre indica establecen relaciones lineales entre las variables, como es el caso de la regresión lineal multivariable, mientras que los modelos no lineales establecen relaciones no lineales entre las variables como los árboles de decisión o las redes neuronales [68].

Un factor a tener también en cuenta a la hora de seleccionar el modelo es la información disponible en la planta [69–71]. Existen estudios de los que se pueden extraer recomendaciones sobre qué información es a aconsejable medir [72–74]. La norma IEC 61724 [75] dedicada a los aspectos de la monitorización en las plantas fotovoltaicas también hace referencia a que información es aconsejable medir en una planta. El manual de buenas prácticas para la monitorización y recogida de información de una planta fotovoltaica proporciona recomendaciones sobre qué se debe medir en las plantas. Este manual es el resultado de una tarea de los grupos de trabajo de la IEA “International Energy Agency”, del programa “Photovoltaic Power Systems Programme” [19]. La disponibilidad de las diferentes medidas del sistema de monitorización condiciona el tipo de modelo a aplicar. Si se dispone de medidas eléctricas los modelos serán diferentes a si se dispone de medidas meteorológicas. La frecuencia de medida también lleva a desarrollar modelos diferentes y resultados con incertidumbres diferentes [76–78]. Además, en los modelos que se entrenan a partir de un histórico de información, la longitud del periodo de entrenamiento y la calidad de los datos pueden influir notablemente en su precisión [79].

Los datos meteorológicos que recomiendan, según los estudios mencionados, para realizar la caracterización de la energía producida son: la irradiancia, la temperatura ambiente y la velocidad de viento. A veces, si se trata de una estructura fija, la temperatura de módulo puede suplir la velocidad del viento e incluso la temperatura ambiente. La información eléctrica incluye las potencias o en su defecto las corrientes del inversor. El punto donde se registra la información

influye en las pérdidas que se pueden identificar. Así, cuanto más cerca del punto de la generación se adquiera la información, más fácil será la identificación de las pérdidas.

Los modelos ML tienen algunas limitaciones, destacando entre ellos su falta de generalidad, dado que son modelos desarrollados para cada planta. Un modelo entrenado en una planta no puede aplicarse de manera inmediata en otra, ni aunque las características de los equipos sean similares [80]. Otra limitación es la dependencia de un histórico de información necesario para disponer de un periodo de entrenamiento con buena calidad de datos. El periodo de entrenamiento debe ser representativo en cuanto a casuística de información y su longitud debe ser suficiente para la caracterización de los parámetros [81].

Pese a sus limitaciones, los resultados que proporcionan, la ventaja de depender exclusivamente de la información disponible en la propia planta, y la capacidad para interpretar la especificidad de cada equipo, convierten a los modelos de ML en la mejor opción para estimar la producción fotovoltaica. En este capítulo se va a realizar una revisión bibliográfica centrada en los modelos de aprendizaje automático ML aplicados a plantas fotovoltaicas para estimar la producción fotovoltaica y la irradiación solar. En esta revisión se incluyen los métodos para la optimización del aprendizaje de estos modelos y las variantes probabilísticas de los mismos que permiten definir una incertidumbre o probabilidad del resultado.

Modelos para determinar la producción fotovoltaica

Conocer la producción de la planta, tal y como se ha mencionado, va a permitir conocer su estado, ajustar los balances de resultados económicos de los proyectos, mejorar la integración en red y detectar fallos o pérdidas de producción [82–88]. El cálculo de la producción con un modelo puede ser, además, la referencia para estimar las pérdidas en las plantas debidas a fallos, sombras o ineficiencias [59,89–92].

Existen referencias bibliográficas que demuestran que los modelos de ML se han aplicado para conocer el comportamiento teórico-ideal de la producción de la planta, o de una parte de ella, al haber sido aplicados solo a alguno de sus componentes. Las variables dependientes de entrada de estos modelos cambian de un estudio a otro [92–94]. La energía producida o la potencia generada está influida por variables como la radiación, la temperatura, etc. [95].

Los modelos que trabajan con un histórico de variables climatológicas medidas en la propia planta son especialmente interesantes para predecir la producción de una planta fotovoltaica porque no necesitan información adicional. Entre las variables meteorológicas que se pueden medir en las plantas [69], las más importantes son la irradiancia y la temperatura ambiente por ser las más influyentes en la generación [96]. Otra variable también importante es la temperatura de los módulos, que tiene en cuenta el comportamiento de los equipos de generación [97,98].

La experiencia en estos modelos de producción de plantas fotovoltaicas está referida a los proyectos que se instalaron durante la primera década del siglo XXI, que son de un tamaño reducido en comparación con las plantas que se han instalado en estos últimos años, hecho que influye en la decisión de la elección del modelo. Las plantas estudiadas van desde unos cuantos paneles [99] hasta unas decenas de kilovatios [98,100]. El estudio de estas pequeñas plantas, ha

dado lugar a una extensa base de propuestas de modelos de ML con aplicación en diagnostico o en conexión a red [63,64,101–103].

Uno de los modelos más sencillos de ML y que se usa para estimar la producción es la regresión lineal multivariable, "Multiple Linear regresión" (MLR), que puede aplicarse en cualquier punto de la planta en el que se disponga de un histórico de medidas, ya sea en la parte de DC o en la parte AC. Este modelo funciona muy bien cuando hay medida de irradiancia muy cerca de la generación, condición de pequeñas plantas, ya que en esos casos hay una alta correlación entre la irradiancia y la producción. Si además se tiene en cuenta en el modelo la temperatura ambiente o de módulo los resultados mejoran disminuyendo el error del modelo [104]. Sin embargo a medida que las plantas aumentan de tamaño o aparecen más factores influyentes en la producción, se necesitan modelos más evolucionados, y con dependencias más complejas [105].

Modelos que permiten relaciones más complejas entre las variables son los modelos no lineales, entre los cuales se encuentran los algoritmos basados en árboles de decisión. Los árboles de decisión se han utilizado en otros campos, pero existen pocos estudios en plantas fotovoltaicas. Los algoritmos basados en esta metodología se han aplicado a la identificación de anomalías, trabajando en este caso el modelo como modelo de clasificación. En otras ocasiones los árboles han servido para analizar la influencia de otras variables sobre la producción [106–108] y en [84] se han utilizado como modelo regresivo estimando la producción con un buen resultado. También se han comparado los árboles de decisión (DT) y "Random Forest" (RF) con regresiones lineales como "Lasso" y "Ridge" [109] dando mejores resultados los modelos basados en árboles de decisión.

Los modelos de ML más utilizados en los últimos años son las redes neuronales ANN [110–112], denominados también "Deep Learning" (DL). El uso de las redes neuronales está muy extendido y se pueden encontrar estudios donde se combinan éstas con otros modelos de ML [71,113,114] o entre ellas formando un conjunto de redes [100]. También existen estudios comparativos entre redes neuronales y otros modelos como una regresión lineal multivariable (MLR), "Support Vector Machine" (SVM) y "Dynamic Kalman Filters" (KF) [99,115]. En otros dos estudios más se han comparado las redes neuronales con los modelos basados en arboles de decisión. Así, en el primero de ellos se han comparado las ANNs con "Support Vector Regression" (SVR) y con "Regression Trees" (RT) [116], dando mejor resultado las redes neuronales. En el segundo estudio se han comparado en 5 plantas diferentes, los modelos: "Lasso Regression", "K-Nearest Neighbors", "Gradient Boosting" (GB), "Regression Trees" y ANNs, siendo los mejores GB y ANN [117].

Todos los estudios presentados hasta el momento corresponden a plantas de pequeño tamaño, de unos cuantos kilovatios, siendo uno de los estudios de mayor potencia el que corresponde a una planta conectada a red de 2640 kWp [49]. En cuanto a plantas de mayor tamaño, en las que se estime la producción se han encontrado en la bibliografía unos pocos estudios con potencias del entorno de los 50 MW, superiores a las anteriores pero todavía lejos de las potencias que se instalan actualmente, como la planta de 500 MW en Golmud (China) [72].

Así, en un primer estudio, se trabaja con un modelo "Random Forest", para estimar la producción de la planta de 50 MW a partir de datos de temperatura e irradiancia de la propia planta. Este

estudio trabaja con datos diarios, un escenario diferente en cuanto a precisión y error de los modelos, y no comparable a la precisión y error que se obtendría al trabajar con intervalos de 10 minutos, que son necesarios si el objeto es que la producción estimada sirva para identificar posibles anomalías [84]. Un segundo estudio, también de una planta de 50 MW [118], aplica redes neuronales. En este estudio se ha realizado una predicción de la producción en un escenario máximo de una hora vista con tres modelos diferentes. La diferencia entre los tres modelos se encuentra en las variables de entrada al algoritmo. En el primero de ellos, el modelo es monovariable trabajando solo con la propia producción. El segundo modelo incluye los ángulos de elevación acimutal del sol. Por último, en el tercer modelo se incluye además información de satélite. En ninguno de ellos se relaciona la producción con variables medidas en la planta. Como último estudio de plantas de gran tamaño, existe otro centrado también en una planta de 50 MW [119]. En este estudio se comparan tres tipos de redes neuronales, "Multi Layer Feed Forward Neural Network" (MLFFNN), "Recurrent Neural Network" (RNN), y "Nonlinear Auto Regressive eXogenous Neural Network" (NARXNN). Los modelos se comparan con las métricas RMSE y MSE, dando como resultado que la red más adecuada para esta aplicación la MLFFNN. El periodo de datos utilizado para entrenar la red es de 45 días y la verificación de la misma se ha realizado con 15 días, periodo muy pequeño y que no contempla el año solar completo, por lo que su aplicación generalista no parece viable.

La revisión realizada en esta sección muestra que los modelos no paramétricos y no lineales de ML son adecuados para determinar la producción de una planta fotovoltaica, puesto que son adaptables a diferentes tipos de información y a las características de la planta. Pese a que estos modelos han sido ampliamente probados en pequeñas plantas fotovoltaicas, es necesario verificar la validez de los mismos en plantas grandes donde la correlación entre la producción y los datos de irradiancia puede ser baja. Los estudios de plantas de gran tamaño son escasos, y los datos utilizados en ellos no se ajustan a los requerimientos del problema que se plantea en el desarrollo de esta tesis. Por todo ello, se demuestra la necesidad y oportunidad de los estudios desarrollados con el fin de adaptar los modelos de ML al cálculo de la producción de las plantas de gran tamaño, con la finalidad de que esta pueda ser aplicada a temas de diagnóstico y cálculo de pérdidas [72].

A la vista de la revisión bibliográfica realizada, los modelos que se seleccionan para analizar son modelos de ML. Por un lado, se estudia el modelo más sencillo MLR que servirá como referencia para el resto de los modelos. Por otro lado, se trabaja también el modelo más complejo y utilizado, las ANN, pero además se usarán modelos basados en árboles de decisión que, tal y como se ha observado en varios trabajos, pueden ser muy adecuados para estimar la producción de una planta FV. Entre estos modelos de árboles de decisión se han escogido los modelos RF y GB. Cabe destacar que no se han encontrado estudios en la literatura científica en los que se haya hecho una comparación de los modelos seleccionados.

Modelos para determinar la irradiancia solar

La irradiancia solar que recibe una planta es otra de las variables clave en los proyectos fotovoltaicos. Primero porque es la variable más influyente en el cálculo de la producción de las plantas [120], y segundo porque es necesaria tanto en la fase de desarrollo del proyecto, como en la de operación. En la fase de desarrollo permite identificar ubicaciones óptimas de la plantas,

pronosticar la producción de energía [77,121] y garantizar la inversión [47]. En la fase de operación, es relevante en el cálculo del "Performance Ratio" (PR), ayuda a cumplir los requisitos de la red [122] y sirve como apoyo en el diagnóstico precoz de fallos de la planta [95,123].

Se considera necesario puntualizar dos términos que a veces se encuentran intercambiados en algunos textos. La irradiancia solar (W/m2) es la potencia por unidad de área recibida del Sol, frente a la radiancia, que es la potencia que emite el sol (W/m2) [124]. Estos dos conceptos tienen sus homólogos en energía si se tiene en cuenta un período de tiempo determinado. Así, la irradiación es la energía recibida del sol, que se mide en (J/m2) y, análogamente a la energía que sale del sol se le llama radiación (J/m2). A la irradiancia solar se le suele conocer como Irradiancia Global, por ser el total de la radiancia que llega a un lugar en un instante, y según como se mida puede ser Normal, Horizontal o Inclinada.

La Irradiancia Global tiene dos componentes [125], la Irradiancia Directa, que es la radiancia que llega a un determinado lugar procedente del disco solar y que, a su vez, puede ser también Normal, Horizontal o Inclinada, y la Irradiancia Difusa, que es la radiancia procedente de toda la bóveda celeste excepto la procedente del disco solar. Los valores más comúnmente medidos en las plantas fotovoltaicas son la Irradiancia Global Normal, la Irradiancia Global Inclinada y la Irradiancia Global Horizontal.

El sensor para medir la irradiancia es el piranómetro. Los piranómetros son instrumentos delicados, que deben alinearse correctamente [46] y limpiarse regularmente, ya que están expuestos al viento, la lluvia y el polvo, que ensucian constantemente la lente. Para que las mediciones sean de calidad, deben estar bien ubicados y calibrados [126,127], y además tener un mantenimiento adecuado [128,129]. Las células calibradas [130], sin tener la precisión de un piranómetro, tienen un uso extendido en plantas fotovoltaicas. Las medidas registradas por ambos sensores, piranómetro o célula, tienen una alta probabilidad de fallo [131], lo que implica la falta de datos o que estos sean erróneos [132]. Por eso es necesario verificar la validez de las mediciones de la irradiancia.

En el ámbito de la verificación de la irradiancia existen pocos estudios publicados que describan métodos para la verificación de esta medida. En los años 90, se llevó a cabo el proyecto internacional "Baseline Surface Research Network" (BSRN), perteneciente al programa "World Climate Research Programme" de la Organización meteorológica mundial "World Meteorological Organisation". Este proyecto instaló estaciones meteorológicas por todo el mundo, para validar la información que llegaba vía satélite. El fin era crear modelos climáticos y estudiar las variaciones temporales de la irradiancia, sobre la superficie de la tierra [133]. Dentro del desarrollo del proyecto, era necesario validar la calidad de las mediciones. Para ello se dispusieron varias medidas de irradiancia directa y difusa sobre el mismo punto, lo que permitió establecer relaciones entre las medidas verificando así su validez [134].

Los estudios más recientes, que identifican errores en la medida de irradiancia son una mejora de las pruebas realizadas en BSRN, incluyendo nuevas variables, como la irradiancia extraterrestre en la parte superior de la atmósfera, el índice de claridad [132], o la hora del día de la medida de irradiancia a estudiar [135]. Los resultados de estos estudios evalúan la precisión de la medición anual de la radiación y están orientados al análisis de las variaciones en el largo plazo de la

radiación sobre la tierra.

Una estación meteorológica de una planta solar fotovoltaica comercial dispone de medida de irradiancia, presión, temperatura, humedad, velocidad y dirección del viento. Sin embargo, no suele disponer de pirheliómetros para medir la irradiancia directa, ni de sensores redundantes. Las relaciones y límites propuestos en los estudios descritos previamente para analizar la calidad de las medidas de irradiancia, no se pueden aplicar a la información de una planta fotovoltaica convencional, por lo que es necesario un modelo que permita estimar una referencia y así poder verificar el valor medido. La revisión de esta sección se focaliza por tanto en la búsqueda de modelos que estimen la irradiancia a partir de las medidas de una estación meteorológica de una planta fotovoltaica.

Existen numerosos modelos que calculan la irradiancia [92,122,136–140], diferenciándose según su aplicación y la metodología seguida. Las ventanas temporales de trabajo varían desde los segundos, con aplicaciones relacionadas con la integración en red, a los años con aplicaciones en análisis climatológicos [141]. Lo mismo sucede con la cobertura espacial del modelo. Se encuentran escenarios de trabajo que corresponden a una localización concreta de una planta, y escenarios que comprenden toda una región con aplicaciones meso-escalares [142].

Las diferentes metodologías se pueden clasificar en cuatro grupos. El primer grupo son los modelos teóricos, modelos paramétricos basados en las coordenadas geográficas del emplazamiento, la posición del sol y la irradiancia teórica que llega a la tierra [143–145]. Estos modelos pueden incluir correcciones que mejoren los resultados, como la turbiedad del aire, la cantidad de aerosoles en el ambiente [146,147] y adaptaciones trigonométricas a la posición de los módulos [148]. Pero no son modelos de ML por lo que no serán considerados en la revisión.

El segundo grupo son los modelos cuya metodología se basa en el tratamiento de imágenes. Estos modelos están orientados a la detección y predicción de la nubosidad, y algunos de ellos se basan en ML. Trabajan con la información de las imágenes del cielo, bien captadas por satélites, o por cámaras [137,143,149,150]. La caracterización de la nubosidad por sí misma no determina el valor de la irradiancia, pero mejora notablemente el resultado. Por eso estos modelos suelen usarse como complemento de otros [151]. Estos modelos tienen una ventana de trabajo temporal muy estrecha, acorde con el movimiento de las nubes por lo que no son aplicables en escenarios de trabajo de medias de diez minutos que suelen ser las medidas registradas en el “Supervisory Control And Data Acquisition” SCADA de las plantas.

El tercer grupo de modelos incluye la física de la atmosfera y se denominan “Numerical Weather Prediction” (NWP) [137,141,149,152–154]. No utilizan datos locales sino datos globales y sirven generalmente para elaborar modelos de irradiancia meso-escalares, por su capacidad para analizar grandes áreas [142], por lo que no pueden ser aplicados para la verificación de las medidas de irradiancia a partir de los datos medidos en la propia planta.

El cuarto grupo son los modelos en los que se centra la revisión, los modelos de aprendizaje automático, regresivos, basados en el análisis de los datos históricos medidos en las estaciones meteorológicas de las propias plantas y cuyo resultado depende de las observaciones previas [155],[94]. Estos modelos tienen la ventaja de proporcionar mejores resultados que los modelos

teóricos y trabajan con datos adquiridos en la planta sin necesidad de sensores, cámaras o información adicional.

Los modelos de aprendizaje automático regresivos se utilizan principalmente en predicción, "forecasting", y trabajan solo con el histórico de la irradiancia, siendo por tanto monovariables. Entre estos destacan los modelos de series temporales lineales como "Auto-Regressive" (AR), "Moving-Average" (MA), "Auto-Regressive Moving-Average" (ARMA), "Auto-Regressive Integrated Moving-Average" (ARIMA), y "Seasonal Auto-Regressive Integrated Moving-Average" (SARIMA) [137,156,157]. Los modelos de series temporales monovariable no lineales han sido menos estudiados [158].

Los modelos de aprendizaje automático regresivos multivariable incluyen variables adicionales mejorando así el resultado del modelo. Los más sencillos son los lineales, como el "Multiple Linear Regression" (MLR), que relaciona la irradiación con otras variables como la temperatura, presión, humedad y velocidad del viento [159], el "Quantile Regression" (QR) que es una variación del anterior trabajando con percentiles [152] y el "Auto-Regressive Integrated Moving-Average with eXogenous variables" (ARIMAX), que es una modificación del modelo monovariable ARIMA que permite trabajar con variables externas o exógenas [160]. Algoritmos de aprendizaje automático regresivos multivariable lineales más complejos con aplicación también en el cálculo de la irradiancia son el "Support Vector Machine" (SVM) [161–163], y el "Support Vector Regressor" (SVR) [164].

Los modelos de aprendizaje automático regresivos multivariable no lineales, son los que presentan mejores resultados. Es por ello por lo que han dado lugar a la mayoría de los modelos desarrollados. Así se pueden encontrar modelos que son adaptación de los mencionados anteriormente como es el caso de incluir en un algoritmo ARIMA, "Wavelets" y variables exógenas convirtiéndose en un "Wavelet Auto-Regressive Integrated Moving-Average with eXogenous variables" (WARIMAX) [158], hasta algoritmos más complejos como las redes neuronales [165–173].

Estos últimos, las redes neuronales, son los algoritmos más estudiados porque pueden relacionar numerosas variables como la temperatura, la humedad, la presión y la velocidad del viento [174]. Además, se pueden combinar con otros modelos, por ejemplo, con wavelets dando lugar al modelo "Wavelets and Artificial Neural Networks" (WANN) [158], o con un modelo NWP [165,175],o con un modelo de cálculo de nubosidad como "Auto-Regressive with eXogenous variables" ARX [176]. Entre los algoritmos ANN resulta interesante para el cálculo de la irradiancia el "Multilayer Perceptron Neural Networks" (MPNN) [177].

Otras alternativas utilizadas para el cálculo de irradiancia, menos estudiadas pero de resultados interesantes, son el algoritmo "Extreme Learning Machines" (ELM) [178], el "K-Nearest Neighbors" (KNN) [179], algoritmos de lógica difusa como el "Adaptive Network-based Fuzzy Inference Systems" (ANFIS) [180,181] y algoritmos basados en árboles de decisión "Quantile Regression Forests" (QRF) [49]. Estos últimos presentan resultados con un rango de error similar a las ANN, siendo además más sencilla su aplicación en los cálculos.

De la revisión de los algoritmos basados en técnicas de ML para la predicción de irradiancia a

partir de señales meteorológicas se concluye que los modelos regresivos de aprendizaje automático pueden modelizar la irradiancia que recibe una planta fotovoltaica. También se observa que de estos algoritmos los modelos más estudiados son las redes neuronales. Por ello se propone trabajar con las redes neuronales, que permitirá verificar los resultados con los estudios de la bibliografía, con un MLR como referencia, y para completar el análisis comparativo trabajar también con árboles de decisión en dos de sus modalidades GB y RF. De esta manera los algoritmos son coincidentes a los seleccionados para modelizar la producción de la planta lo que va a permitir conocer el comportamiento de los mismos en dos aplicaciones relacionadas con las plantas fotovoltaicas.

Técnicas de búsqueda de hiperparámetros

Los hiperparámetros determinan la estructura y la manera en que aprende el algoritmo, como, por ejemplo, el número de árboles y su tamaño en modelos basados en arboles de decisión o el número de capas ocultas y de neuronas en cada capa de los modelos de redes neuronales. Por ello es importante determinarlos con precisión para obtener un rendimiento óptimo de los modelos obtenidos. Tradicionalmente, los hiperparámetros se han definido mediante la técnica de prueba y error [115], pero actualmente se están utilizando técnicas más avanzadas de optimización de hiperparámetros (HPO) [182]. Los algoritmos de optimización de hiperparámetros son un conjunto de algoritmos de búsqueda sistemática o de técnicas de optimización [183].

La técnica más utilizada en HPO es la búsqueda en red “Grid Search”, que consiste en generar un espacio de resultados con todas las posibles combinaciones de valores que se desean probar. Esta técnica se ha utilizado con éxito, por ejemplo, para encontrar el valor óptimo del número de clusters en un k-means [184] o para encontrar el número óptimo de capas ocultas y neuronas en redes neuronales [115]. Sin embargo, esta técnica puede volverse computacionalmente costosa cuando el número de hiperparámetros es alto o cuando los intervalos de los hiperparámetros se discretizan en muchos valores.

Una alternativa es la búsqueda aleatoria, "Random Search", donde las combinaciones de hiperparámetros se generan aleatoriamente y es posible controlar el número final de valores. Esta técnica puede alcanzar buenos resultados con menos valores posibles, ya que no requiere probar todas las combinaciones del espacio de búsqueda [185]. Además, esta técnica puede obtener buenos resultados cuando los hiperparámetros tienen diferente influencia sobre el resultado de los modelos. Sin embargo, tiene el inconveniente de no proporcionar la seguridad de haber encontrado realmente los hiperparámetros óptimos.

La técnica de búsqueda que mejor soluciona el problema de combinaciones elevadas sin aplicar optimizaciones complejas es el método de hiper-banda, "Hyper-band",[186] [186] porque trabaja con conjuntos de combinaciones en lugar de combinaciones individuales, que además va discriminado mediante separaciones sucesivas. Este método es mucho más rápido que los anteriores [187].

Se han utilizado otros métodos más complejos, como el análisis de sensibilidad con parada automática temprana, para determinar el número de neuronas de las capas ocultas de un modelo de redes neuronales [188], o el uso de "Ant Colony Optimisation" (ACO) para optimizar los

hiperparámetros de un "Support Vector Machine" (SVM) [189]. También se han aplicado algoritmos de optimización, como por ejemplo la optimización Bayesiana. Ésta consiste en un algoritmo iterativo en el que se unen un modelo probabilístico y una función de adquisición que decide el valor de la variable en el que se aplica el modelo. En cada iteración, el modelo se ajusta a todas las observaciones de la función objetivo realizadas hasta el momento y la función de adquisición, que utiliza la distribución predictiva del modelo probabilístico, determina el siguiente valor candidato a ser ajustado con el modelo. [190–192].

La determinación de la metodología de optimización más adecuada para cada modelo de ML requiere de estudios comparativos entre las diferentes técnicas, y no existen muchos de este tipo. El estudio más completo de comparación no pertenece al ámbito fotovoltaico, y en él se analizan cuatro técnicas diferentes: “Random Search”, “Grid Search”, “Tree Structured Parzen Estimators” (TPE), y “Covariance Matrix Adaptation-Evolutionary Estrategy” (CMA-ES). Los hiperparámetros a optimizar pertenecen al modelo “Long Short Term Memory neural network” (LSTM). El resultado de este estudio concluye que TPE es la técnica más adecuada para optimizar el modelo por ser la técnica más rápida y con menos demanda computacional [183]. Sin embargo, existen otros dos estudios más, que también optimizan los hiperparámetros de LSTM. El primero de ellos presenta un algoritmo propio “Adaptative Hyper Paramter Adjustment” (AHPA) como mejor solución [193] y el segundo propone como solución, una optimización bayesiana para optimizar los hiperparámetros del mismo modelo, el LSTM [194]. Dado que los resultados no son los mismos en todos los estudios, no se puede llegar a concluir qué metodología de optimización de hiperparámetros es mejor en el caso de un LSTM. De la misma manera sucede con el SVM, sus hiperparámetros han sido definidos con diferentes técnicas. En un primer estudio se aplica una búsqueda en red “Grid Search” [195], un segundo estudio aplica el algoritmo de optimización “Grass Hopper Optimization Algorithm” (GOA) [196], y un tercer estudio un método de prueba y error [115]. El hecho de existir varios estudios de diferentes modelos que optimizan los hiperparámetros de diferentes formas, revela que no hay una tendencia clara ni una solución única en la metodología de optimización de hiperparámetros para un modelo concreto.

La búsqueda de hiperparámetros que optimicen la configuración del algoritmo del modelo de ML, apenas se ha considerado en la fotovoltaica. Existe un estudio que analiza los hiperparámetros en un modelo SVM [195], y otro estudio que también ajusta los hiperparámetros de SVM y de las ANN [115]. Sin embargo, en las grandes plantas que se están instalando, la sensibilidad del resultado es más crítica que en plantas pequeñas existentes, porque el grado de correlación entre las variables del sistema es menor y el impacto del error en los modelos mayor. Por estos motivos es importante garantizar que se está trabajando con los hiperparámetros convenientemente ajustados.

En resumen, es necesario aplicar técnicas sistemáticas de búsqueda u optimización de hiperparámetros en los modelos de aprendizaje automático. Aunque no hay una metodología clara ni una identificación de los hiperparámetros en algunos de los modelos, los análisis de los estudios revelan que encontrar la configuración adecuada de hiperparámetros mejora el rendimiento de los modelos.

En esta tesis, se propone utilizar la técnica de búsqueda en red ("Grid Search") siempre que sea

posible, ya que permite explorar todas las opciones. Sin embargo, cuando el número de modelos resultantes sea muy alto y la demanda computacional sea un problema, se utilizarán otras técnicas de búsqueda como la aleatoria ("Random Search") o el método de hiper-banda ("Hyper-band").

Variante probabilística de los modelos

Los resultados de los modelos previamente revisados son la aplicación de un modelo a una muestra de la población que es lo que se denomina una estimación puntual del modelo, pero se desconoce la representatividad de este resultado frente al resultado de trabajar con la población completa.

En el ámbito de la estadística, cuando la población es muy grande o infinita, se trabaja con muestras y se calculan estimaciones puntuales [197]. Diferentes muestras dan lugar a diferentes estimaciones puntuales. Para conocer el resultado de un modelo de manera global sin influencia de la muestra de entrenamiento es necesario recurrir a la inferencia estadística [198].

La inferencia estadística es el conjunto de métodos que permiten conocer, a través de una muestra estadística, el comportamiento de una determinada población con un riesgo de error medible en términos de probabilidad, permitiendo extraer conclusiones sobre la población así como, el grado de fiabilidad de los resultados del estudio [199].

Uno de los métodos de trabajo en inferencia, consiste en definir un intervalo de confianza del resultado, basado en el abanico de resultados proveniente de diferentes muestras de la población [200]. Un intervalo de confianza indica el margen de variación que puede tener el resultado con una probabilidad determinada. A dicha probabilidad se le denomina nivel de confianza del intervalo, y la precisión del modelo viene definida por la amplitud del intervalo. El nivel de confianza y la longitud del intervalo varían conjuntamente para un mismo modelo, de forma que un intervalo más amplio tendrá más probabilidad, mayor nivel de confianza, mientras que un intervalo de menor amplitud, tendrá menor nivel de confianza [200]. Si se comparan diversos modelos, el modelo más preciso será el que para un mayor nivel de confianza tenga una amplitud de intervalo menor.

La bibliografía describe diversas maneras de calcular un intervalo de confianza. La más sencilla se basa en la distribución Normal, ya que existe una relación entre la desviación estándar y el nivel de confianza [201], que permite definir el intervalo con unos límites de $\pm k\sigma$, donde k es cualquier número entero positivo mayor que la unidad y σ es la desviación estándar de la media de la población. Para garantizar un nivel de confianza del orden de un 95% se tendría que trabajar con un valor de k de 2.5. Para saber si una distribución es Normal se realiza el diagnostico a los datos, mediante la gráfica de cuantiles o un test de normalidad [197,202].

En el caso de no disponer de una distribución Normal, existen diferentes alternativas. El método más inmediato y sencillo, para determinar intervalos de confianza, es el teorema de Chebyshev [197], que puede ser aplicado a cualquier distribución y permite definir el intervalo también con unos límites de $\pm k\sigma$. En este caso para garantizar un nivel de confianza del orden de un 95% se tendría que trabajar con un valor de k igual a 4.5. La desventaja de este método es que el nivel de confianza es aproximado y el valor de k tiende a ser más elevado de lo necesario, dando lugar a intervalos con amplitud sobrestimada.

Una segunda opción cuando la distribución no se ajusta a una Normal, consiste en la aplicación de técnicas de "Boostraping". El "Boostraping" es un método de re-muestreo, que consiste en generar un gran número de subconjuntos a partir de los datos originales del estudio. Esta estrategia implica la selección repetida de elementos, obteniendo numerosas muestras más pequeñas formadas por partes de la muestra inicial. Al aplicar el modelo a todas las nuevas muestras se genera una distribución de resultados, uno para cada nueva muestra. Esto es posible porque la inferencia o comportamiento de una población puede ser modelada mediante un nuevo muestreo de los datos de la misma muestra [203]. Cuando el conjunto de resultados generados con el re-muestreo es lo suficientemente grande, se puede aplicar el teorema del límite central, que dice que si una muestra es lo bastante grande, sea cual sea la distribución de la media muestral, la muestra seguirá aproximadamente una distribución Normal [204]. Aplicando el teorema del límite central, el conjunto de los resultados seguirá una distribución Normal con μ como valor medio y σ como desviación estándar, y los límites del intervalo quedarán definidos en función del nivel de confianza y σ. La desventaja de este procedimiento es que hace falta un número muy grande de muestras en el re-muestreo para que la precisión sea buena. Está técnica ha sido empleada en el ámbito de la fotovoltaica [205].

Una tercera forma es definir un intervalo de confianza basado en el uso de percentiles [200]. El percentil es una medida estadística utilizada para clasificar y comparar datos en un conjunto de observaciones. Se trata de un valor que divide la distribución de datos en 100 partes iguales. Cada percentil representa la posición relativa de un valor dentro de una distribución de datos. El percentil de orden τ de una distribución, es el valor de la variable que divide los datos en dos partes, una con τ % observaciones por debajo y otra con (1- τ) % por encima. Los intervalos de confianza definidos mediante percentiles tienen un límite inferior de $\tau/2$ y un límite superior de $1-(\tau/2)$.

La ventaja de definir intervalos a partir de un modelo de percentiles respecto a realizar un "Boostraping", es que se entrena una única vez al modelo lo que supone un importante ahorro computacional. Además, los intervalos calculados con percentiles son asimétricos lo que permite ajustar mejor los resultados. La limitación de usar percentiles es que se necesita un conjunto representativo de observaciones en el resultado, lo que es crítico para nuestra aplicación dado que a radiaciones altas hay menos observaciones.

Otros métodos que permiten calcular intervalos de confianza son el método "Parametric Prediction Interval" (PPI), basado en la función de distribución de Student [206] y el método "Interval Quadratic Programming" [207]. Estos métodos no representan una mejora sobre los anteriores, pero se mencionan porque han sido utilizados para generar intervalos de confianza de la producción fotovoltaica.

En los modelos que trabajan con redes neuronales, su incertidumbre se calcula generando multitud de resultados que caracterizan la aleatoriedad del dato y la variabilidad del propio modelo [208], y sobre los cuales se puede aplicar el teorema del límite central. Para facilitar el cálculo de los resultados, sin llevar a cabo la repetición sucesiva de los modelos, se modifican la naturaleza de las capas del modelo [209]. Esta metodología es mucho más eficiente desde el punto de vista computacional que el "Boostraping", que necesitaría repetir todos los modelos con las diferentes combinaciones.

En cuanto a la predicción climatológica, también existen referencias de estudios en los que se calculan intervalos. Por ejemplo un estudio que trabaja con un modelo numérico de predicción climatológica "Numerical Weather Prediction" (NWP) considera una distribución de "Kernel Gausiano" para definir el intervalo del resultado de la predicción [210]. O un estudio en el que se define un intervalo con un 95% en base a la distribución de t-student para la irradiancia con la que posteriormente se calcula la producción [206].

El uso de intervalos de confianza como resultado de un modelo está muy extendido en otras áreas de la ciencia. Sin embargo, en el ámbito de las plantas fotovoltaicas no está tan generalizado su utilización. Pese a que no existe una extensa bibliografía de estudios en los que se hayan estimado intervalos de confianza en modelos aplicados a las plantas fotovoltaicas, es necesario considerar el intervalo de confianza como parte de la definición de modelos en las aplicaciones de las plantas fotovoltaicas. De las técnicas para el cálculo de intervalos de confianza revisadas siempre que no se cumplan criterios de normalidad se preferirá usar percentiles porque, además de la agilidad de cálculo que presentan, son capaces de mostrar la heterocedasticidad, es decir, permite que la varianza de los errores no sea constante en todas las observaciones realizadas. Esto es fundamental porque los modelos no se comportan igual a bajas producciones, a medias y a altas [211,212]. Para el caso concreto de las redes neuronales el cálculo se llevará a cabo con "Tensor Flow Probability" (TFP) [208,209,213].

1.4. Estructura de la tesis

Resumiendo lo analizado hasta el momento, la energía fotovoltaica va a estar muy presente en los próximos años, con gran potencial y va a ser parte de la base de la generación de energía mundial. La tesis se enfoca en el desarrollo de modelos de aprendizaje automático para determinar la producción y la irradiancia en grandes plantas fotovoltaicas. La producción de energía en estas plantas está influenciada por diversas variables, como la irradiancia, las irregularidades del terreno, las sombras y la nubosidad, lo que requiere modelos más complejos para una estimación precisa, como los modelos de aprendizaje automático.

El estudio tiene como objetivos crear un modelo que, basándose en datos medidos directamente de la planta, estime su producción o la de una parte de la misma. Y crear un segundo modelo que estime la irradiancia a partir de la irradiancia teórica y de las variables meteorológicas registradas en la planta.

En la revisión bibliográfica se han evaluado los modelos y técnicas existentes aplicados a la producción e irradiancia, así como las técnicas para optimizar hiperparámetros y calcular intervalos. En los siguientes capítulos se describirá la base matemática de los modelos seleccionados y su aplicación a la estimación de la producción y la irradiancia.

El capítulo dos se centra en describir teóricamente los modelos, mientras que en el capítulo tres se presentan los resultados de la aplicación de los modelos a tres plantas fotovoltaicas para determinar la producción de un inversor en cada una de ellas. También se definen los hiperparámetros de cada modelo y planta, se calcula la estimación puntual del modelo utilizando

una muestra de prueba y se determina su intervalo de confianza. El capítulo cuatro se dedica a la aplicación de los modelos a la medida de irradiancia, utilizando una estación meteorológica que proporciona medidas de irradiancia y otras variables climatológicas. Al igual que en la estimación de producción, se optimizan los hiperparámetros y se calculan los intervalos. Finalmente, el último capítulo extrae las conclusiones obtenidas a partir de los resultados obtenidos en el estudio.

2. Marco teórico. Modelos y su optimización

2.1. Introducción

En este capítulo se describen los algoritmos de aprendizaje automático, "Random Forest" (RF), "Gradient Boosting" (GB) y redes neuronales (ANN), seleccionados en el capítulo anterior, sección 1.3. Los dos primeros modelos están basados en árboles de decisión, por lo que se explica el funcionamiento y las peculiaridades específicas de estos dos modelos. En cuanto a las redes neuronales, la tipología es muy amplia, por lo que la descripción se centrará en el funcionamiento de las "Multilayer Perceptron Neural Networks" (MPNN), que son las utilizadas en la tesis, por su sencillez y eficacia para resolver problemas de regresión. Además de estos modelos de "Machine Learning" (ML), se utilizará como modelo de referencia la regresión lineal multivariable (MLR).

También forman parte del marco teórico de la tesis las metodologías de búsqueda de los hiperparámetros de los modelos de ML mencionados que, es un paso fundamental para conseguir la mejor precisión de ajuste de cada uno de los modelos. En este capítulo se describen las técnicas de búsqueda de los hiperparámetros (Sección 1.3): "Grid Search", "Random Search" e "Hyper-band".

Para una toma de decisión adecuada, la estimación del método debe venir acompañada de la estimación de su incertidumbre. Completando los modelos de ML seleccionados, se describe el funcionamiento de las variantes de los mismos que permiten calcular los intervalos de confianza de la predicción del modelo.

Para terminar el marco teórico de la tesis se explican las métricas que se usan para poder comparar el comportamiento de los diferentes modelos.

2.2. Regresión Lineal Multivariable

La Regresión Lineal Multivariable (MLR) [214], es un modelo lineal que trabaja de forma similar a una regresión lineal simple del tipo y=a·x + b, pero utilizando más de una variable independiente, tal y como se describe en la ecuación (1):

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_K \cdot x_K + \varepsilon \quad \textbf{(1)}$$

Siendo "y" la respuesta, "x_i" las variables independientes, β_0 el termino independiente, β_1, β_2, ..., β_k los coeficientes de la regresión y ε el error. El error, generalmente se supone que tiene una esperanza nula y una varianza de valor σ^2, (2), presentando una distribución Normal [197].

$$E(\varepsilon)=0;\ Var(\varepsilon)=\sigma^2 \quad \textbf{(2)}$$

En el entrenamiento del modelo se calculan los coeficientes β_i, mediante el método de mínimos cuadrados, minimizando la suma de los cuadrados de los residuos "Residual Sum of Squares" (RSS) [215]. Los valores de los coeficientes β_i se obtienen igualando a cero las derivadas parciales de la función RSS (3) respecto a dichos coeficientes.

$$RSS=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \rightarrow \left.\frac{\delta RSS(\beta_i)}{\delta \beta_i}\right]_{\beta_i=cte:j\neq i}=0 \qquad (3)$$

Siendo n el número de observaciones, "$\hat{y}_i$" el valor estimado "y_i" el valor real, y "β_i" el parámetro objeto, que en este caso son los coeficientes de la ecuación (1).

2.3. Árboles de decisión

Los árboles de decisión estiman el valor de una variable desconocida (dependiente) a partir de variables conocidas (independientes), mediante reglas binarias del tipo "si" o "no" con las que se consigue repartir las observaciones de las variables conocidas en función de sus atributos y estimar así el valor de la respuesta.

Los árboles de decisión tienen una representación gráfica intuitiva, (ver Figura 4) donde cada regla o declaración se representa en un nodo con las posibles respuestas "si" / "no". Las respuestas unen los nodos a través de ramas generando la estructura del árbol. Por convenio, la respuesta afirmativa se dirige hacia la izquierda y la respuesta negativa hacía la derecha. Existen tres tipos de nodos, el nodo inicial del árbol, que se denomina raíz, su entrada está formada por los datos iniciales y su salida por dos ramas la opción "si" y la opción "no". Los nodos intermedios que están comunicados siempre con un nodo previo y con sus dos salidas. Y, por último, están los nodos terminales del árbol, en los que sus salidas corresponden al resultado propagado por el árbol y se denominan hojas.

En el entrenamiento de un árbol de decisión, se determinan las reglas para la raíz y los nodos del árbol siguiendo un proceso recursivo. Durante este proceso, las variables independientes, "X_i" se distribuyen en los nodos del árbol y se definen los valores discriminatorios para cada variable en cada nodo. Estos valores discriminatorios se utilizan para separar las muestras de entrenamiento en grupos más homogéneos en términos de la variable dependiente.

A medida que se desciende por el árbol, las muestras de entrenamiento se dividen en función de las reglas definidas en los nodos anteriores. En cada nodo, se busca la mejor división posible que maximice la homogeneidad de las muestras dentro de cada subgrupo resultante. Esto se logra mediante la selección de un valor discriminatorio óptimo para una variable específica en ese nodo.

En las hojas del árbol, se almacenan los valores correspondientes de la variable dependiente "Y" para cada observación de la muestra de entrenamiento. El conjunto de resultados en cada nodo terminal puede tener uno, dos o toda una distribuión de valores diferentes, (ver Figura 4).

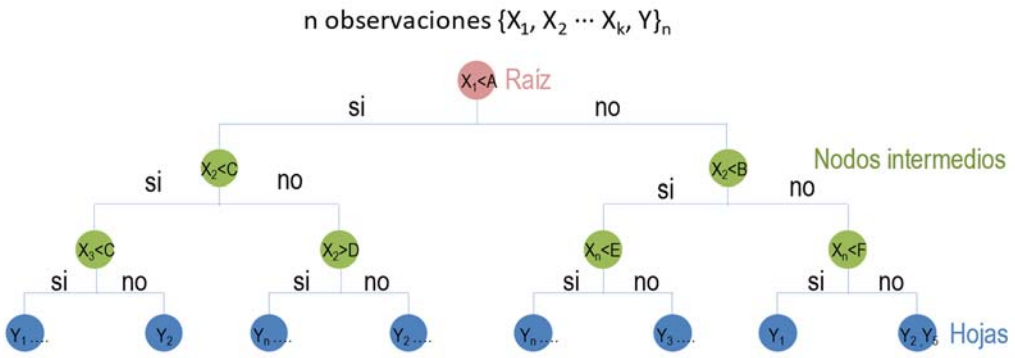


Figura 4. Estructura de un árbol de decisión

Esta Figura 4 muestra las diferentes partes de un árbol. En los nodos, se realiza una discriminación basada en el valor de la variable "X_i". Los resultados, que son los valores de las variables dependientes "Y_j" para cada observación, se acumulan en las hojas del árbol. El índice 'i' (que varía de 1 a k) se refiere a las variables independientes, como por ejemplo la temperatura. Por otro lado, el índice 'j' (que varía de 1 a n) indica el número de la observación.

El proceso de definición de las reglas se describe a continuación:

- Se selecciona una cualquiera de las variables independientes "X_i" y se le asigna un valor umbral aleatorio "A" (ver Figura 5 (a)). Este valor umbral divide a las observaciones de la variable "X_i" en dos grupos (ver Figura 5 (b)).
- Se repite el proceso variando el valor umbral, buscando el valor umbral optimo "$A_{óptimo}$" que minimiza el RSS. Para ello se calcula la derivada de la función de pérdida RSS respecto "A" y se iguala a cero para minimizarla (3). Cuando la ecuación resultante no tiene solución, el valor del parámetro "A" se busca mediante un método de optimización, como es el gradiente descendente [66]. El gradiente descendente es un algoritmo de optimización que calcula la derivada en un punto cualquiera de la función, y mediante pasos cada vez más pequeños va aproximando la derivada a un valor nulo consiguiendo encontrar el valor de la variable que minimiza la función, (ver Figura 5 (c)).
- Los pasos anteriores se realizan para cada variable "X_2"···"X_k". La variable que presente un RSS mínimo es la variable que se ubica en la raíz y la regla queda definida mediante su umbral óptimo.
- Una vez definida la raíz, el proceso se repite para cada nodo del árbol, buscando los valores umbrales y las variables "X_i" a las que se aplica, hasta alcanzar el criterio de parada.
- En cada hoja se acumulan los valores "Y_j", correspondientes a las observaciones. Estas observaciones se tienen que agregar dando lugar a un único valor por hoja. Lo más habitual es considerar el valor de la hoja como el promedio de todas sus observaciones, pero puede ser también el resultado un cuantil de las observaciones [52]. La predicción del árbol es el valor de una de sus hojas.

Este proceso se representa en las tres secciones que forman la Figura 5. La primera sección, (a), muestra la estructura del árbol. Para este ejemplo se considera el árbol más sencillo, sin nodos intermedios, 8 observaciones y se toman como variables la variable independiente "X_1" y la

dependiente "Y". En la sección central, (b), se muestra la gráfica que representa los pares de puntos formados por los valores de las variables en cada observación $(X_1, Y_1) \cdots (X_1, Y_8)$, y el valor de "A" que divide en dos conjuntos dichas observaciones según sean menor o mayor que el umbral. De cada subconjunto se calculan las estimaciones "$\hat{Y}_1$" e "$\hat{Y}_2$". En la gráfica de la derecha, (c), se muestra la búsqueda del valor de "A" minimizando el RSS.

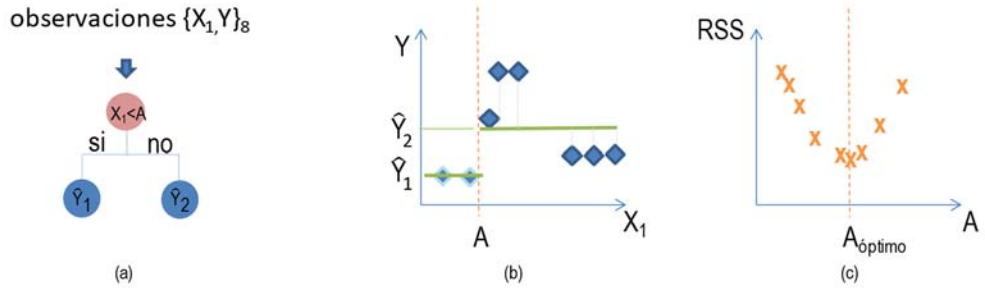


Figura 5. Estructura del árbol (a). Grafica de disgregación de observaciones según el valor de "A" (b). Gráfica de búsqueda del valor óptimo del umbral "A" (c)

Los árboles pueden formar estructuras complejas que permiten reducir rápidamente el error, es decir, el modelo se puede ajustar muy bien a las observaciones empleadas como entrenamiento con las que se ajustan los parámetros. Pero en algunas ocasiones, el ajuste genera un "overfitting" [216], que significa que el modelo se ha sobre-entrenado, es decir, es capaz de reproducir casi a la perfección el periodo de entrenamiento pero se reduce su capacidad al aplicarlo a nuevos datos. Para prevenir el problema de "overfitting" de los árboles, se suele limitar el tamaño del árbol.

El tamaño del árbol se controla mediante criterios de parada. Los criterios de parada o atienden directamente al tamaño del árbol, limitando el número de nodos directamente, definiendo que la incorporación de un nuevo nodo reduzca el error en al menos un valor mínimo, o indirectamente a través de limitaciones del número mínimo de observaciones para generar una nueva división en el nodo.

El tamaño del árbol también afecta al sesgo, "Bias", y a la varianza del modelo. El sesgo es la diferencia entre la esperanza matemática del estimador y el valor numérico que estima, mide cuánto se alejan en promedio las estimaciones de un modelo respecto a los valores reales. El sesgo refleja cómo de capaz es el modelo de aprender la relación real que existe entre las variables independientes y la variable respuesta. La varianza mide cuánto cambia el modelo en función de los datos utilizados en su entrenamiento. Idealmente, un modelo no debería modificar su resultado por pequeñas variaciones en los datos de entrenamiento. Si esto ocurre, es porque el modelo está memorizando los datos en lugar de aprender la verdadera relación entre las variables independientes y la variable respuesta.

Los árboles pequeños, de pocas ramificaciones, tienen poca varianza, pero no consiguen representar bien la relación entre las variables, es decir, tienen sesgos altos. En contraposición, los árboles grandes se ajustan mucho a los datos de entrenamiento, por lo que tienen muy poco sesgo pero mucha varianza.

La forma de conseguir el equilibrio entre el sesgo y la varianza es con la unión de varios árboles

que compensen ambos parámetros y esto se consigue con los métodos de agregación de modelos "ensemble" [217]. Estos métodos de agregación combinan múltiples modelos en uno nuevo con el objetivo de lograr el equilibro y consiguiendo a la vez mejores resultados que cualquiera de los modelos individuales originales. La clave para que los métodos de "ensemble" consigan mejores resultados es que los modelos que los forman sean lo más diversos posibles (sus errores no estén correlacionados).

Una vez definido el árbol, tanto su tamaño como los valores umbrales de cada nodo, para estimar una variable desconocida a partir de otras variables conocidas, se recorre el árbol hasta alcanzar uno de los nodos terminales (ver Figura 6).

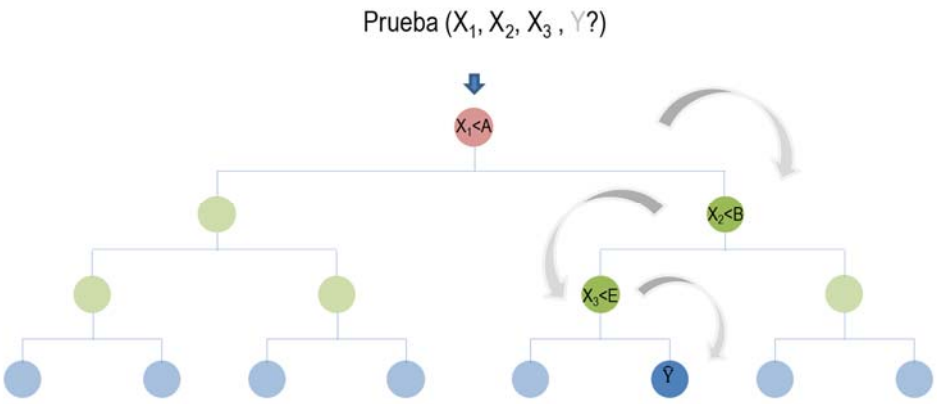


Figura 6. Estimación de la variable desconocida con un árbol de decisión

Los modelos basados en arboles de decisión presentan numerosas ventajas como son: su fácil interpretación y la capacidad para trabajar con variables numéricas o categóricas sin una distribución específica. Son modelos bastante insensibles a la influencia de "Outliers" y además pueden aplicarse tanto a problemas de regresión como de clasificación Como desventajas se pueden citar su sensibilidad cuando los datos del entrenamiento no son representativos.

La razón de por la que se han seleccionado en el estudio dos modelos basados en árboles de decisión "Random Forest" (RF) y "Gradient boosting" (GB) es porque generan los árboles de manera diferente y equilibran el sesgo y la varianza también con diferente método [218]. A continuación, se describen las particularidades de cada uno de ellos.

"Random Forest"

El algoritmo "Random Forest" (RF) [219], es un método no lineal supervisado que consta de un conjunto de árboles de decisión que se agregan mediante "Bagging". Cada árbol individual se entrena con un subconjunto distinto de los datos de entrenamiento, obtenido mediante muestreo aleatorio. Esto hace que cada árbol sea en sí mismo un modelo individual y distinto al resto. Estos modelos individuales tienen muy poco sesgo pero mucha varianza, y agregándolos se consigue reducir la varianza sin apenas aumentar el sesgo.

La mejora de RF con respecto a otros modelos de "Bagging" radica en su capacidad para seleccionar aleatoriamente el número de variables que intervienen en cada árbol. El trabajar con

subconjuntos de variables ayuda a de-correlacionar los árboles y, como resultado, se logra una mayor reducción de la varianza.

Esta mejora es crucial en la estimación de la producción de plantas fotovoltaicas, donde la radiación es una variable influyente. Cuando un modelo de RF está dominado por una variable, los árboles tienden a ser muy similares entre sí. La alta correlación entre los árboles dificulta el proceso de "Bagging" y, por lo tanto, no mejora significativamente el modelo. En los modelos RF regresivos, se considera óptimo trabajar en cada árbol con un tercio del número de variables.

El resultado del modelo es el promedio de las respuestas de todos los árboles individuales, como se muestra en la Figura 7.

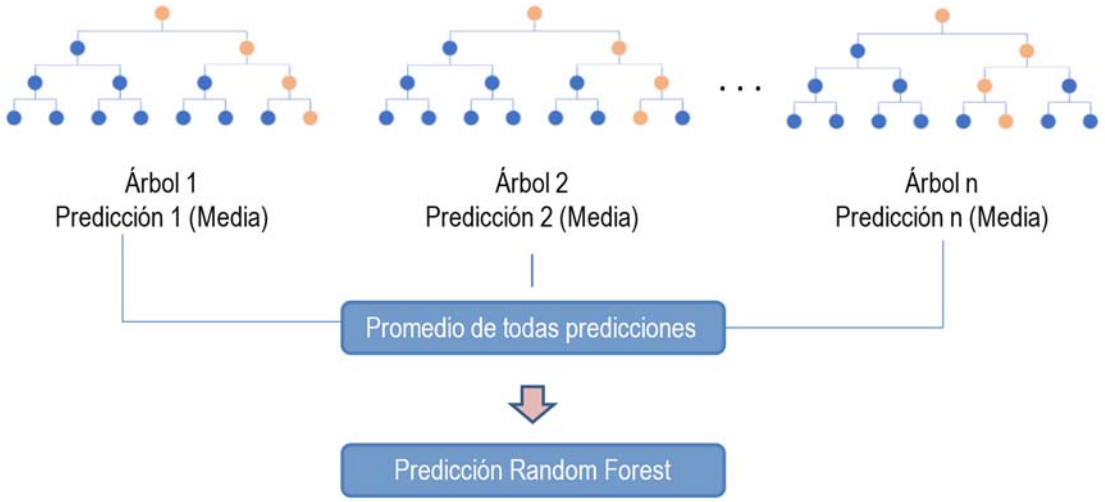


Figura 7. Funcionamiento del algortimo "Random Forest"

El algoritmo de "Random Forest" (RF) presenta una ventaja importante, y es que no sufre de "Overfitting" al aumentar el número de árboles creados en el proceso. A partir de un número determinado de árboles, la reducción del error se estabiliza, por lo que un exceso de árboles solo conlleva un esfuerzo computacional mayor. El problema de sobre-ajuste en este modelo está relacionado con el tamaño de los árboles, y tiene que ser controlado en la selección de los hiperparámetros del modelo. En general, RF es un modelo robusto y flexible que puede manejar diferentes tipos de datos y variables, lo que lo hace muy útil en muchas aplicaciones de aprendizaje automático.

"Gradient Boosting"

El algoritmo "Gradient Boosting" (GB) [220] es un método no lineal basado en árboles de decisión. GB realiza un "Boosting" como método de agregación de árboles. Este método consiste en un ajuste secuencial de árboles sencillos, de forma que cada árbol aprende de los errores del anterior. En cada iteración se ajusta un nuevo árbol a los residuos del anterior haciendo que el peso de las observaciones cambie en función de la bondad del ajuste de los árboles nuevos. Los árboles individuales no suelen ser muy grandes y tienen la misma estructura de ramas y hojas, tienen mucho sesgo, pero poca varianza. El sesgo se puede reducir ajustando secuencialmente los árboles, y además con ello se consigue reducir el error del modelo.

El resultado del algoritmo es una combinación de la primera predicción y las predicciones de los diferentes árboles generados.

El proceso de construcción del modelo comienza con un árbol de una única hoja "A_1", que suele tomar el valor medio de la variable dependiente en las observaciones del periodo de entrenamiento. La predicción "$\hat{y}_1$" de cualquier observación con este árbol da como resultado el mismo valor medio "$\bar{y}$". Al disponer de una predicción inicial se puede calcular el primer residuo "Res_1". El segundo árbol se construye buscando predecir "Res_1" a partir de las variables independientes. La predicción "$\hat{y}_2$" es la suma de la predicción "$\hat{y}_1$" y la predicción "$\widehat{Res}_1$", y con ella se puede volver a calcular un nuevo residuo. Este proceso iterativo se muestra matemáticamente en las ecuaciones (4). El criterio de parada puede ser un número máximo de árboles o bien un residuo mínimo. Cabe destacar que el residuo decrece a medida que se aumenta el número de árboles debido a que la predicción se aproxima gradualmente al valor buscado.

Árbol inicial	$A_1(x) \rightarrow \hat{y}_1 = \bar{y} \rightarrow Res_{1i} = y_i - \hat{y}_1$	
Iteración 1	$A_2(Res_1) \rightarrow \widehat{Res}_1 \rightarrow \hat{y}_2 = \hat{y}_1 + \lambda \cdot \widehat{Res}_1$ $Res_{2i} = y_i - \hat{y}_2$	
Iteración 2	$A_3(Res_2) \rightarrow \widehat{Res}_2$ $\hat{y}_3 = \hat{y}_2 + \lambda \cdot \widehat{Res}_1 + \lambda \cdot \widehat{Res}_2$ $Res_{3i} = y_i - \hat{y}_3$	**(4)**
	…	
Iteración n-1	$A_{n-1}(Res_{n-2}) \rightarrow \widehat{Res}_{n-2}$ $\hat{y}_{n-1} = \hat{y}_1 + \lambda \cdot \sum_1^{n-2} \widehat{Res}_i \rightarrow Res_{(n-1)i} = y_i - \hat{y}_{n-1}$	
Iteración n	$A_n(Res_{n-1}) \rightarrow \widehat{Res}_{n-1}$ $\hat{y}_n = \hat{y}_1 + \lambda \cdot \sum_1^{n-1} \widehat{Res}_i \rightarrow Res_n \approx 0$	

El valor λ introducido en las ecuaciones se denomina tasa de aprendizaje, "Learning Rate", y controla la velocidad de aprendizaje del algoritmo. Es notable que en GB los árboles no predicen los valores de "y", sino que predicen los residuos "Res", y la predicción "$\hat{y}$" se conforma con la agregación de la predicción inicial "$\hat{y}_1$" y las predicciones de los residuos de cada árbol. La representación gráfica del proceso se muestra en la Figura 8.

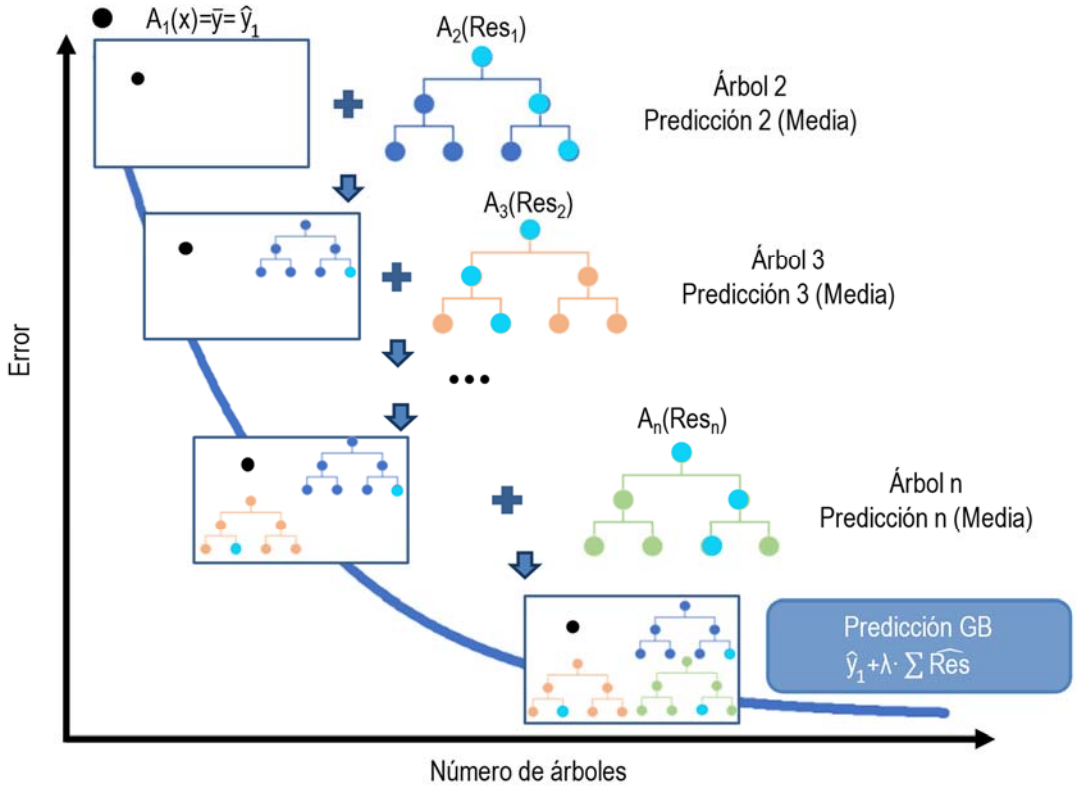


Figura 8. Funcionamiento del algortimo Gradient Boosting

GB es más rápido que RF y produce un modelo final más pequeño. Sin embargo, es susceptible de sobre-ajuste y más sensible al ruido de los datos de entrada que RF [221].

En general, los modelos basados en árboles de decisión son muy flexibles y se pueden aplicar a una amplia gama de problemas de regresión y clasificación con varias variables, lo que los convierte en uno de los métodos de "Machine Learning" más exitosos.

2.4. Redes Neuronales. "Multilayer Perceptron"

Las redes neuronales son algoritmos no lineales supervisados que imitan el funcionamiento del cerebro humano [66]. Están compuestas por unas unidades llamadas neuronas, que se conectan entre sí para reconocer patrones y relaciones en los datos. Estas neuronas y sus relaciones se organizan en capas, de manera que las capas iniciales extraen las características más simples de los datos, y a medida que se aumenta el número de capas, el algoritmo es capaz de captar relaciones más complejas.

Los algoritmos MPNN constan de una capa de entrada, una capa de salida y una o varias capas ocultas, cada una de ellas con sus neuronas. La conexión entre capas consecutivas se lleva a cabo relacionando las neuronas mediante pesos "W_i ", "V_i" y sesgos "b_i", (ver Figura 9).

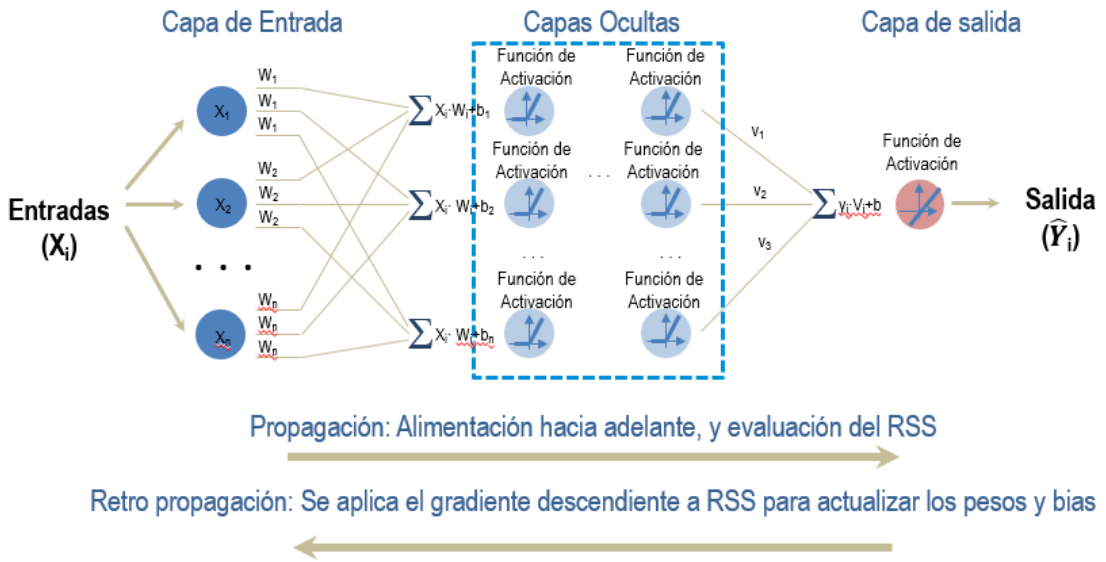


Figura 9. Funcionamiento del algoritmo Redes Neuronales

En las capas ocultas y en la de salida se encuentra la función de activación, que activa o desactiva las diferentes neuronas de la capa. Existente varias funciones de activación, la función escalón, la función limitante, la tangencial hiperbólica, la sigmoidal, etc. Entre ellas se escoge la función "Rectified Linear Unit" (ReLU) como función de activación en las capas ocultas, y para la capa de salida la función identidad, funciones sencillas y eficientes en los casos de estudio de esta tesis.

La función ReLU se define como el máximo entre cero y el valor de la variable de entrada, ecuación (5). Si la entrada es negativa, la salida es nula, y si es positiva, la salida es igual a la entrada, esto permite decidir si se activa o no una determinada neurona.

$$f(a) = \max(0, a);\ \text{siendo "a" el valor de la variable de entrada} \quad \textbf{(5)}$$

La función identidad devuelve el valor de la variable, ecuación (6).

$$f(a) = a\ \text{siendo "a" el valor de la variable de entrada} \quad \textbf{(6)}$$

El algoritmo MPNN ajusta sus parámetros, que son los valores de los pesos "W_i" y "V_i" y de los sesgos "b_i", durante el entrenamiento, consiguiendo disminuir el error del modelo. El proceso de ajuste se realiza en dos fases. La primera fase, la propagación, parte de los valores de la muestra de las variables de entrada al algoritmo "X_i", estas se propagan hasta la salida de la red obteniendo un resultado "$\widehat{Y}_i$". Con ese resultado se estima un RSS. La segunda fase, la retro propagación, minimiza el RSS, que actúa de función de pérdidas, "Loss Funtion" [222], modificando los valores de los pesos y "Bias" (ver Figura 9). A continuación, se explican con más detalle los procesos de propagación y retro propagación del entrenamiento de la red.

Propagación

La muestra de entrenamiento tiene "j" observaciones de las "i" variables conocidas "X_{ij}", y de la variable objetivo "Y_j". Para iniciar el proceso se asignan valores aleatorios a los pesos y se

consideran nulos los sesgos. Con esta premisa, se calculan los valores de las neuronas de la primera capa oculta, como el sumatorio del producto del valor de las variables del problema "X_{ij}" por su peso "W_i". Además, a este sumatorio se le añade el sesgo "b_i", que en la primera iteración tiene el valor de cero. Los valores de las neuronas son la entrada de la función ReLU. La función de activación da como resultado los valores "y_j" que son los diferentes valores de salida de cada neurona de la primera capa oculta.

Estos valores de salida de las neuronas de la primera capa oculta se multiplican por los pesos, correspondientes, se agrupan en sumatorios, se les añade un sesgo, que en la primera iteración tiene el valor nulo, pasando a ser los valores de las neuronas de la capa oculta siguiente. Esas neuronas, de nuevo serán activadas o desactivadas por la función ReLU dando como resultado el valor de salida de las neuronas de la capa. Este proceso se repite tantas veces como capas ocultas existan y para todas las neuronas de cada capa.

La capa de salida dispone solo de una única neurona, por lo que todos los valores de salida de todas las neuronas de la última capa oculta se multiplican por sus correspondientes pesos y se agrupan en un único sumatorio, ajustando el resultado con un sesgo, que también en la primera iteración tiene el valor nulo. Este valor es la entrada de la neurona de la capa de salida, que se activa en este caso con la función identidad, el resultado de la función es la estimación de la variable desconocida "$\widehat{Y}_j$" para la observación de las variables conocidas "X_{ij}". Con la predicción "$\widehat{Y}_j$" y el valor real "Y_j", se calcula el RSS.

Para entender mejor el proceso se expone un ejemplo, (ver Figura 10), de una red con dos neuronas en la capa de entrada, una capa oculta con una única neurona y la capa de salida con su neurona. La muestra de entrenamiento tiene seis observaciones de tres variables, dos conocidas, "X_1" y "X_2" y una variable objetivo "Y", tal y como se muestra en la tabla de datos de la Figura 10. Se asignan unos valores aleatorios a los pesos y sesgos, mostrados todos ellos en la figura. Se denomina "a_j" y "a'_j" a los valores de entrada de la neurona de la capa oculta y de la capa de salida respectivamente de la observación "j" y existen tantos valores de ellas como observaciones hay en la muestra. El valor de la neurona a la salida se le denomina "y_j" en la capa oculta y en la capa de salida en este caso es la propia predicción buscada para cada observación "$\widehat{Y}_j$". El RSS se calcula con la ecuación (3).

En este ejemplo se expone el caso de la quinta observación, se muestra cómo se calcula el valor de "a_5", se le aplica la función ReLU obteniéndose un valor de salida "y_5". Este valor se multiplica por el peso "V_1" y se le suma el valor del sesgo "b" para tener el valor de "a'_5". En la capa de salida, se aplica la función identidad y se obtiene el valor de la predicción para la observación quinta "$\widehat{Y}_5$". El proceso se repite con todas las observaciones, obteniendo el conjunto de las seis predicciones

"$\widehat{Y}_j$" correspondientes a cada observación de la muestra. Con las predicciones y los valores de "Y_j" se calcula el RSS del modelo.

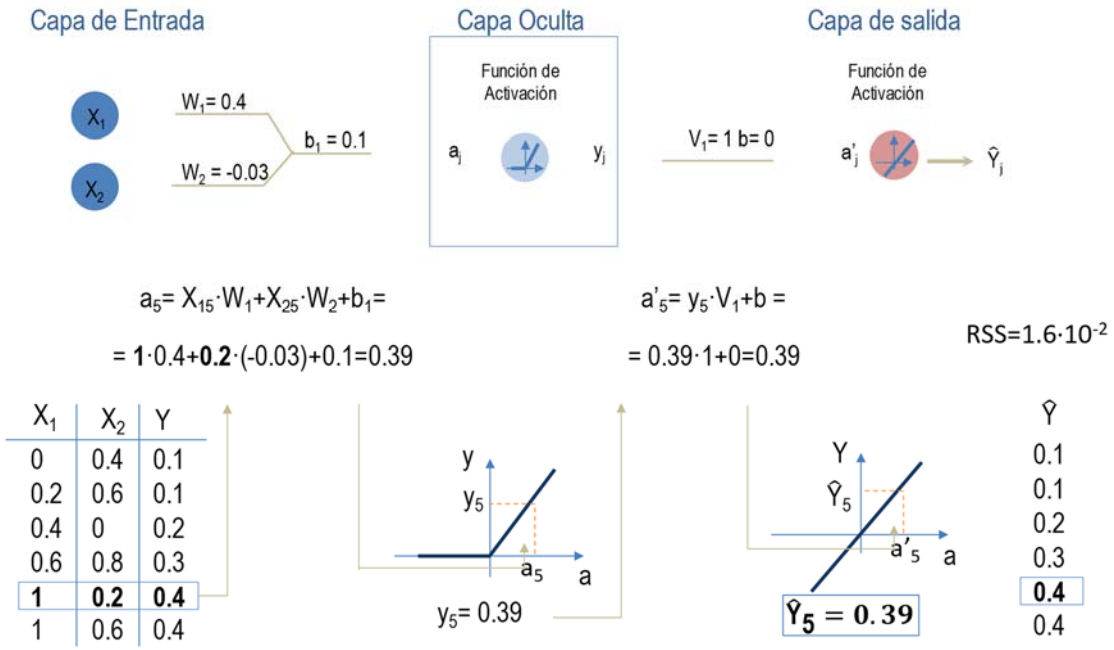


Figura 10. Ejemplo del proceso de propagación

Retro-propagación

El objetivo es ajustar los valores de los parámetros para minimizar el valor del RSS, partiendo del último sesgo, se va retrocediendo hasta llegar al primer peso.

Para encontrar los valores de los parámetros que minimizan el valor de RSS, hay que derivar la función respecto a cada uno de los pesos y sesgos de la red e igualarla a cero para obtener los valores ajustados (3).

Este método no es inmediato por dos motivos. El primero, es que la función RSS no depende directamente de los pesos y sesgos por lo que para derivar respecto a ellos hay que aplicar la regla de la cadena, que consiste en enlazar derivadas sucesivamente. El segundo se debe a que la ecuación resultante de igualar la derivada a cero no tiene solución por lo que hay que buscar el valor del parámetro aplicando el método del gradiente descendente. Para ajustar el paso del gradiente juega un papel importante en la convergencia el valor del hiperparámetro Tasa de Aprendizaje, "Learning Rate".

Continuando con el ejemplo Figura 10, en este caso hay que ajustar, los valores de los parámetros: "W_1", "W_2", "b_1", "V_1" y "b", comenzando por el último sesgo, "b" y acabando por el primer peso W_1.

Para minimizar el valor de RSS respecto de "b" se calcula la derivada de RSS respecto a él. Al no depender RSS directamente de "b", se aplica la regla de la cadena, ecuación (7).

$$\frac{dRSS}{db} = \frac{dRSS}{d\hat{Y}} \cdot \frac{d\hat{Y}}{db} \qquad (7)$$

El cálculo de cada uno de los dos factores de la ecuación (7) se presenta en las ecuaciones (8) y

(9):

$$\frac{dRSS}{d\hat{Y}} = \frac{d(Y-\hat{Y})^2}{d\hat{Y}} = \sum_{j=1}^{6} -2\cdot\left(Y_j-\hat{Y}_j\right) \quad \textbf{(8)}$$

$$\frac{d\hat{Y}}{db} = \frac{d(\sum y_j\cdot V_1+b)}{db} = 1 \quad \textbf{(9)}$$

Sustituyendo (8) y (9) en la ecuación (7), se obtiene la ecuación (10) para calcular el valor de "b". Esta ecuación se resuelve por aproximación a través del algoritmo del gradiente descendente, que busca el valor de "b" que anula la derivada.

$$\frac{dRSS}{db} = \sum_{j=1}^{6} -2\cdot\left(Y_j-\hat{Y}_j\right) \quad \textbf{(10)}$$

En el cálculo del resto de pesos y sesgos de la red, se trabaja de manera análoga, derivando RSS respecto de cada uno de ellos y recurriendo para ello a la regla de la cadena y el gradiente descendente para calcular la solución, ecuaciones (11) a (14).

$$\frac{dRSS}{dV_1} = \frac{dRSS}{d\hat{Y}}\cdot\frac{d\hat{Y}}{dV_1} = \sum_{j}^{6} -2\cdot\left(Y_j-\hat{Y}_J\right)\cdot\frac{d(\sum y_j\cdot V_1+b)}{dV_1} \quad \textbf{(11)}$$

$$= \sum_{i}^{6} -2\cdot\left(Y_j-\hat{Y}_J\right)\cdot\sum_{i}^{6} y_j$$

$$\frac{dRSS}{db_1} = \frac{dRSS}{d\hat{Y}}\cdot\frac{d\hat{Y}}{dy_j}\cdot\frac{dy_j}{da}\cdot\frac{da}{db_1} = \sum_{i}^{6} -2\cdot\left(Y_j-\hat{Y}_J\right)\cdot V_1\cdot 1\cdot 1 \quad \textbf{(12)}$$

$$\frac{dRSS}{dW_2} = \frac{dRSS}{d\hat{Y}}\cdot\frac{d\hat{Y}}{dy_j}\cdot\frac{dy_j}{da_j}\cdot\frac{da_j}{dW_2} = \sum_{i}^{6} -2\cdot\left(Y_j-\hat{Y}_J\right)\cdot V_1\cdot 1\cdot X_{2i} \quad \textbf{(13)}$$

$$\frac{dRSS}{dW_1} = \frac{dRSS}{d\hat{Y}}\cdot\frac{d\hat{Y}}{dy_j}\cdot\frac{dy_j}{da_j}\cdot\frac{da_j}{dW_1} = \sum_{i}^{6} -2\cdot\left(Y_j-\hat{Y}_J\right)\cdot V_1\cdot 1\cdot X_{1i} \quad \textbf{(14)}$$

Estos pasos, de propagación y retro propagación se repiten hasta que los valores de los residuos son muy bajos o se aplica un criterio de parada de iteraciones. Una vez entrenado el modelo, los parámetros quedan definidos y se mantendrán fijos en su aplicación a cualquier otra muestra de observaciones. Una nueva muestra relacionará sus observaciones a través de los pesos predefinidos en el entrenamiento, dando como resultado la predicción para esa muestra.

Las redes neuronales son muy flexibles debido a su gran número de parámetros, pesos y sesgos, lo que permite resolver prácticamente todos los problemas que se plantean tanto lineales como no lineales. Precisamente por su gran número de parámetros, son muy complejas y pueden no

funcionar en su punto óptimo si no se ajustan debidamente los hiperparámetros: número de capas, número de neuronas, y tasa de aprendizaje entre otros [190]. El ajuste de los hiperparámetros de estos modelos seleccionados es muy crítico y ayudará a su mejor convergencia y a minimizar su error.

2.5. Búsqueda de los hiperparámetros óptimos de los modelos

La selección adecuada de hiperparámetros consigue minimizar el error de los modelos. Existen diferentes técnicas de búsqueda de hiperparámetros (Sección 1.31.3, presentando cada una de ellas ventajas e inconvenientes cuando se aplican a un tipo de modelo por lo que es necesario seleccionar la técnica más adecuada para cada caso. En este apartado se explican las técnicas de búsqueda de hiperparámetros [223] utilizadas en esta tesis: "Grid Search", "Hyper-band" y "Random Search".

Búsqueda mediante red "Grid Search"

La técnica de búsqueda "Grid Search" consiste en una búsqueda sistemática y completa de todas las posibles combinaciones de un conjunto de valores de hiperparámetros predefinidos dentro de una retícula cartesiana. Esta metodología tiene la ventaja de que el óptimo encontrado es el óptimo global dentro del rango de búsqueda utilizado. A modo de ejemplo, si un modelo tuviera tres hiperparámetros diferentes, la malla de búsqueda de la solución sería la representada en la Figura 11. En esta figura cada uno de los ejes representa la variación de un hiperparámetro diferente y los puntos son cada una de las combinaciones de hiperparámetros que se utilizan en la búsqueda de la combinación óptima para un modelo determinado.

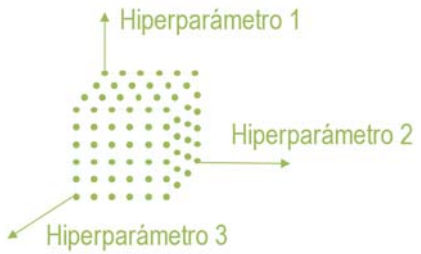


Figura 11. Búsqueda en red de una combinación de tres hiperparámetros

La limitación de la búsqueda en red es el número de combinaciones a estudiar. Si hay un número muy alto de combinaciones, la carga computacional es muy grande y es imposible llevarla a cabo. Esto sucede cuando el número de hiperparámetros a optimizar es elevado o hay hiperparámetros que presentan un rango amplio de valores posibles.

En el caso del rango amplio de posibles valores, se puede dar solución a la limitación mencionada, realizando la búsqueda mediante un proceso iterativo, en el que en cada iteración se modifica el tamaño de la retícula de la malla. El tamaño de la retícula es la diferencia entre los valores consecutivos de cada hiperparámetro (delta). El proceso iterativo se describe a continuación:

- Primera iteración: se trabaja con un mallado con un delta grande. La combinación de hiperparámetros que se obtienen como respuesta no es una optimización precisa, para afinarla es necesario realizar una nueva búsqueda.
- Segunda iteración: la red de trabajo tiene un delta más fino y el rango de valores de la misma se ajusta alrededor de los valores de los hiperparámetros con mejores resultados de la primera iteración.
- Sucesivas iteraciones: El proceso puede repetirse varias veces, reduciendo en cada iteración el tamaño del delta y el rango de valores de los hiperparámetros.

Al reducirse el valor del delta en cada iteración se mejora la precisión de la optimización. El número de iteraciones queda definido por la estabilidad en los errores de los modelos con combinaciones de valores consecutivos de hiperparámetros.

En la Figura 12 se representa el proceso iterativo de la búsqueda de la combinación de dos hiperparámetros, en el cual se realizan tres iteraciones para encontrar la combinación definitiva.

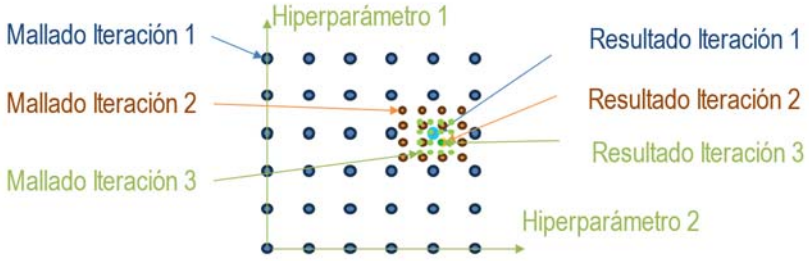


Figura 12. Búsqueda en red con iteración de una combinación de dos hiperparámetros

Existen casos, en los que el número de hiperparámetros es muy elevado y con un rango muy amplio de valores posibles. En estos casos, ni siquiera la fragmentación de la búsqueda en iteraciones sucesivas resulta suficiente para optimizar los hiperparámetros mediante una búsqueda en malla, por lo que se deben utilizar otro tipo de metodologías como las que se describen a continuación.

Búsqueda aleatoria

La búsqueda aleatoria consiste en una búsqueda de valores de hiperparámetros en un conjunto de rangos predefinidos [187], donde no se prueban todas las combinaciones sino solamente un número concreto de valores aleatorios. Al realizar una búsqueda parcial hay un ahorro de coste computacional.

En el proceso, se eligen aleatoriamente valores de entre todos los hiperparámetros que forman el espacio de búsqueda hasta alcanzar un número máximo. Estos valores se combinan generando un conjunto de modelos que se entrenan y se validan estimando el error de cada uno de ellos. El resultado óptimo buscado es el modelo que presenta un menor error.

La gran ventaja de este algoritmo es que al definir el número total de pruebas estas se pueden ajustar a la capacidad computacional de los equipos, por lo que es una alternativa de búsqueda cuando existen demasiadas combinaciones de hiperparámetros o limitaciones informáticas.

La aleatoriedad en la elección de los valores de los hiperparámetros puede suponer una ventaja adicional cuando en un hiperparámetro la diferencia entre los posibles valores de su rango no es constante, su delta es variable, ya que podría llegar a caracterizar la muestra de prueba mejor que la búsqueda sistemática de delta fijo. Sin embargo, a su vez, la aleatoriedad es una desventaja, por la dificultad para cubrir el espacio de búsqueda con una densidad de combinaciones constante, lo que genera nuevos valores sin cubrir entre los cuales se podría encontrar el conjunto óptimo.

Para ilustrar las ventajas y desventajas del método aleatorio, se presenta un ejemplo con dos hiperparámetros, ambos con un rango de valores de 0 a 1. El hiperparámetro 1 tiene un incremento constante de 0.2, resultando en un conjunto de 6 valores: {0, 0.20, 0.40, 0.60, 0.80, 1}. Por otro lado, el hiperparámetro 2 tiene un incremento variable, con un delta mayor en la primera mitad de su rango y un delta decreciente en la segunda mitad. Esto resulta en una mayor concentración de posibles valores para el hiperparámetro 2 en los valores más altos. El conjunto de valores para este hiperparámetro consta de 11 elementos: {0, 0.20, 0.40, 0.56, 0.68, 0.77, 0.85, 0.90, 0.95, 0.98, 1}.

Se establece un límite máximo de pruebas en 36 y se seleccionan aleatoriamente los valores de los hiperparámetros para cada prueba. La Figura 13 muestra una representación gráfica de los valores de los hiperparámetros seleccionados en el escenario de trabajo. En esta figura se destacan áreas donde este método permite aumentar el número de pruebas de modelos, así como áreas donde no se realizarán pruebas.

Para solucionar la falta de caracterización de algunas zonas, se puede aplicar la metodología de manera iterativa lo que permitiría aumentar la densidad de modelos en las zonas despobladas, mediante una nueva búsqueda centrada en un área concreta de valores de hiperparámetros.

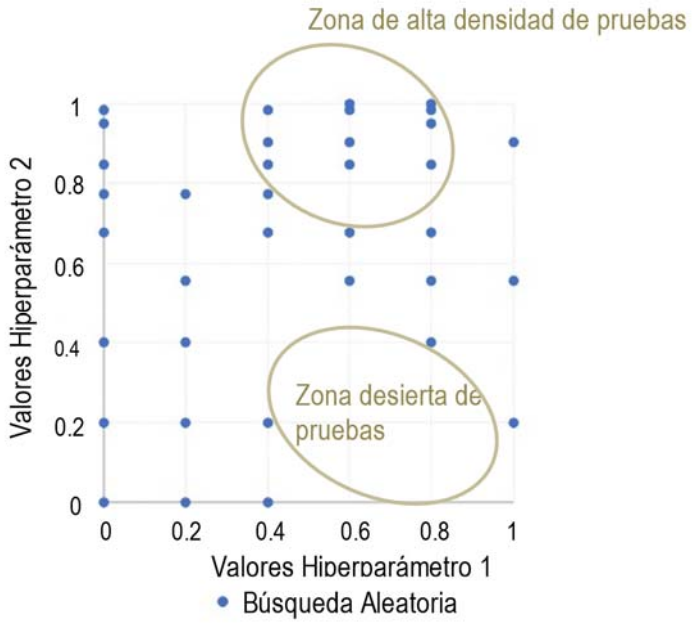


Figura 13. Búsqueda aleatoria

El inconveniente de este sistema es que se pueden no estar probando las mejores combinaciones. Por ese motivo se desaconseja su uso siempre que sea viable utilizar una red. Sí que este método

se considera complemento en los métodos de búsqueda, por su capacidad para aumentar de manera controlada poblaciones muestrales en zonas previamente acotadas mediante otras metodologías.

Búsqueda de hiper-banda

La búsqueda de hiper-banda [186] es un método basado en una búsqueda aleatoria y preparado para trabajar sobre un volumen importante de posibles combinaciones de hiperparámetros, porque utiliza la técnica de reducción mediante sucesivas separaciones en mitades iguales, "Successive Halving", lo que permite agilizar el proceso de búsqueda aunque existan numerosas combinaciones. Esto la hace especialmente recomendable cuando se trabaja con un gran número de hiperparámetros y cuando estos tienen un rango de valores amplio. Esta estrategia ha demostrado obtener resultados de alta calidad en un tiempo de cálculo reducido.

A continuación, se describe el proceso de búsqueda llevado a cabo por el método de hiper-banda:

- Se seleccionan al azar diferentes combinaciones de hiperparámetros y se calculan los resultados de los modelos a que dan lugar las combinaciones y el error que presentan, pero llevando a cabo solo unas pocas iteraciones, sin permitir la convergencia de los modelos. De esta forma no hay una gran carga computacional y el tiempo de cálculo es relativamente rápido.
- Se desestiman la mitad de los modelos que presentan un mayor error.
- Se les incrementa ligeramente el número de iteraciones, y los modelos que permanecen en la búsqueda continúan minimizando su error. Es importante resaltar que los modelos no comienzan los cálculos de nuevo, sino que continúan desde la última iteración. Cuando se alcanza el número de iteraciones definido, nuevamente se desestiman la mitad de los modelos que presentan un error más alto.
- El proceso se repite de manera recurrente, aumentando las iteraciones, analizando errores y eliminando en cada valoración la mitad de los modelos. Por eso, a la técnica de optimización empleada se le denomina reducción a la mitad sucesiva.

El diagrama del proceso de búsqueda se muestra a continuación, Figura 14:

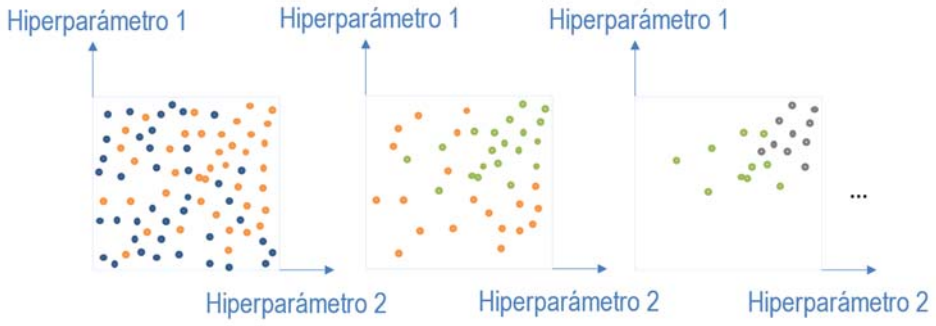


Figura 14. Búsqueda con el método de hiper-banda

La gran ventaja de esta metodología es su velocidad, ya que cuando se trabaja con un volumen considerable de combinaciones el número de iteraciones para encontrar el resultado de cada una

de ellas es muy bajo compensando el esfuerzo computacional, y cuando se van reduciendo modelos aumentan el número de iteraciones para mejorar la convergencia de los modelos que permanecen como candidatos idóneos.

El método de hiper-banda puede proporcionar buenos resultados, pero no garantiza la selección de los mejores valores de hiperparámetros, ya que favorece a los modelos que convergen rápidamente, descartando combinaciones de hiperparámetros que podrían generar un modelo excelente, pero que requieran de más iteraciones para converger. El hecho de que un modelo presente un error alto en las primeras iteraciones, no necesariamente indica que sea un mal modelo.

2.6. Cálculo del intervalo de confianza de la predicción.

El intervalo de confianza del resultado de los modelos permite determinar la incertidumbre de los mismos. En este apartado se van a describir cómo se calculan los intervalos de confianza de los resultados de los modelos objeto de estudio en esta tesis.

El primer paso para definir intervalos de confianza consiste en verificar si las variables de interés siguen una distribución Normal. En una distribución normal existe una relación entre la desviación estándar, el nivel de confianza y los valores límite. Así se puede definir un intervalo de amplitud $\pm 2\sigma$ que tendrá un nivel de confianza del 95.5%, o un intervalo de amplitud $\pm 3\sigma$ cuyo nivel de confianza es 99.7% [201].

Existen multitud de pruebas para verificar la normalidad, siendo las más sencillas los métodos gráficos. Entre ellos se encuentran el histograma y el denominado gráfico de cuantiles, Q-Q, donde se representan los datos estandarizados frente a la distribución normal estándar.

Cuando las variables no siguen una distribución normal, los modelos de ML seleccionados tienen algoritmos propios que permiten calcular los intervalos de confianza de la predicción. Estos algoritmos son el Quantile Regression (QR) para MLR, el Quantile Regression Forest (QRF) para RF, y el Quantile Gradient Boosting (QGB) para GB. Para las redes neuronales no hay un algoritmo concreto, sino que se modifica la red neuronal incluyendo capas específicas que permiten presentar la solución del modelo como una distribución de posibles soluciones en lugar de un único valor y, a partir de la distribución de soluciones, se calcula el intervalo de confianza.

Cálculo del intervalo de confianza del algoritmo de Regresión Lineal Multiple

El algoritmo que se utiliza es el Quantile Regression (QR) [224]. Este algoritmo es una adaptación del algoritmo MLR, que permite obtener el resultado en forma de percentiles, con los cuales se puede obtener el intervalo de confianza.

La diferencia entre ambos algoritmos está, no solo, en la forma de presentación de los resultados, sino también en la forma de definición de los parámetros del modelo. MLR estima como resultado la media de la variable objetivo, optimizando los coeficientes de la función presentada en la

ecuación (1) con el método de mínimos cuadrados. QR estima como resultado cualquier cuantil de la variable objetivo a través de las mismas variables que MLR, pero los coeficientes de la ecuación (15) que relacionan las variables, dependen del cuantil que se desea calcular.

Es reseñable que el resultado puntual del MLR es la media de la producción mientras que en el QR se estima como resultado medio ajustado, la mediana, que es el cuantil con "τ" igual a 50%, por lo que los resultados de la estimación puntual no serán coincidentes con los obtenidos mediante MLR.

$$y=\beta_0(\tau)+\beta_1(\tau)\cdot x_1+\ldots+\beta_p(\tau)\cdot x_p+\varepsilon \quad \textbf{(15)}$$

Siendo "y" la respuesta, "x_i" las variables independientes, $\beta_0(\tau)$, el termino independiente, $\beta_i(\tau)$, los coeficientes de la regresión que dependen del percentil "τ" y "ε" el error. Los coeficientes β, son calculados para los cuantiles de manera que se minimice la desviación absoluta de la mediana.

QR es un modelo que permite calcular los intervalos de confianza de la predicción en base a los cuantiles con una única ejecución del modelo, lo que supone un ahorro de tiempo y demanda computacional y, además el resultado no tiene por qué ser un intervalo simétrico pudiendo así representar mejor algunos comportamientos de variables concretas.

Cálculo del intervalo de confianza del algoritmo "Random Forest"

El algoritmo que se utiliza es el "Quantile Regression Forest" (QRF) [225,226]. Este algoritmo es el mismo algoritmo RF, pero presentando la respuesta de la predicción con cuantiles en lugar de con valores medios. De esta forma es posible evaluar intervalos de confianza de la predicción llevada a cabo con modelos RF.

Para explicar cómo se calculan los intervalos del resultado de una muestra test con un QRF se supone el modelo RF definido con sus hiperparámetros y parámetros que ha sido previamente aplicado a la muestra y que ha generado una predicción.

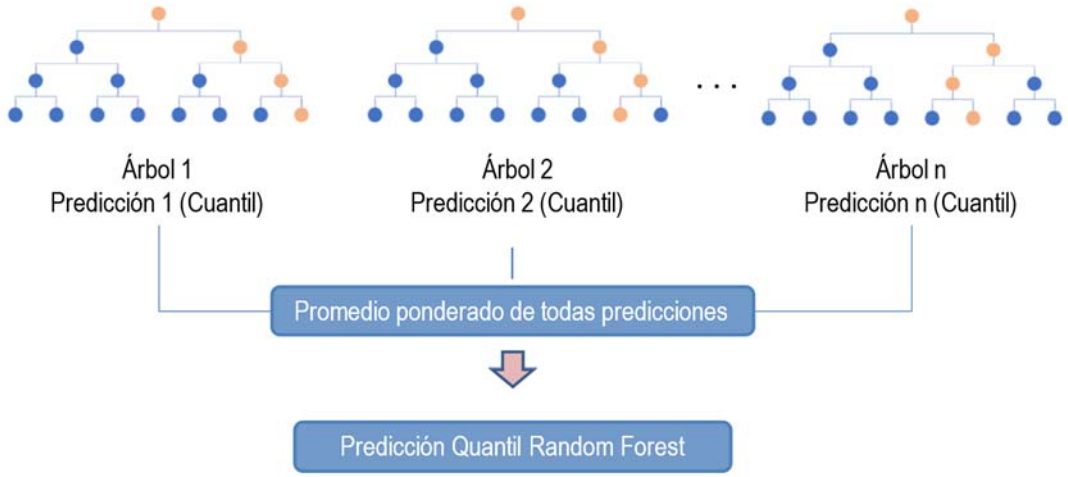


Figura 15. Funcionamiento del algortimo "Quantil Random Forest"

Las diferentes observaciones de la muestra test recorren los mismos árboles que se habían configurado en el modelo RF hasta llegar a una hoja. Hasta aquí el proceso es idéntico, la diferencia está en cómo se trabajan los conjuntos de valores de cada hoja para dar una respuesta. En el QRF la respuesta de cada árbol es la probabilidad de que la predicción tenga un valor menor a la observación "y" o el valor de la variable "y" por debajo del cual se encuentra un porcentaje "τ" dado de observaciones, lo que se denomina cuantil. La respuesta global del modelo es el promedio ponderado de la respuesta de todos los árboles. El coeficiente de ponderación es el porcentaje de muestras que llegan a cada hoja.

El esquema de trabajo se representa en la Figura 15. Esta figura es una modificación de la Figura 7, donde se indica que el resultado de cada árbol es el cuantil en lugar de la media, y como resultado final el promedio ponderado del resultado de todos los árboles del modelo en lugar de un promedio.

QRF permite obtener intervalos asimétricos de las predicciones del modelo RF. Así por ejemplo si se pretende obtener el intervalo de confianza del 90%, se tomará como límite superior del intervalo el percentil 95 de las respuestas de cada árbol y como límite inferior del intervalo el percentil 5 de las respuestas de cada árbol.

Cálculo del intervalo de confianza del algoritmo "Gradient Boosting"

El algoritmo "Quantil Gradient Boosting" (QGB) es el algoritmo GB, pero la predicción se presenta como el valor del cuantil asociado a una incertidumbre definida permitiendo así calcular el intervalo de confianza de la predicción realizada con GB.

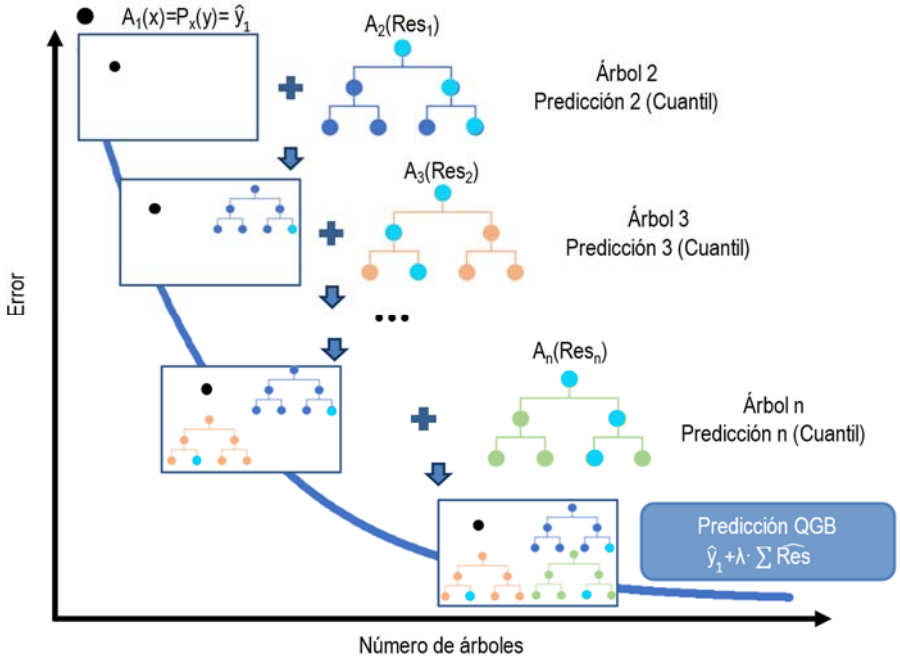


Figura 16. Funcionamiento del algortimo "Quantil Gradient Boosting"

El modelo QGB queda definido con los hiperparámetros y los parámetros que definen el recorrido de las observaciones modelo GB asociado a la misma predicción. La variación reside en la manera

de extraer la respuesta, el valor que se considera resultado de las hojas. En el QGB el valor de las hojas de los árboles es el cuantil especifico de los valores de las observaciones de cada hoja. Los árboles están igualmente relacionados, y la solución es la predicción inicial más la tasa de aprendizaje multiplicada por el sumatorio de las soluciones de cada árbol individual.

Si se pretende estimar un intervalo de confianza de un 90% de la predicción con un modelo QGB, se calculará el valor del cuantil 95 del conjunto de respuestas de las hojas terminal de cada árbol para definir el límite superior, y de igual forma se calculará el cuantil 5 para el límite inferior. El proceso es análogo al algoritmo GB y se representa en la Figura 16.

Los intervalos definidos con este modelo son también asimétricos por tener un origen basado en cuantiles.

Cálculo del intervalo de confianza del algoritmo de redes neuronales

Las redes neuronales no disponen de un algoritmo específico para calcular el intervalo de confianza de la predicción, sino que este se calcula modificando la propia red para que el resultado del modelo se obtenga como una distribución de posibles resultados. A partir de esta distribución, se puede estimar un intervalo de confianza.

El intervalo contemplará los dos tipos de incertidumbre, la incertidumbre aleatoria, que está asociada a la aleatoriedad inherente a los datos y la incertidumbre epistémica, que se refiere a la incertidumbre en el ajuste de los parámetros del modelo. En las redes neuronales es posible calcular un intervalo de confianza específico para cada tipo de incertidumbre y también hacerlo de manera conjunta.

Intervalo de confianza aleatorio:

Para calcular el intervalo de confianza asociado a la aleatoriedad del dato, se modifica la capa de salida de la red neuronal. En lugar de obtener un único valor para cada observación, se obtienen los parámetros que caracterizan una distribución Normal de posibles valores resultado para cada observación. Al representar el resultado de la red en forma de una distribución Normal con parámetros μ y σ para cada observación, el cálculo del intervalo de confianza es directo.

La presencia de un gran número de valores posibles para cada observación permite la aplicación del teorema del límite central, lo que nos permite considerar que la distribución de los resultados es una distribución Normal.

Esta modificación no afecta los a valores de los hiperparámetros y parámetros de la red, que mantienen su estructura original. La única alteración se da en la forma de expresar el resultado.

En la Figura 17, se muestra un ejemplo en el que se ha calculado la incertidumbre aleatoria de un conjunto de datos cualesquiera. En esta gráfica se representan las observaciones con puntos negros, la tendencia de estas con una línea azul y la incertidumbre del modelo, sombreada en gris. Si se analizan las observaciones se ve como a medida que crece el valor de la variable genérica representada, la dispersión de predicción es mayor, y la incertidumbre también se amplía. El intervalo tendrá mayor o menor amplitud en función de su nivel de confianza coincidiendo el 99%

prácticamente con el final de la zona gris.

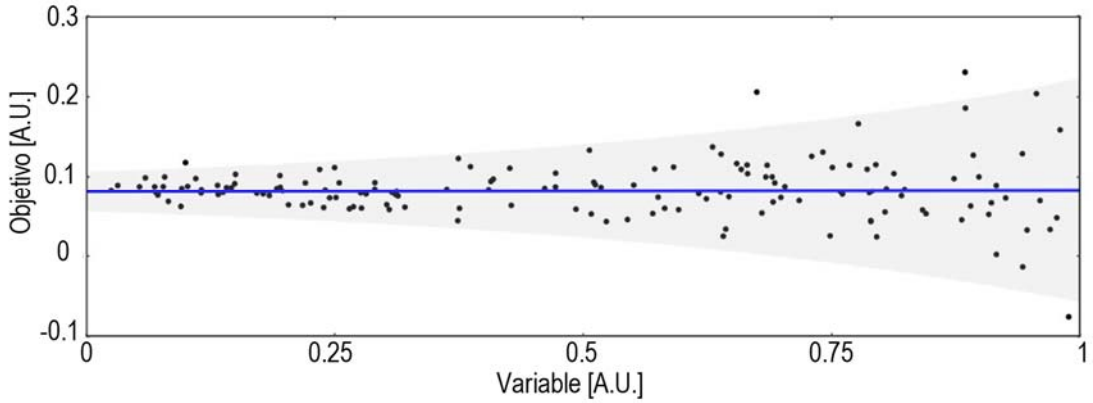


Figura 17. Incertidumbre aleatoria de una red neuronal

Intervalo de confianza epistémico:

Para calcular el intervalo de confianza asociado al modelo y tener en cuenta la sensibilidad de las redes neuronales a los cambios y desviaciones en los parámetros, se llevan a cabo múltiples pruebas del modelo utilizando la misma muestra de datos [208]. Durante estas pruebas, los parámetros del modelo se van modificando para obtener diferentes resultados. La cantidad de pruebas realizadas debe ser lo suficientemente grande para que, al aplicar el teorema del límite central, los resultados puedan aproximarse a una distribución normal. Esto permite calcular directamente el intervalo de confianza.

En la Figura 18 se representan los resultados del abanico de cien modelos generados al introducir variaciones en los parámetros del modelo para el mismo ejemplo anterior. En forma de puntos negros se representan las observaciones reales, la línea azul de mayor grosor representa la tendencia de ellas y cada línea de color es el resultado de uno de los modelos probados. También, como en la incertidumbre aleatoria, la incertidumbre epistémica aumenta cuando la dispersión de los datos reales es mayor y cuando hay menos datos.

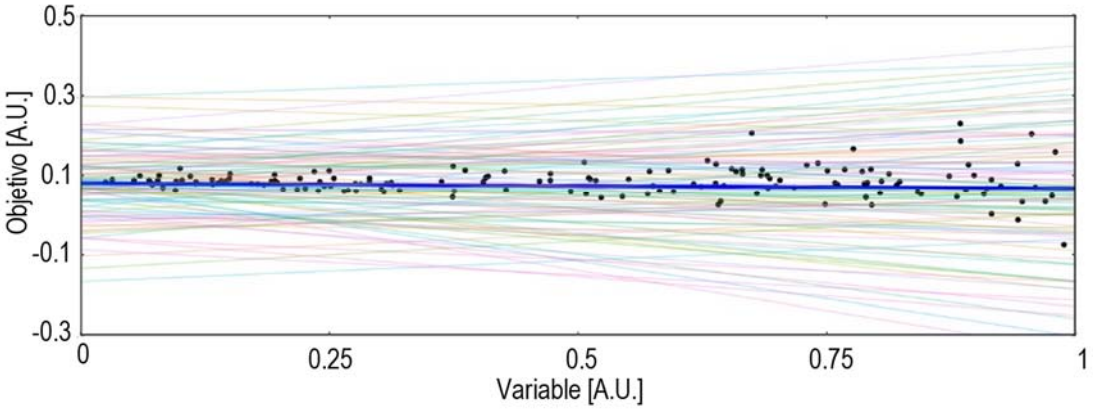


Figura 18. Incertidumbre epistémica de una red neuronal

Intervalo de confianza combinado:

El intervalo de predicción del modelo engloba a ambos intervalos, aleatorio y epistémico. Para calcular este intervalo, existen dos enfoques: uno manual, que implica analizar en cada observación cuál de los dos intervalos calculados de manera independiente contiene al otro; y otro enfoque que implica incorporar en el algoritmo de manera conjunta las dos modificaciones ya descritas, obteniendo el resultado directamente.

Estas modificaciones combinadas generan numerosas pruebas, cada una con una variación en los parámetros del modelo. Pero además los resultados de estas pruebas son distribuciones Normales. Por lo que cada prueba posee su propia incertidumbre, que se calcula a partir de la distribución Normal de sus resultados.

Siguiendo con el ejemplo, se calcula el intervalo combinado. En esta ocasión no se representan las observaciones ni su tendencia, sino solamente el resultado de las dos incertidumbres conjuntas. En la Figura 19, se representan en color, los resultados puntuales de los diferentes modelos y con sombreado gris la incertidumbre aleatoria de cada uno de ellos.

En la mayoría de los modelos se solapan las incertidumbres aleatorias a excepción de los modelos limítrofes de la incertidumbre epistémica. Será pues la incertidumbre aleatoria de estos modelos limítrofes la envolvente de la incertidumbre global. La amplitud del intervalo de confianza dependerá del nivel de confianza seleccionado coincidiendo el 99% prácticamente con los límites de la envolvente de la incertidumbre.

Mediante la metodología descrita es posible, por tanto, calcular intervalos de confianza de la predicción de las redes neuronales.

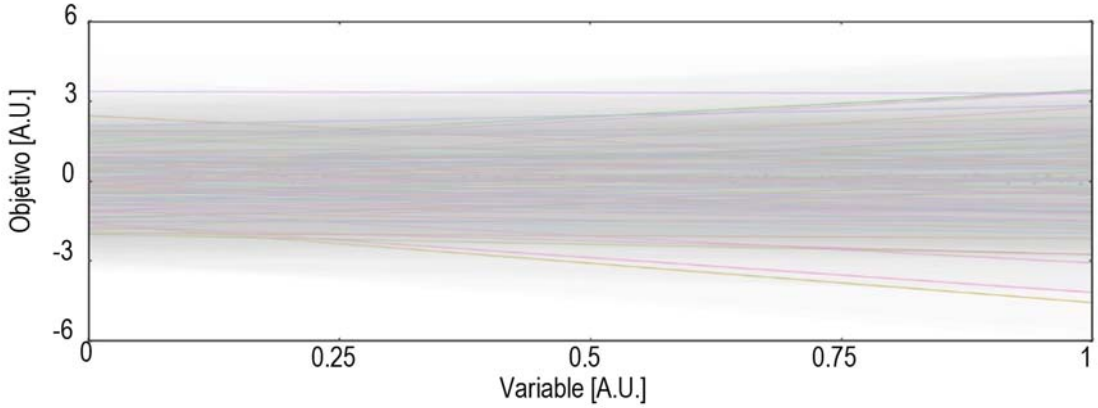


Figura 19. Incertidumbre aleatoria y epistémica de una red neuronal

2.7. Métricas

Para evaluar el comportamiento de cualquier modelo se utilizan las métricas e índices que proporcionan una estimación de las diferencias entre los valores calculados por el modelo y los valores reales. Existen numerosas métricas para evaluar los modelos como el RMSE, MAE, coeficientes de correlación, etc [227]. Las métricas que se van a utilizar en este trabajo para

evaluar la estimación puntual de los modelos son: nRMSE, nMAE, nBias y SS; y en el caso de la valoración del intervalo de confianza se tomará como métrica la mediana de la amplitud del intervalo en todas las observaciones.

La raíz cuadrada del error cuadrático medio normalizada, "Normalised root-mean-square error" (nRMSE), ecuación (16), es la raíz cuadrada de las medias de las desviaciones al cuadrado, siendo las desviaciones las diferencias del valor estimado y la medida real. Está métrica identifica especialmente desvíos en las estimaciones.

$$nRMSE=100\cdot\sqrt{\frac{1}{N}\cdot\sum_{i=1}^{N}\left(\hat{y}_i-y_i\right)^2} \quad \textbf{(16)}$$

El Error Absoluto medio normalizado, "Normalised Mean Absolute Error" (nMAE) es el promedio del valor absoluto de las desviaciones. Representa la divergencia entre los resultados obtenidos en valor absoluto y se calcula según la ecuación (17).

$$nMAE=100\cdot\frac{1}{N}\sum_{1}^{N}\left|\hat{y}_i-y_i\right| \quad \textbf{(17)}$$

Donde $\hat{y}_i$ e y_i son los valores estimados y reales respectivamente, y N el tamaño de la muestra de prueba.

El sesgo normalizado (nBias): es la media de las desviaciones. Representa el error sistemático, y se calcula con la ecuación (18).

$$nBias=100\cdot\frac{1}{N}\sum_{1}^{N}\left(\hat{y}_i-y_i\right) \quad \textbf{(18)}$$

Donde $\hat{y}_i$ e y_i son los valores estimados y reales respectivamente y N el tamaño de la muestra de prueba.

Además, para comparar los modelos entre sí se puede utilizar el "Skill Score" (SS) que evalúa el resultado con respecto al modelo de referencia. Cuanto mayor es el valor, mejor es el modelo, sin embargo, cuando el valor es bajo significa que no existe mejora con respecto al modelo de referencia. En este caso, como modelo referencia se utilizará el MLR, por ser el más sencillo de todos los que se han probado. La relación matemática que define esta métrica se presenta en la ecuación (19).

$$SS=100\cdot\frac{1\text{-}RMSE_{modelo}}{RMSE_{referencia}} \quad \textbf{(19)}$$

Para evaluar el intervalo de confianza de un modelo se utilizan métricas específicas y diferentes a las que han valorado la estimación puntual. La métrica principal que caracteriza la bondad de un intervalo es su amplitud.

La amplitud del intervalo (L), mide la precisión de la estimación. Se calcula como la diferencia entre

el límite superior L_s y el inferior L_i, ecuación (20). Intervalos amplios proporcionan estimaciones imprecisas, mientras que intervalos estrechos proporcionan estimaciones precisas. La amplitud del intervalo da idea del margen de error de la estimación.

$$A_L(x_1,x_2\ldots x_n)=L_s(x_1,x_2\ldots x_n)-L_i(x_1,x_2\ldots x_n) \quad \textbf{(20)}$$

La mediana de la amplitud del intervalo es la mediana de las diferencias de los límites superior e inferior de cada una de las observaciones, ecuación (21). Este indicador define la precisión de cada uno de los intervalos calculados.y se presentará en unidades porcentuales

$$\text{Mediana}(A_L(x_1,x_2\ldots x_n))\cdot 100 \quad \textbf{(21)}$$

El análisis del conjunto de todas estas métricas permite valorar el modelo más adecuado para aplicar en la predicción de la producción de las plantas fotovoltaicas y en la irradiancia.

3. Aplicación de los modelos ML a la generación fotovoltaica

3.1. Introducción

En este capítulo se evalúan los algoritmos Random Forest (RF), Gradient Boosting (GB) y Artificial Neural Network (ANN) aplicados a la simulación de la producción de plantas fotovoltaicas. El objetivo de esta evaluación es determinar cuál de estos métodos se adapta mejor a esta tarea, comparándolos entre sí y con el algoritmo de referencia Multiple Linear Regression (MLR).

Para llevar a cabo la evaluación, se trabajará con tres plantas fotovoltaicas diferentes. En cada planta, se evaluará la producción de uno de sus inversores utilizando los cuatro modelos mencionados anteriormente, dando lugar a doce escenarios de trabajo diferentes.

En cada escenario será necesario seleccionar los hiperparámetros óptimos. Para ello se buscará entre miles de modelos el conjunto de modelos que serán entrenados y evaluados. Posteriormente, se entrenarán los modelos para definir sus parámetros internos y se predecirá la producción y su intervalo de confianza para una muestra de prueba. Los resultados serán evaluados mediante diferentes métricas para determinar la eficacia y rendimiento de cada modelo.

El mejor modelo resultante de esta evaluación proporcionará una referencia respecto a la cual se puede comparar la producción real, lo que permite identificar pérdidas e indisponibilidades en períodos cortos de tiempo. Al conocer la existencia de una pérdida de energía, se pueden tomar medidas rápidas para solucionar las deficiencias de la planta. Esto es especialmente relevante en plantas de gran tamaño, donde es difícil identificar una pérdida de generación por la disparidad de la producción en los diferentes inversores. Esta disparidad en la producción puede originarse porque la irradiancia recibida en los módulos de la planta y la medida en las estaciones no es homogénea, por la extensión y el enorme número de paneles. Cuando una pérdida de generación a nivel de inversor o de caja perdura en el tiempo, deriva en pérdidas significativas de la planta a largo plazo.

En las siguientes secciones de este capítulo, se describirán las plantas de trabajo, la metodología utilizada y se presentarán y discutirán los resultados obtenidos de los escenarios de estudio.

3.2. Metodología

En este apartado se explica la metodología que se ha seguido para ensayar los algoritmos seleccionados que se aplicarán a un inversor de cada una de las tres plantas fotovoltaicas estudiadas.

Primero se describen las plantas fotovoltaicas, detallando las distancias que hay entre el inversor seleccionado y los puntos de medición en cada planta. Luego se explica el proceso que se sigue para definir el modelo que servirá para predecir la producción.

Descripción de las plantas

Se han seleccionado tres plantas de gran tamaño, que presentan diferencias en la topografía, en el número de estaciones de medición y en la distancia de estas estaciones de medición a los módulos de generación. Estas diferencias influyen en el comportamiento de los modelos y van a

permitir ensayar los algoritmos en distintas condiciones. En cada una de las plantas se ha elegido un inversor y su campo solar asociado. Con sus datos se generarán los modelos.

La diferente topografía entre las plantas implica que habrá variaciones en el comportamiento de los modelos, especialmente entre las plantas situadas en zonas más planas y aquellas que presenten terrenos más complejos con altas pendientes. Las pendientes en el terreno generan irregularidades en la irradiancia recibida, debido a pequeños desalineamientos de los “strings” lo que provoca una producción desigual de energía fotovoltaica en los diferentes puntos de la planta e incluso en el campo solar del inversor.

El número de estaciones de medición influye en los modelos ya que disponer de varias estaciones permite considerar posibles variaciones debidas, por ejemplo, a la nubosidad con mayor precisión.

La distancia entre la estación de medición y los módulos del campo solar del inversor también influirá en el comportamiento de los modelos. La nubosidad no se percibirá de la misma forma a diferentes distancias, lo que afectará la precisión de los modelos. Por lo tanto, la distancia entre los módulos del inversor que se estudia y la torre de medición es un factor determinante en el comportamiento de los modelos.

A continuación, se detallan brevemente las características de las plantas con las que se ha trabajado.

La planta PV1 se encuentra en un terreno de topografía compleja, con colinas que alcanzan los 25 metros de altura. Esta característica provoca que la irradiancia solar recibida varíe en diferentes puntos de la planta. La estación meteorológica de la planta está situada a 400 metros del inversor seleccionado para el estudio y a unos 600 metros de los módulos generadores conectados a este inversor.

Por otro lado, la planta PV2 se sitúa un terreno plano, por lo que recibirá una irradiación uniforme en toda su extensión. Esta planta cuenta con dos estaciones meteorológicas y el inversor seleccionado se encuentra junto al campo solar. Las estaciones están a una distancia de aproximadamente 250 y 800 metros del campo solar, respectivamente.

Finalmente, la planta PV3 se localiza en un terreno con una ligera pendiente, pero con la misma orientación, lo que permite una irradiación uniforme cuando el sol está perpendicular a la pendiente del terreno. Esta planta cuenta con cuatro estaciones meteorológicas y el inversor elegido se encuentra en el centro del campo solar. Las estaciones están a una distancia que varía entre los 450 metros y 1 kilómetro del campo solar.

En la Figura 20, se muestran las distancias existentes entre las estaciones de medición, el inversor y el punto más cercano de los módulos que forman el campo solar en cada planta.

La información con la que se configuran los modelos se toma de las estaciones de medición y de los inversores. En la Tabla 1, se enumera para cada planta sus características principales, el rango de potencia, el periodo de información y la disponibilidad de datos, así como el número de estaciones meteorológicas que dispone cada planta.

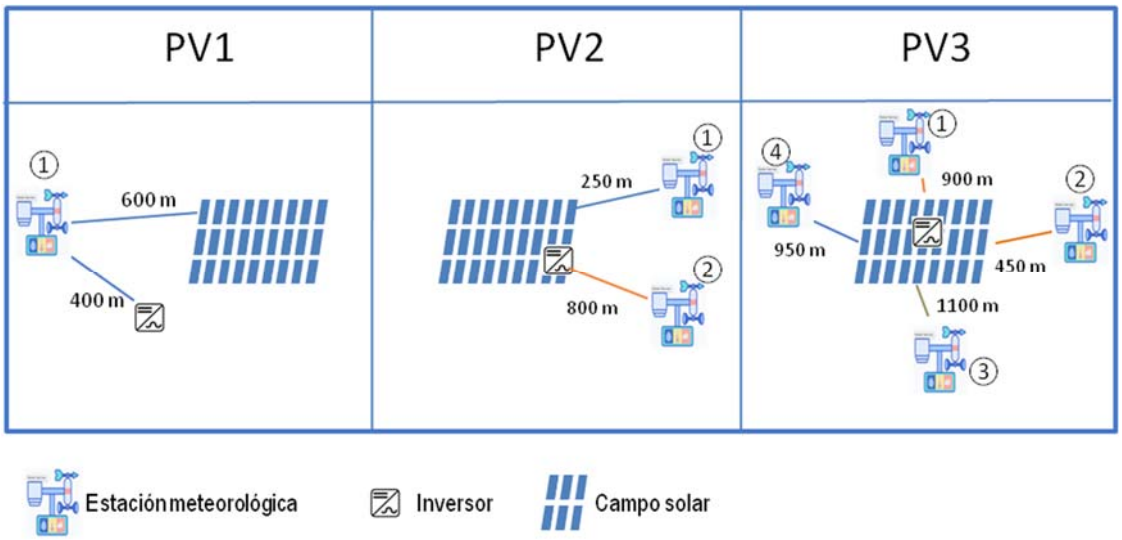


Figura 20. Distancias entre los elementos de las plantas

Tabla 1. Información general de las plantas

		PV1	PV2	PV3
Características Generales	Potencia Instalada (MW)	+15	+50	+100
	Periodo de información (meses)	36	28	30
	Disponibilidad de información (%)	97	97.5	97.5
	Nº de estaciones de medida	1	2	4

En la Tabla 2, Se indican para cada planta las variables medidas tanto en las estaciones de medida como en el inversor. Los datos se registran con una frecuencia de un minuto, pero se trabaja con medias de diez minutos que es el estándar del sector.

Tabla 2. Resumen de las variables medidas en las plantas

		PV1	PV2	PV3
Medidas en Estación Meteorológica	Irradiancia (IR)	X	X	X
	Temperatura ambiente (T_{amb})	X	X	X
	Temperatura módulo (T_{mod})	X	X	X
	Presión (B)	X	-	-
	Humedad (H)	X	-	-
	Velocidad del viento (Vel)	X	-	-
	Dirección del viento (Dir)	X	-	-
Medidas en Inversor	Potencia de salida (P_{ac})	X	X	X
	Potencia de entrada (P_{dc})	-	X	X

Las variables de trabajo elegidas para configurar los algoritmos son las comunes a todas las plantas del estudio. Estas variables tal y como puede verse en la Tabla 2 son: la medida de irradiancia por estación meteorológica, temperatura ambiente, temperatura de módulo y la potencia a la salida del inversor. La irradiancia proporciona información sobre la cantidad de energía solar que reciben las placas solares. La temperatura ambiente y la temperatura del módulo pueden afectar a la eficiencia de los paneles solares y, por lo tanto, a la producción. La potencia a

la salida del inversor es la variable que se busca caracterizar, es la variable dependiente u objetivo

Metodología de trabajo

En esta sección se describe el proceso de trabajo para la predicción de la producción a la salida de un inversor. El esquema de la Figura 21 representa el resumen del flujo de trabajo que se lleva a cabo.

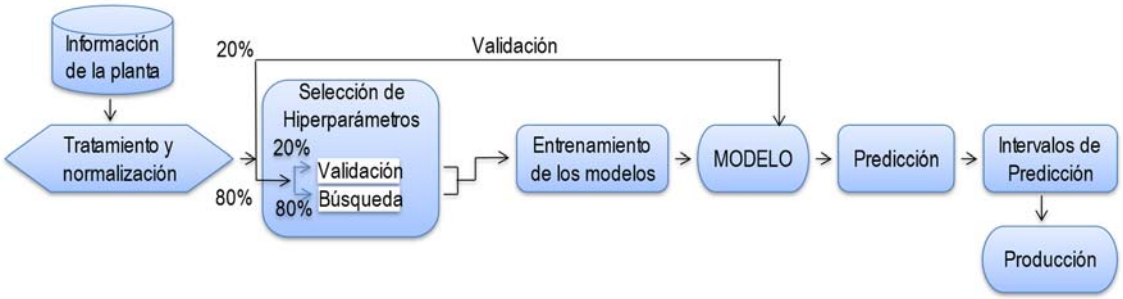


Figura 21. Flujo de trabajo para estimar la generación fotovoltaica

Recepción de la información

En esta primera etapa, se genera la base de datos con la información relevante que permite resolver el problema de predicción de la producción. Estos datos contienen la información histórica de producción, y de las variables de entrada del modelo.

Tratamiento y normalización

La información de las plantas se trata y adecúa al formato requerido para el estudio. Esto puede implicar la conversión de unidades, la corrección de errores o la estandarización de los datos. Además, se debe realizar un filtrado de los valores que estén fuera del rango esperado o que sean claramente errores de medición. Esto es necesario para evitar introducir datos erróneos que puedan afectar al entrenamiento y confundir a los algoritmos utilizados en el estudio.

Una vez eliminados los datos anómalos se normalizan los datos. Los modelos trabajan con datos normalizados para evitar sesgos, y permitir la comparación de las variables en los modelos. La comparación equitativa evita que variables no dominantes tengan un peso desproporcionado y distorsionen el modelo solo por su valor numérico.

La normalización consiste en transformar los valores numéricos de las variables originales a una escala común. Para ello se realiza un cambio de escala de manera que todas las variables se encuentren dentro del rango de 0 a 1, Existen diferentes métodos de normalización, el utilizado en este estudio es la escala min-max de acuerdo con la ecuación (22).

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (22)$$

Siendo X cualquier tipo de variable y X' su correspondiente normalizada.

En el proceso de normalización de variables, se considera la naturaleza y características específicas de cada una. En el caso de medidas comunes a toda la planta como irradiancia y

temperatura ambiente y de módulo, los valores máximos y mínimos son determinados conjuntamente de todas las estaciones. Sin embargo, las producciones de los inversores se normalizan individualmente para cada inversor.

El periodo completo de medida, en todas las estaciones y plantas se desordena y se divide aleatoriamente en dos partes en la proporción 80/20. El primer conjunto, de mayor longitud, se utiliza como el período de entrenamiento de los modelos. El segundo conjunto se usa como el período de validación de los modelos.

Desordenar los datos y la división aleatoria, ayudan a conseguir una mezcla más homogénea de los valores de las variables, y esto redunda en la mejora de los modelos porque permite destacar la influencia de las autocorrelaciones de las variables; estableciendo relaciones más estables entre ellas. La producción de una planta o inversor fotovoltaico está íntimamente relacionada con la irradiancia que se recibe del sol, y por tanto tiene un comportamiento, en parte, altamente predecible.

Exclusivamente para la búsqueda de hiperparámetros, el periodo de entrenamiento se subdivide de nuevo aleatoriamente en dos partes, 80/20. La parte más extensa de esta nueva subdivisión se utiliza para entrenar los algoritmos de métodos de búsqueda de hiperparámetros y el 20% restante para validar la selección de hiperparámetros.

Selección de hiperparámetros

Los hiperparámetros se seleccionan aplicando métodos de búsqueda sistemática en configuraciones o conjuntos predefinidos que como se ha explicado en la Sección 2.5, se eligen antes del entrenamiento del modelo y afectan a su rendimiento. Estos son por ejemplo el número y tamaño de árboles, y el número de capas ocultas en una red neuronal, entre otros.

Entrenamiento del modelo

En esta etapa, se utiliza el conjunto de datos de entrenamiento preparado anteriormente para ajustar los parámetros del modelo. El objetivo es encontrar los valores de los parámetros que minimicen una función de pérdida, que mide la discrepancia entre las predicciones del modelo y los valores reales. El entrenamiento del modelo generalmente implica el uso de algoritmos específicos de optimización, para ajustar los parámetros de manera iterativa.

Predicción de la producción

Una vez que el modelo está definido, se realizan predicciones de la producción. Los datos de prueba o validación alimentan al modelo, que genera predicciones basadas en los patrones aprendidos durante el entrenamiento.

Intervalos de predicción

El resultado de la predicción se complementa con el cálculo de su intervalo de confianza. Tal y como se indica en la Sección 2.6, el primer paso es verificar si la producción fotovoltaica sigue una distribución Normal [197]. Para ello se realiza la gráfica de su histograma y el gráfico de cuantiles “Q-Q plot” (ver Figura 22).

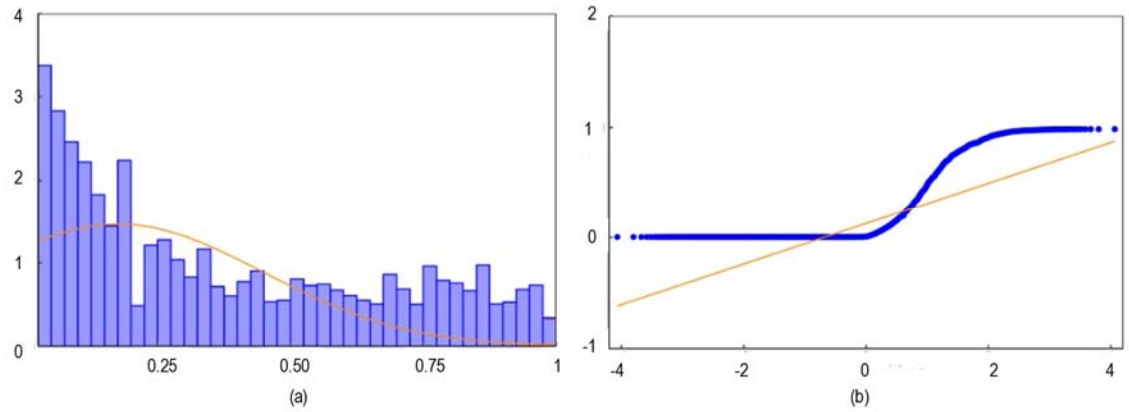


Figura 22. Histograma de la producción (a) y grafica de cuantiles (b) de la producción normalizada

En la Figura 22 (a), se representa la variable producción normalizada y el ajuste a una normal, la forma de la gráfica y la diferencia entre la distribución y el ajuste denota que no hay un comportamiento Gaussiano en esta variable. La gráfica de cuantiles, (ver Figura 22 (b)), lo corrobora ya que tampoco se ajusta a una línea recta (anaranjada) que sería lo esperado si fuera una distribución Normal. Por tanto, el cálculo del intervalo de confianza de los diferentes modelos tendrá que estimarse mediante otras metodologías.

Cálculo de la producción y análisis de desviaciones

Una vez descrita la metodología que permite definir los modelos, se realiza una evaluación de los mismos. La evaluación implica comparar las predicciones del modelo con los valores reales de producción y calcular métricas.

Todos los algoritmos se han programado en R, "R programming language" [228]. R es un software abierto que dispone de paquetes de librerías entre las cuales se incluyen las herramientas que se usan en las técnicas de "Machine Learning".

3.3. Aplicación de los algoritmos a las plantas fotovoltaicas

En esta sección se aplican los algoritmos a las diferentes plantas y se presentan los resultados de la predicción de la producción con los modelos resultantes y su intervalo de confianza. Además, en los casos en los que es necesaria la optimización de hiperparámetros se presentan también el proceso y los resultados de la selección de estos.

Regresión lineal multivariable

El algoritmo MLR [214] no tiene hiperparámetros. Las variables independientes son la irradiancia, la temperatura ambiente y la temperatura de módulo, y el número de estas depende de las estaciones de medición que tiene cada planta. La expresión matemática mediante la cual el modelo calcula la producción del inversor de las plantas fotovoltaicas se presenta en la ecuación

(23), que es una particularización de la ecuación (1) Sección 2.2. En esta ecuación (i) es el identificador de la planta, y recorre los valores de 1 a 3 según la planta de trabajo, y (j) el número de estaciones meteorológicas de la planta.

$$y_i = \alpha_{io} + \sum_j \beta_{ij} \cdot IR_{ij} + \sum_j \gamma_{ij} \cdot T_{amb_{ij}} + \sum_j \delta_{ij} \cdot T_{mod_{ij}} \qquad \textbf{(23)}$$

El ajuste de los coeficientes α, β, γ y δ se realiza mediante el método de mínimos cuadrados. Los valores de estos coeficientes obtenidos ajustando el modelo para cada planta se presentan en la Tabla 3.

Tabla 3. Parámetros del modelo MLR. Aplicación generación PV

		PV1	PV2	PV3
Término independiente (α_{io})		0.018	-0.009	-0.012
Coeficiente de la Irradiancia (β_{ij})	Estación 1	1.196	0.926	0.144
	Estación 2	-	0.132	1.170
	Estación 3	-	-	0. 139
	Estación 4	-	-	-0. 204
Coeficiente de la Temperatura ambiente (γ_{ij})	Estación 1	0.083	-0.466	0. 184
	Estación 2	-	0.340	-0. 007
	Estación 3	-	-	-0. 204
	Estación 4	-	-	0. 021
Coeficiente de la Temperatura de módulo (δ_{ij})	Estación 1	-0.181	0.284	-0. 262
	Estación 2	-	-0.063	0. 007
	Estación 3	-	-	0. 298
	Estación 4	-	-	0. 021

Analizando los valores de estos coeficientes se puede ver como la irradiancia es la variable más significativa. También se observa que en algunos casos el incluir más medidas de otras estaciones puede ser redundante. Sin embargo, al analizar los resultados de los modelos de ML se observa que son fundamentales para mejorar la precisión.

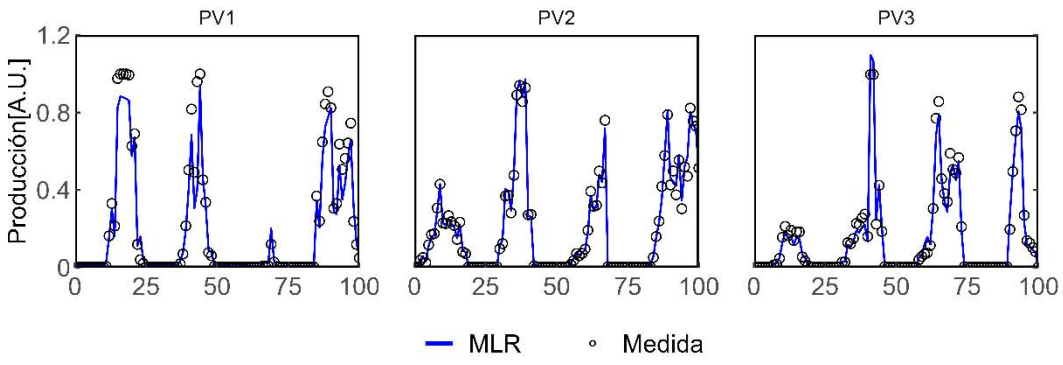


Figura 23. Muestra del resultado del modelo MLR. Aplicación generación PV

Una vez obtenidos los coeficientes del algoritmo, se pueden aplicar a los periodos de prueba en cada planta. Una muestra de los resultados de este modelo para cada una de las plantas se puede ver en la Figura 23, La muestra seleccionada no es una representación temporal de datos

correlativos, sino una muestra aleatoria de valores no consecutivos.

La relación entre las mediciones de producción en el periodo de prueba y el resultado del modelo se muestra en la Figura 24. En esta figura además de asociar a cada valor medido el valor estimado con el modelo MLR de la producción, se representa la función identidad y la línea de tendencia.

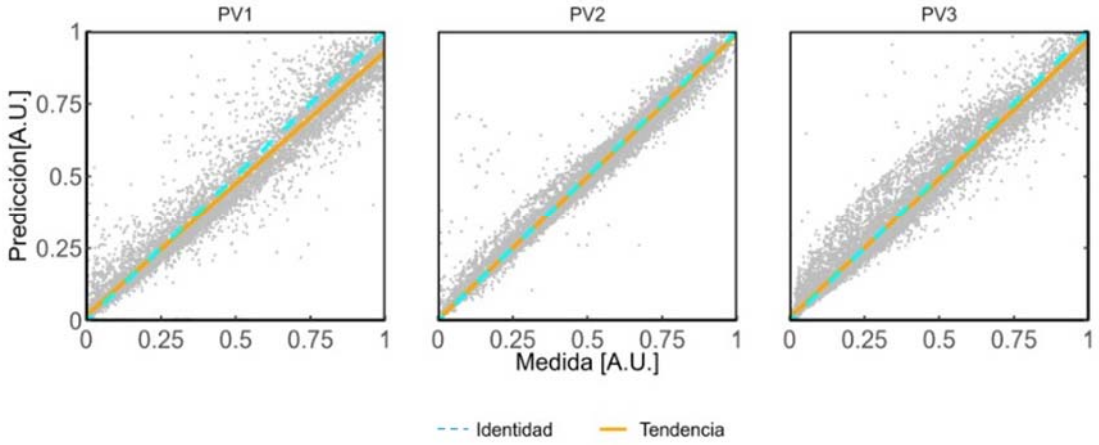


Figura 24. Gráfica de dispersión del resultado del modelo MLR. Aplicación generación PV

Analizando la Figura 24, se observa que en la planta PV1, el modelo tiene menor precisión, porque hay mayor divergencia entre las rectas de identidad y tendencia del resultado con respecto a los modelos de las otras dos plantas. El modelo generado para la planta PV2, además de tener una buena precisión, tiene la menor incertidumbre porque presenta menor dispersión en su nube de datos. El modelo de la planta PV3 tiene también una buena precisión porque la línea de identidad es coincidente con la tendencia del modelo, sin embargo, su incertidumbre es superior a la que presenta el modelo aplicado en la planta PV2 al presentar mayor dispersión.

Para calcular los intervalos de confianza se utiliza el algoritmo “Quantile Regression” (QR) [224], que permite calcular los cuantiles del resultado. Los niveles de confianza que se evalúan en los intervalos de predicción son los del 90%, 95% y 99%.

En la Figura 25, se presenta la muestra de los resultados de la Figura 23 incluyendo los intervalos para los tres niveles de confianza estimados. En esta muestra se puede ver como son las amplitudes de los intervalos y cómo se comportan con respecto a la estimación del modelo y a las medidas reales.

La representación del comportamiento de los intervalos en el total de la muestra se representa en la Figura 26. En esta figura se indican la relación entre las mediciones de producción y el resultado del modelo tal y como se mostraban en la Figura 24, y además incluyen, la tendencia de la relación entre las mediciones y el resultado de cada uno de los intervalos.

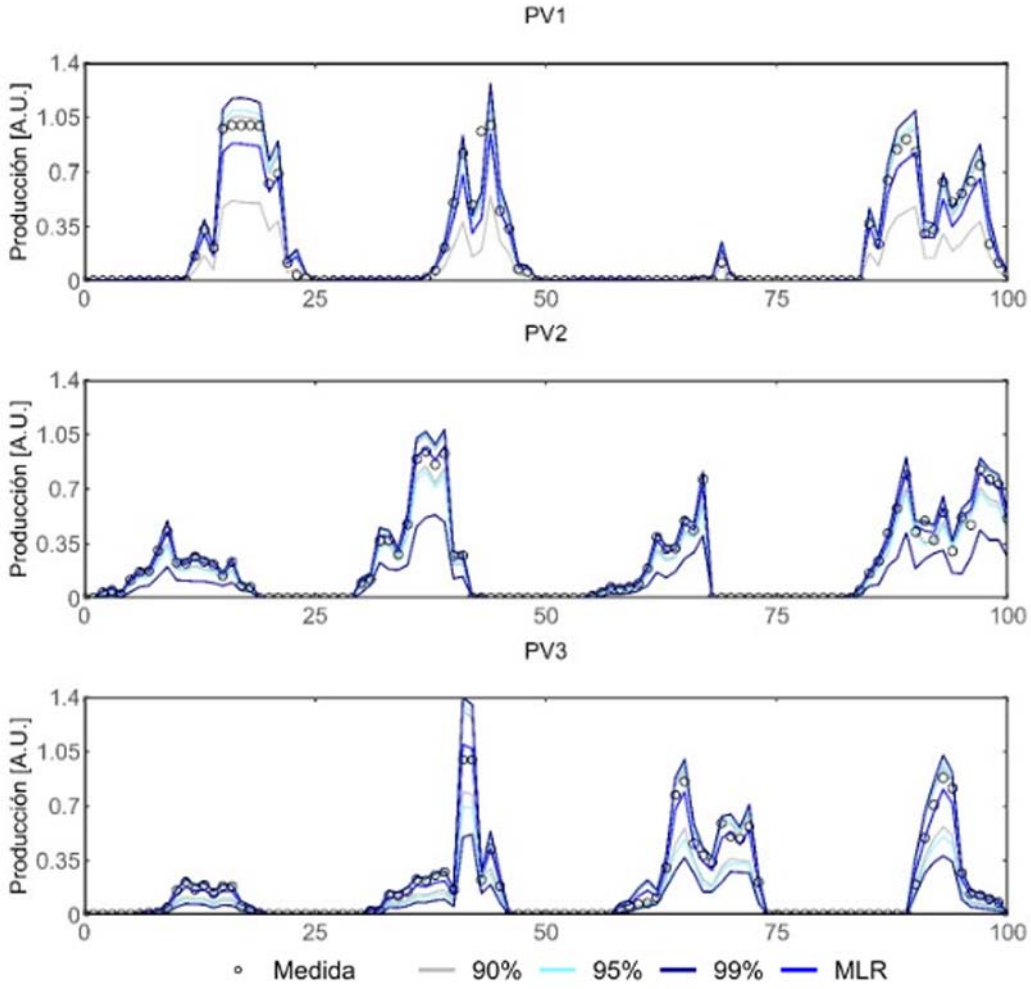


Figura 25.Muestra del intervalo de confianza del modelo MLR. Aplicación generación PV

En la Figura 26 se observa la asimetría de los intervalos, los límites superiores están más próximos a la tendencia del resultado del modelo, debido a que la producción está limitada por la irradiancia máxima teórica que llega a la tierra. Sin embargo, los limites inferiores de los intervalos son más amplios debido a que la nubosidad parcial hace que exista dispersión en los resultados del modelo. Este hecho se presenta más acusado en el ejemplo de la planta PV1, de tal manera que para los niveles de confianza superiores al 90% en la planta PV1, este límite inferior es una valor constante e igual a cero, incluyendo así todo el rango de posibles valores hasta la parada del inversor.

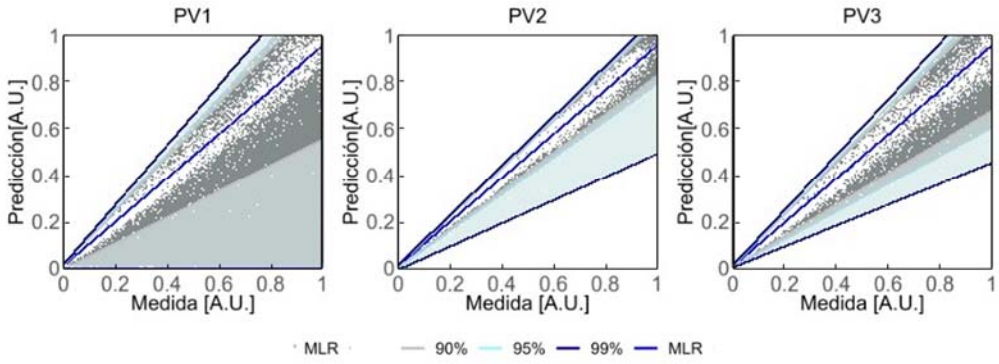


Figura 26. Resultado del intervalo de confianza del modelo MLR. Aplicación generación PV

También se observa como en la planta PV2 la amplitud del intervalo es menor ya que sus líneas de tendencia se cierran sobre la tendencia del modelo. La planta PV3, tiene una amplitud intermedia y la mayor amplitud la tiene la planta PV1. Estos resultados son coherentes ya que el intervalo de confianza es una medida de la incertidumbre y por tanto la amplitud del intervalo es un reflejo de las dispersiones de datos ya comentadas en el análisis del resultado del modelo

"Random Forest"

En el algoritmo Random Forest [219] hay que seleccionar el valor de los hiperparámetros que definen el modelo. Estos hiperparámetros son el número de árboles (Nº Árboles), la profundidad de cada árbol "Depth" (Profundidad), y la tasa de muestreo de datos "Sample Rate", (T. Muestreo) junto con el número de variables (Nº Variables) que intervienen en cada árbol. La combinación óptima de estos cuatro valores permitirá alcanzar una buena precisión. A continuación, se explica la influencia de cada uno de ellos en el algoritmo.

Ajustar el número de árboles es necesario para lograr la convergencia del error, asegurando así la paciencia del modelo o la estabilidad en el error mínimo. El aumento del número de árboles conlleva una disminución en el error del modelo y un aumento del tiempo de computación, pero a partir de un cierto número de árboles la disminución del error tiene un comportamiento asintótico y el tiempo de computación sigue aumentando. Por este motivo es conveniente determinar el número de árboles a partir del cual no existe una disminución significativa del error.

La profundidad define el tamaño, la complejidad de cada uno de los árboles, y permite controlar el sobre-entrenamiento del modelo. A medida que se trabaja con mayor profundidad se consigue una mejor precisión, pero a la vez se corre el riesgo de una sobrecarga computacional y de incurrir en sobreajuste. Por este motivo el ajuste de este hiperparámetro es uno de los más críticos.

La tasa de datos en la muestra y el número de variables que se utilizan en cada árbol individual, influyen en menor medida sobre el error del modelo, pero garantizan que no exista correlación entre los diferentes árboles lo que mejora el rendimiento del modelo, evita el sobreajuste y disminuye el tiempo de cálculo.

Dado que los modelos "Random Forest" requieren una alta potencia computacional se realiza la búsqueda de hiperparámetros con el método sistemático de búsqueda en red [223] pero de forma iterativa, en dos pasos (ver Sección 2.5). En el primer barrido, se utiliza una red extensa con un mallado grueso de valores de hiperparámetros (ver Tabla 4). Una vez identificadas las regiones de interés en el primer barrido, se realiza un segundo barrido con un mallado más fino.

Tabla 4. Rango de valores de los hiperparámetros y número de modelos. Barrido 1 (RF). Aplicación generación PV

	PV1	PV2	PV3
T. Muestreo	0.1-0.9, delta 0.1	0.1-0.9, delta 0.1	0.1-0.9, delta 0.1
Profundidad	5-50, delta 5	5-50, delta 5	5-50, delta 5
Nº Variables	2-3	2-6	2-12
Nº Árboles	10-1500, delta 10	10-1500, delta 10	10-1500, delta 10
Nº Modelos	27000	67500	148500

Como puede verse en la Tabla 4, el número de modelos difiere según la planta, porque el número

de variables en cada planta varía. La planta PV1 tiene una estación meteorológica, hay por tanto un máximo de tres variables, la planta PV2 tiene dos estaciones, por lo que hay un máximo de seis variables y la planta PV3 tiene cuatro estaciones por lo que el algoritmo puede tener en cuenta hasta doce variables. El valor mínimo de variables para tener un árbol multivariable es 2. El resto de hiperparámetros tienen el mismo rango de variación en las tres plantas.

La representación gráfica de todos los modelos del primer barrido (243000) se muestra en la Figura 27. Para poder discriminar los resultados según los diferentes valores de los hiperparámetros la figura se divide en 3 secciones, de forma que cada sección se representa en una fila y corresponde a los resultados de una planta. Para cada planta se realizan cinco gráficas, y en cada una se representa una tasa de muestreo diferente, dando lugar a quince gráficas diferentes. En cada una de estas quince gráficas, se representa el error de cada modelo en el eje de ordenadas en función del número de árboles, en el eje de abscisas. Los valores de las diferentes profundidades de los árboles se indican con diferentes colores, y el número de variables que intervienen en los árboles con diferente símbolo.

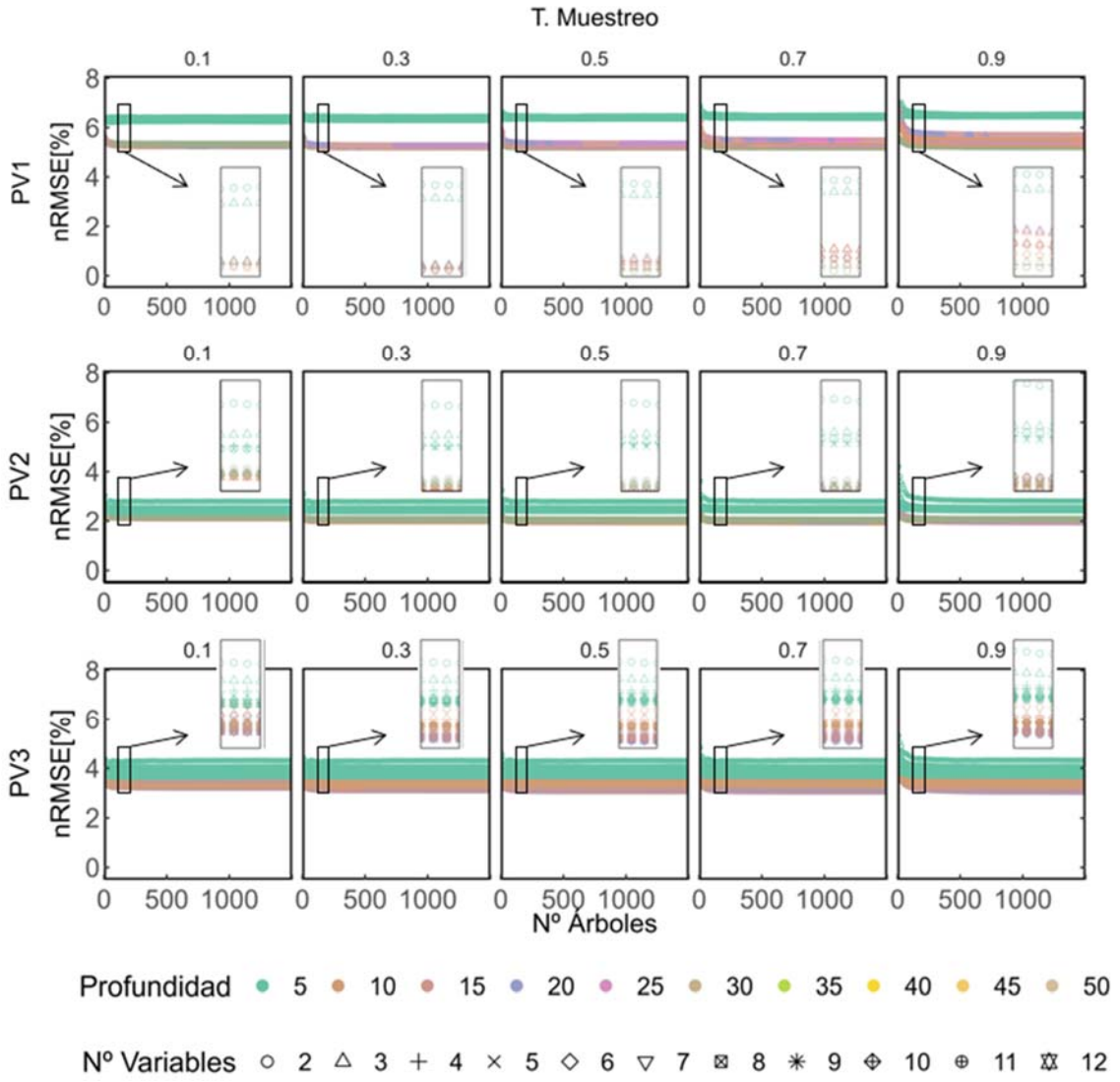


Figura 27. Representación del error de todos los modelos generados en la búsqueda. Barrido 1. (RF). Aplicación generación PV

Se observa una superposición de modelos en algunas gráficas de la Figura 27 debido a que varias combinaciones de hiperparámetros presentan errores muy similares. Al existir más de una alternativa que minimiza el error se considerará óptimo el modelo que presente la combinación de hiperparámetros más sencilla.

Analizando cada una de las plantas por separado, se aprecia, una variación del error según los hiperparámetros considerados, así al aumentar el tamaño de los árboles y el número de los mismos en cada una de las plantas se aprecia una disminución del error hasta un determinado valor, a partir de éste el error satura no consiguiéndose una mejora del mismo al aumentar los valores de los hiperparámetros. También se observa que la tasa de muestreo y el número de variables que intervienen en cada árbol tienen una menor influencia sobre el error.

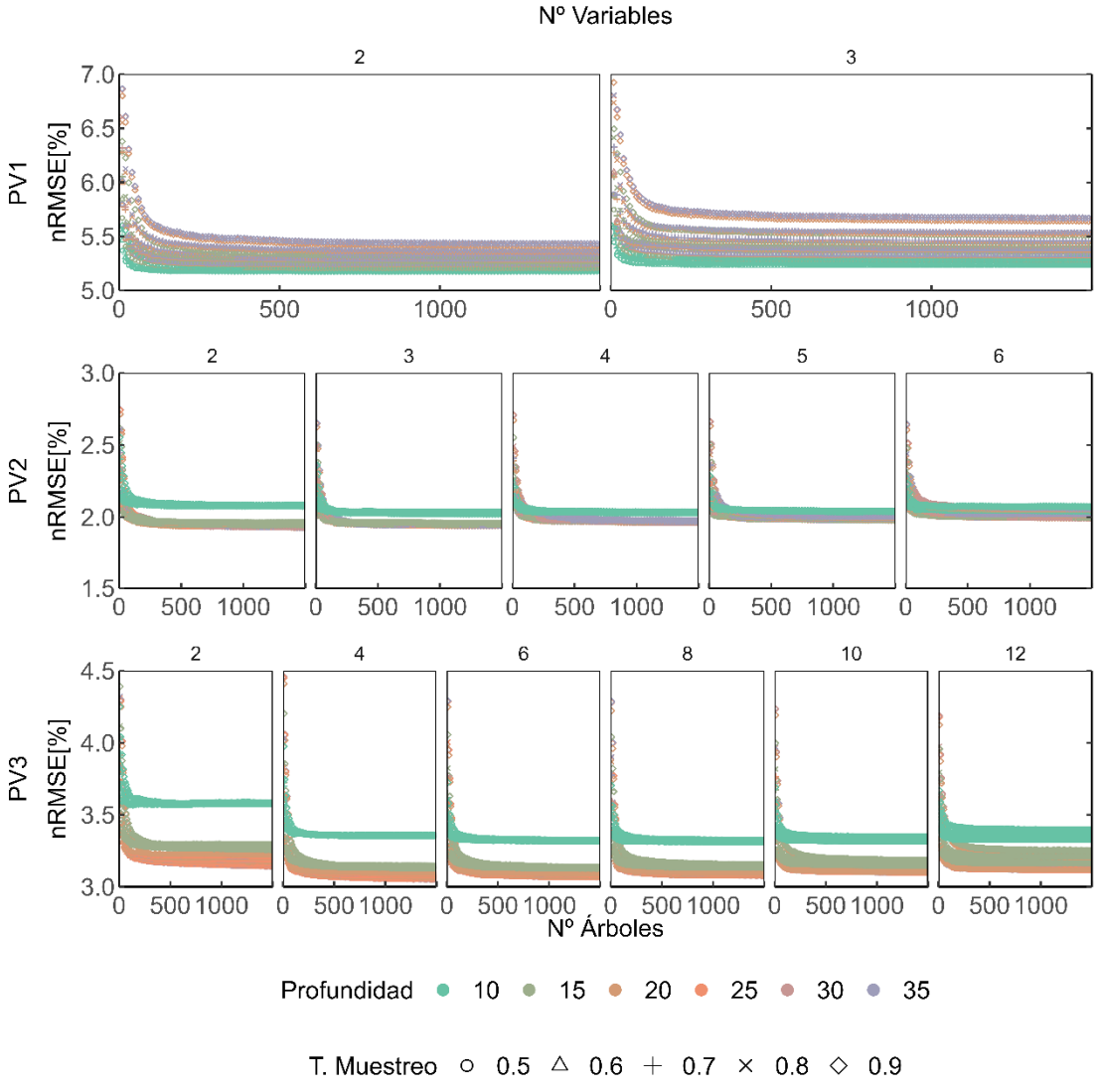


Figura 28. Detalle del error de los modelos generados en la búsqueda. Barrido1 (RF, Profundidad entre 10 y 35 y T. Muestreo mayor de 0.5). Aplicación generación PV

El número recomendado de variables para conformar los árboles individuales en modelos regresivos es de un tercio del número de variables que intervienen en el modelo. Para verificar, si esta premisa se cumple en las plantas de estudio se selecciona una muestra de modelos, en concreto los que tienen una tasa de muestreo superior a 0,5 y una profundidad entre 10 y 35, ya

que según la Figura 27, valores inferiores de tasa de muestreo y de profundidad tienen errores más altos y las profundidades por encima del valor de 35 no aportan una disminución del error. El error de estos modelos frente al número de árboles se muestra en diferentes gráficas que componen la Figura 28. Las gráficas están ordenadas por filas, y cada fila corresponde a una planta. Así, en la primera fila que corresponde a la planta PV1 hay dos gráficas una para cada posible cantidad de variables que pueden conformar los árboles. En la segunda fila dedicada a la planta PV2 hay cinco graficas porque el número de variables con las que se pueden conformar los árboles varía entre 2 y 6, y en la tercera fila que corresponde a la planta PV3, y tiene hasta 12 variables, se representan solo las opciones pares dando lugar a seis gráficas. En cada una de estas gráficas los valores de la profundidad de los árboles se muestran con diferente color, y los de la tasa de muestreo con diferentes símbolos.

En la gráfica de la Figura 28 se puede observar cómo el número de variables a considerar en cada árbol que minimizan el error para las plantas PV1 y PV2, es 2 y para la planta PV3 el valor es de 4, valores que coinciden con la aproximación propuesta de un tercio de las variables del estudio. Además, en esta figura se corrobora lo que se observaba en la Figura 27*,* no se aprecian grandes cambios en el error con la variación de la tasa de muestreo, mientras que la profundidad del árbol es un hiperparámetro muy influyente.

Para identificar si un modelo presenta un sobre-ajuste se comparan los errores de los modelos obtenidos en el entrenamiento y al aplicarlos a la muestra de validación. Si hay una diferencia sustancial entre ellos existe un sobre-entrenamiento. Esto se muestra en la Figura 29, donde se representa a través de quince gráficas ordenadas en tres filas y cinco columnas. En cada fila se representan los modelos que corresponden a cada planta y en cada columna los modelos que corresponden a cada tasa de muestreo. En cada una de las gráficas se vuelve a representar el error en función del número de árboles indicando con diferente color los valores de la profundidad y mediante diferentes símbolos el error en cada modelo del entrenamiento o de su validación. En este caso para una mejor visualización se han seleccionado un subconjunto de valores de tasa de muestreo, los valores impares, y un subconjunto de valores de profundidad representativos de todo el rango.

En la Figura 29, se observa que apenas hay diferencia entre los errores del conjunto de datos de entrenamiento y de validación, por lo que no existe sobre-entrenamiento, y por lo tanto el resto de hiperparámetros pueden tomar los valores más sencillos que minimicen el error.

Con las consideraciones realizadas se ha seleccionado el número de variables de los árboles. La tasa de muestreo puede tomar cualquier valor, por lo que es recomendable mantenerse en la zona media alta. La profundidad del árbol es el hiperparámetro más crítico por su influencia en el error y porque su aumento complica el modelo y aumenta los tiempos computacionales. Y el número de árboles no se ha analizado todavía, si bien en todas las figuras anteriores, se observa un comportamiento asintótico a partir de los 200 árboles.

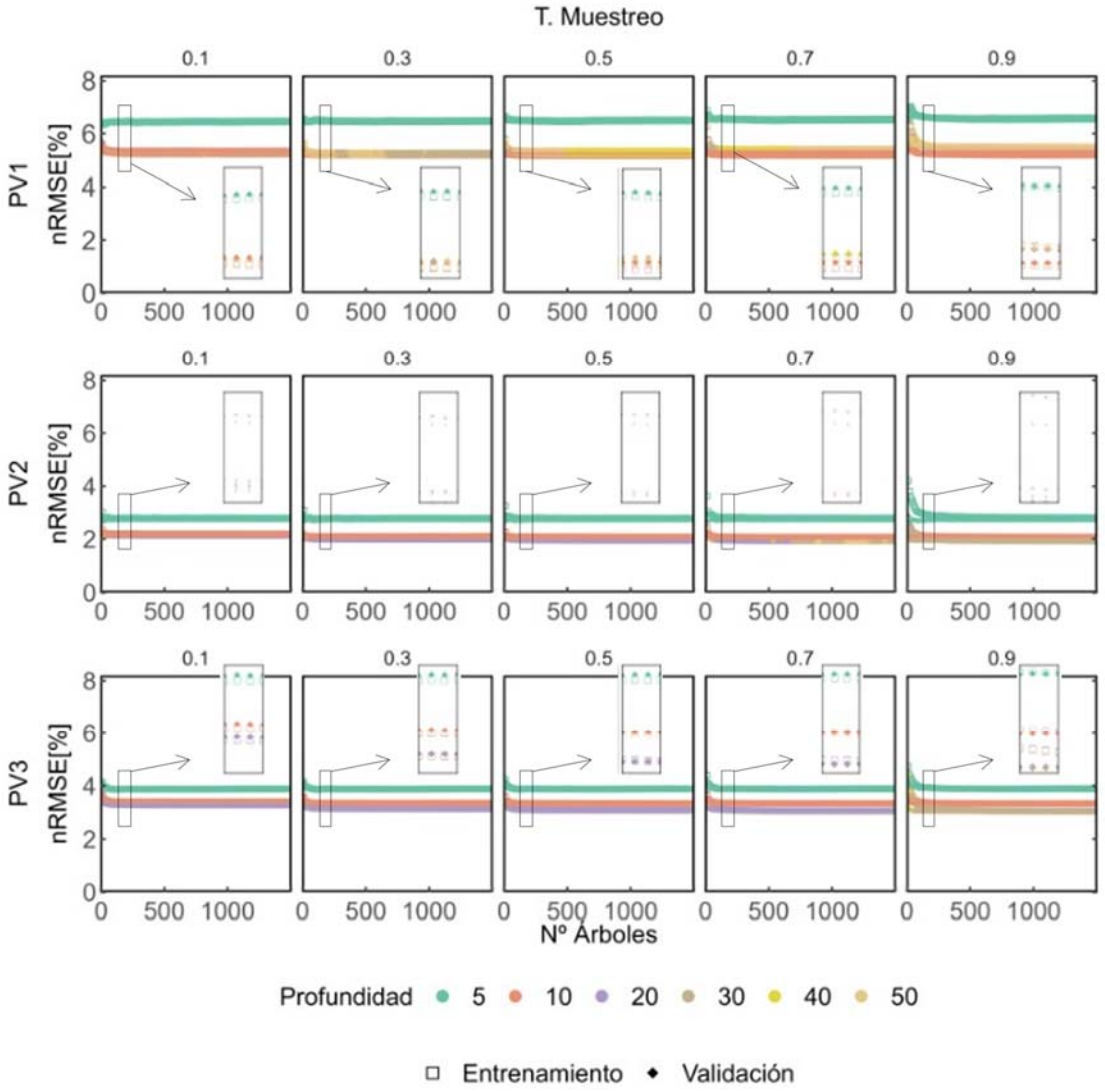


Figura 29. Detalle del error de los modelos generados en la búsqueda. Barrido 1. (RF, Nº Variables: 2 para las plantas PV1 y PV2, y 4 para la planta PV3). Aplicación generación PV

Para afinar el estudio se realiza un segundo barrido de búsqueda de hiperparámetros en red, acotando principalmente la profundidad de los árboles en el caso de la planta PV1, entre los valores menores de 15, en la planta PV2, entre los menores de 25, y en la planta PV3 entre los menores de 30. La tasa de muestreo se limita a los valores medios altos. Los rangos de los hiperparámetros de este segundo barrido se muestran en la Tabla 5.

Tabla 5. Rango de valores de los hiperparámetros y número de modelos. Barrido 2 (RF). Aplicación generación PV

	PV1	PV2	PV3
T. Muestreo	0.6-0.9, delta 0.1 0.97,0.98,0.99	0.8, 0.9, 0.97,0.98,0.99	0.8, 0.9, 0.97,0.98,0.99
Profundidad	5-15, delta 1	15-35, delta 1	25-40, delta 1
Nº Variables	2	2	4
Nº Árboles	10-1500, delta 10	10-1500, delta 10	10-1500, delta 10
Nº Modelos	10500	15000	11250

La representación de todos los modelos del segundo barrido en las tres plantas de estudio se muestra en la Figura 30.

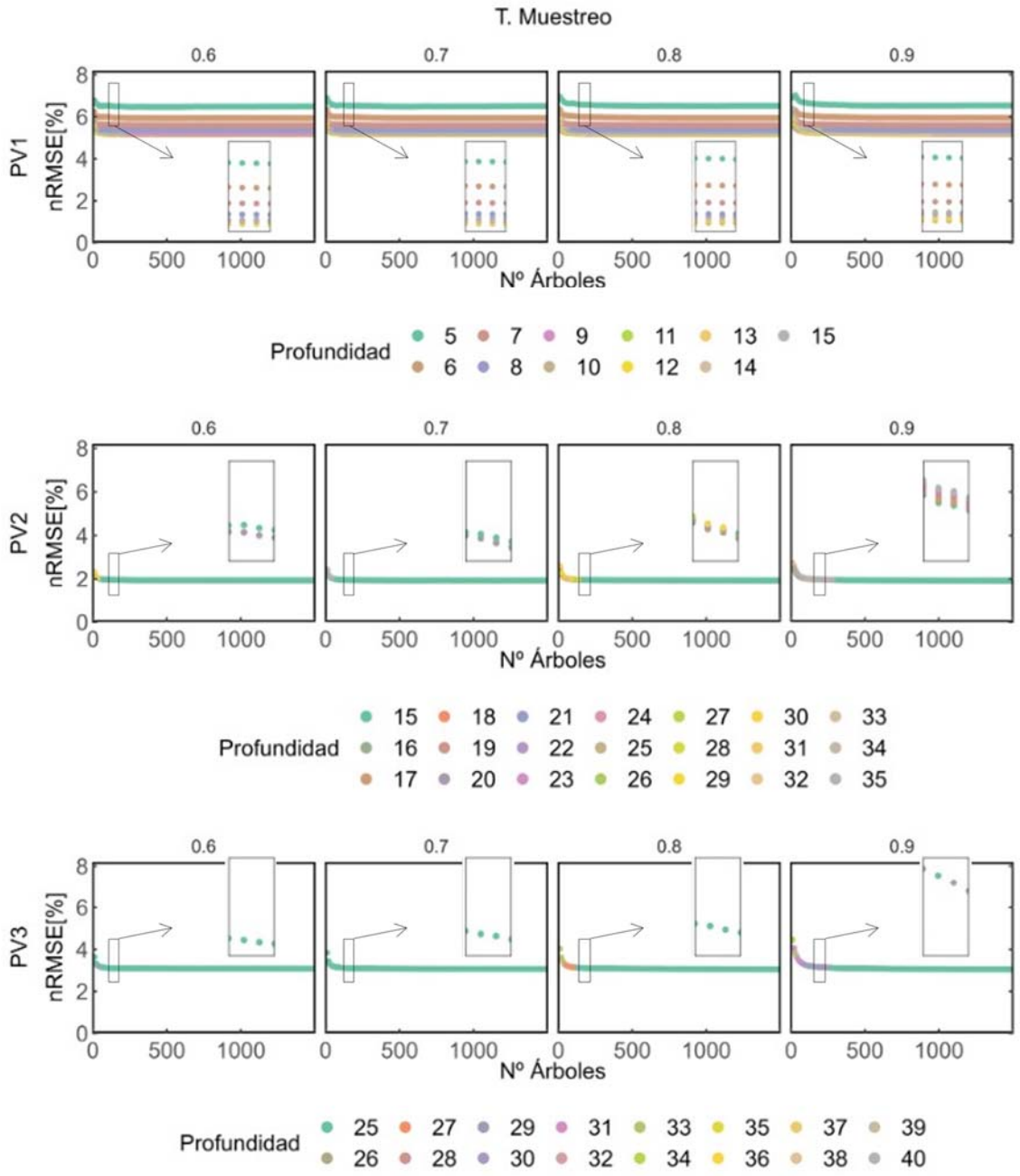


Figura 30. Representación del error de todos los modelos generados en la búsqueda. Barrido 2. (RF). Aplicación generación PV

En esta figura se representan doce gráficas repartidas en tres filas, una para cada planta y cuatro columnas que corresponden a las diferentes tasas de muestreo. En cada gráfica, de manera análoga a las anteriores figuras, se representa el error de los modelos en función del número de árboles e indicando la profundidad de los árboles con diferente color.

De la Figura 30, se observa una vez más la superposición de los valores del error, el dominio de la profundidad y la baja influencia en el error de la tasa de muestreo. Este segundo barrido permite disponer de modelos con hiperparámetros cuyos rangos de valores dan lugar a errores bajos. Para buscar los valores más adecuados se va a considerar las siguientes premisas:

- La profundidad del árbol ha de ser la más pequeña posible, siempre y cuando no suponga un

aumento sustancial del error del modelo.

- La tasa de muestreo debe entrenarse entre los valores medio altos 0.7-0.8 equilibrando el resto de los valores de los hiperparámetros del modelo.
- El número de árboles debe ser lo más pequeño posible siempre y cuando se cumpla el criterio de estabilidad. El número de árboles se considera estable cuando en más de dos modelos consecutivos aumentando el número de árboles, la diferencia del error es inferior a 10^{-3}, ecuación (24). En esa situación se asume que el aumento de árboles conlleva a una mejora no significativa en el modelo.

$$\Delta\, nRMSE < 10^{-3} \qquad \textbf{(24)}$$

En las Tabla 6 a Tabla 8 se presentan los errores normalizados nRMSE del periodo de entrenamiento de los modelos, en los que el número de variables que intervienen en los árboles se ha optimizado, y se han acotado el resto de hiperparámetros a valores razonables: la tasa de muestreo al rango 0.7-0.8, el número de árboles al rango 500-590 y la profundidad se ha particularizado para cada planta.

Tabla 6. Error nRMSE. Selección de hiperparámetros de los modelos RF en PV1. Aplicación generación PV

PV1		Profundidad								
Nº Árboles	T. Muestreo	7	8	9	10	11	12	13	14	15
500	0.7	5.602	5.381	5.244	5.169	5.181	5.185	5.208	5.238	5.262
	0.8	5.618	5.387	5.247	5.174	5.186	5.193	5.224	5.252	5.286
510	0.7	5.603	5.382	5.244	5.168	5.181	5.185	5.208	5.237	5.261
	0.8	5.618	5.387	5.247	5.174	5.186	5.193	5.223	5.251	5.285
520	0.7	5.603	5.382	5.245	5.168	5.181	5.185	5.208	5.237	5.261
	0.8	5.617	5.386	5.246	5.174	5.185	5.192	5.223	5.251	5.285
530	0.7	5.603	5.381	5.244	5.169	5.182	5.185	5.208	5.237	5.261
	0.8	5.616	5.386	5.246	5.174	5.185	5.193	5.223	5.251	5.285
540	0.7	5.603	5.381	5.244	5.169	5.181	5.185	5.208	5.236	5.261
	0.8	5.616	5.385	5.246	5.174	5.185	5.193	5.223	5.252	5.285
550	0.7	5.602	5.380	5.243	5.168	5.181	5.184	5.207	5.235	5.260
	0.8	5.616	5.385	5.245	5.174	5.185	5.193	5.224	5.252	5.285
560	0.7	5.602	5.379	5.243	5.167	5.180	5.184	5.207	5.235	5.259
	0.8	5.615	5.384	5.244	5.173	5.184	5.193	5.224	5.252	5.286
570	0.7	5.601	5.379	5.242	5.167	5.180	5.184	5.206	5.234	5.258
	0.8	5.614	5.384	5.244	5.173	5.184	5.193	5.224	5.252	5.285
580	0.7	5.601	5.378	5.242	5.167	5.180	5.183	5.205	5.233	5.257
	0.8	5.614	5.384	5.245	5.173	5.184	5.193	5.223	5.252	5.284
590	0.7	5.602	5.378	5.242	5.167	5.179	5.183	5.205	5.233	5.257
	0.8	5.613	5.383	5.245	5.174	5.184	5.193	5.223	5.251	5.284

Tabla 7. Error nRMSE. Selección de hiperparámetros de los modelos RF en PV2. Aplicación generación PV

PV2		Profundidad									
Nº Árboles	T. Muestreo	16	18	20	22	24	26	28	30	32	34
500	0.7	1.945	1.941	1.940	1.940	1.941	1.941	1.941	1.941	1.941	1.941
	0.8	1.953	1.949	1.949	1.949	1.949	1.949	1.949	1.949	1.949	1.949
510	0.7	1.945	1.941	1.940	1.941	1.941	1.941	1.941	1.941	1.941	1.941
	0.8	1.954	1.950	1.949	1.949	1.949	1.950	1.950	1.950	1.950	1.950
520	0.7	1.945	1.942	1.939	1.941	1.941	1.941	1.941	1.941	1.941	1.941
	0.8	1.954	1.950	1.949	1.949	1.950	1.950	1.950	1.950	1.950	1.950
530	0.7	1.945	1.942	1.939	1.941	1.941	1.941	1.941	1.941	1.941	1.941
	0.8	1.954	1.950	1.949	1.949	1.949	1.950	1.950	1.950	1.950	1.950
540	0.7	1.945	1.942	1.939	1.941	1.941	1.941	1.941	1.941	1.941	1.941
	0.8	1.954	1.950	1.949	1.949	1.949	1.950	1.950	1.950	1.950	1.950
550	0.7	1.945	1.942	1.939	1.941	1.941	1.941	1.941	1.941	1.941	1.941
	0.8	1.954	1.950	1.949	1.949	1.949	1.950	1.950	1.950	1.950	1.950
560	0.7	1.945	1.941	1.939	1.940	1.940	1.941	1.941	1.941	1.940	1.940
	0.8	1.953	1.949	1.949	1.949	1.949	1.949	1.949	1.949	1.949	1.949
570	0.7	1.944	1.940	1.939	1.940	1.940	1.940	1.940	1.940	1.940	1.940
	0.8	1.953	1.949	1.948	1.948	1.948	1.948	1.948	1.948	1.948	1.948
580	0.7	1.944	1.940	1.939	1.939	1.940	1.940	1.940	1.940	1.940	1.940
	0.8	1.952	1.948	1.948	1.948	1.948	1.948	1.948	1.948	1.948	1.948
590	0.7	1.945	1.941	1.939	1.940	1.940	1.940	1.940	1.940	1.940	1.940
	0.8	1.953	1.949	1.948	1.948	1.948	1.949	1.949	1.949	1.949	1.949

Tabla 8. Error nRMSE. Selección de hiperparámetros de los modelos RF en PV3. Aplicación generación PV

PV3		Profundidad								
Nº Árboles	T. Muestreo	20	25	27	29	31	33	35	37	39
500	0.7	3.081	3.074	3.073	3.073	3.073	3.073	3.073	3.073	3.073
	0.8	3.111	3.104	3.103	3.103	3.103	3.103	3.103	3.103	3.103
510	0.7	3.081	3.073	3.073	3.072	3.072	3.072	3.072	3.072	3.072
	0.8	3.111	3.103	3.103	3.103	3.103	3.103	3.103	3.103	3.103
520	0.7	3.080	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072
	0.8	3.110	3.103	3.103	3.102	3.102	3.102	3.102	3.102	3.102
530	0.7	3.080	3.072	3.072	3.071	3.072	3.071	3.071	3.071	3.071
	0.8	3.110	3.103	3.102	3.102	3.102	3.102	3.102	3.102	3.102
540	0.7	3.080	3.072	3.071	3.071	3.071	3.071	3.071	3.071	3.071
	0.8	3.110	3.103	3.102	3.102	3.102	3.102	3.102	3.102	3.102
550	0.7	3.080	3.071	3.070	3.070	3.070	3.070	3.070	3.070	3.070
	0.8	3.110	3.102	3.102	3.102	3.102	3.102	3.102	3.102	3.102
560	0.7	3.079	3.070	3.069	3.069	3.069	3.069	3.069	3.069	3.069
	0.8	3.110	3.102	3.102	3.102	3.102	3.102	3.102	3.102	3.102
570	0.7	3.079	3.070	3.070	3.069	3.069	3.069	3.069	3.069	3.069
	0.8	3.109	3.102	3.102	3.101	3.101	3.101	3.101	3.101	3.101
580	0.7	3.079	3.069	3.069	3.069	3.069	3.069	3.069	3.069	3.069
	0.8	3.109	3.102	3.101	3.101	3.101	3.101	3.101	3.101	3.101
590	0.7	3.079	3.069	3.069	3.069	3.069	3.069	3.069	3.069	3.069
	0.8	3.109	3.102	3.101	3.101	3.101	3.101	3.101	3.101	3.101

En la Tabla 6, se presentan los errores nRMSE de 180 modelos de la planta PV1. Sobre este conjunto de modelos la combinación de hiperparámetros mejor es la que se conforma con los valores de una profundidad de 10, una tasa de muestreo de 0.7 y 560 árboles.

En la Tabla 7 se presentan los errores nRMSE de 200 posibles modelos con diferentes hiperparámetros para la planta PV2. Dado que el rango de valores de profundidad es amplio, solo se muestran en la tabla los valores pares. El valor la profundidad tiene que ser un compromiso entre el menor posible para bajar la carga computacional y que a la vez minimice el error. Por ese motivo se escoge el valor de 20. La tasa de muestreo que menores errores presenta es 0.7, y en cuanto a la elección del número de árboles se aplica el principio de estabilidad y se selecciona 520 como número adecuado.

En la Tabla 8 se presentan los errores nRMSE de 180 posibles modelos para la planta PV3, en esta planta se muestran solo los valores impares de profundidad, para facilitar la observación y además se incluye el valor de 20 del barrido 1, valor que además es el seleccionado porque los valores obtenidos son, aparentemente, demasiado grandes y requieren una carga computacional muy alta que no compensa la diferencia de error que es prácticamente despreciable. El valor de la tasa de muestreo es de 0.7 como en las plantas anteriores y el número de árboles seleccionado es de 560.

Tabla 9. Hiperparámetros óptimos de cada planta y número de modelos para su obtención (RF). Aplicación generación PV

Hiperparámetros	PV1	PV2	PV3
T. Muestreo	0.7	0.7	0.7
Profundidad	10	20	20
Nº Variables	2	2	4
Nº Árboles	560	520	560
Nº Modelos	37500	82500	159750

En la Tabla 9, se resumen los valores de hiperparámetros seleccionados, junto con el número de modelos que han intervenido en la selección:

Los resultados obtenidos para los errores reflejan las características de la variabilidad de la producción de cada una de las plantas. La planta PV2, que es plana, sin pendientes y dispone de dos estaciones de medida, obtiene los mejores resultados, frente a la PV1 que tiene una topografía más compleja y una única estación de medida. En la planta PV3 se pueden conseguir unos errores intermedios, pero con un modelo más complejo porque tiene más estaciones de medida, lo que ayuda al modelo a tener mejores relaciones, pero a costa de tener árboles más completos y que relacionan más variables.

Una vez definidos los hiperparámetros, se entrenan los modelos para obtener los valores de los parámetros internos y así poder calcular la predicción de la muestra de prueba. Los resultados de la prueba se presentan en la Figura 31 y Figura 32. En la Figura 31 se expone una muestra del conjunto de datos de prueba, mientras que en la Figura 32 se muestra la totalidad del conjunto relacionando el resultado del modelo y la medida.

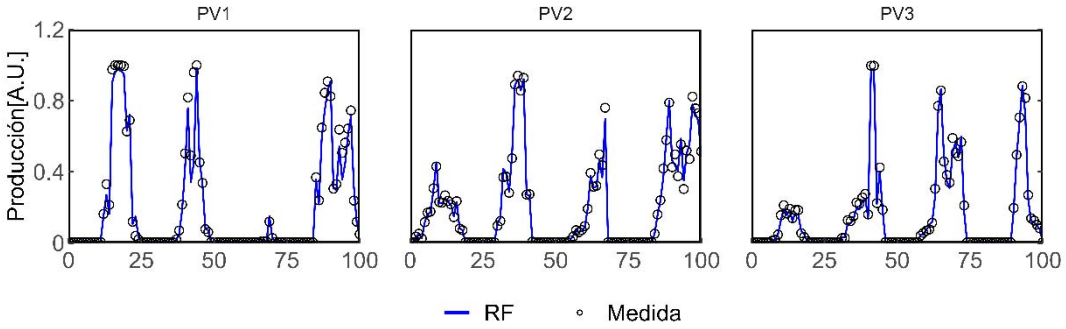


Figura 31. Muestra del resultado del modelo RF. Aplicación generación PV

Tal y como puede verse en la Figura 31, el modelo RF tiene un mejor comportamiento que el modelo MLR. Esto también se observa en la Figura 32 donde junto con la relación de las medidas de producción y el resultado del modelo se presenta la función identidad y la tendencia de la nube de puntos. En las tres plantas puede observarse que la tendencia del modelo es prácticamente la identidad, es decir, el modelo es muy preciso y además la dispersión es muy pequeña, lo que implica que su incertidumbre es baja.

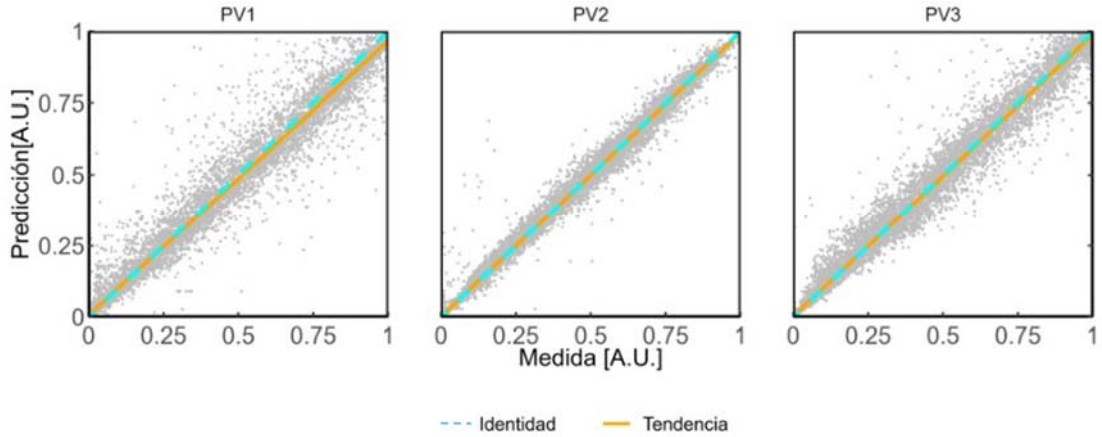


Figura 32. Gráfica de dispersión del resultado del modelo RF. Aplicación generación PV

En la Figura 32 se observa que en general el modelo RF mejora el comportamiento del MLR ya que las líneas de tendencia presentan una mayor similitud con la identidad en las tres plantas. Particularizando a cada una de las tres plantas, el resultado del modelo en la planta PV1 tiene una mayor dispersión, que los resultados de los modelos de las otras plantas, siendo el resultado más preciso el del modelo de la planta PV2.

El intervalo de confianza se estima con el modelo "Quantile Regression Forest" (QRF) [225,226] para evaluar intervalos en base a los percentiles del resultado del modelo RF y con niveles de confianza 90%, 95% y 99%. Una muestra de los valores de los intervalos de confianza resultantes se presenta en la Figura 33, indicando con líneas de diferentes colores, cada uno de los niveles de confianza.

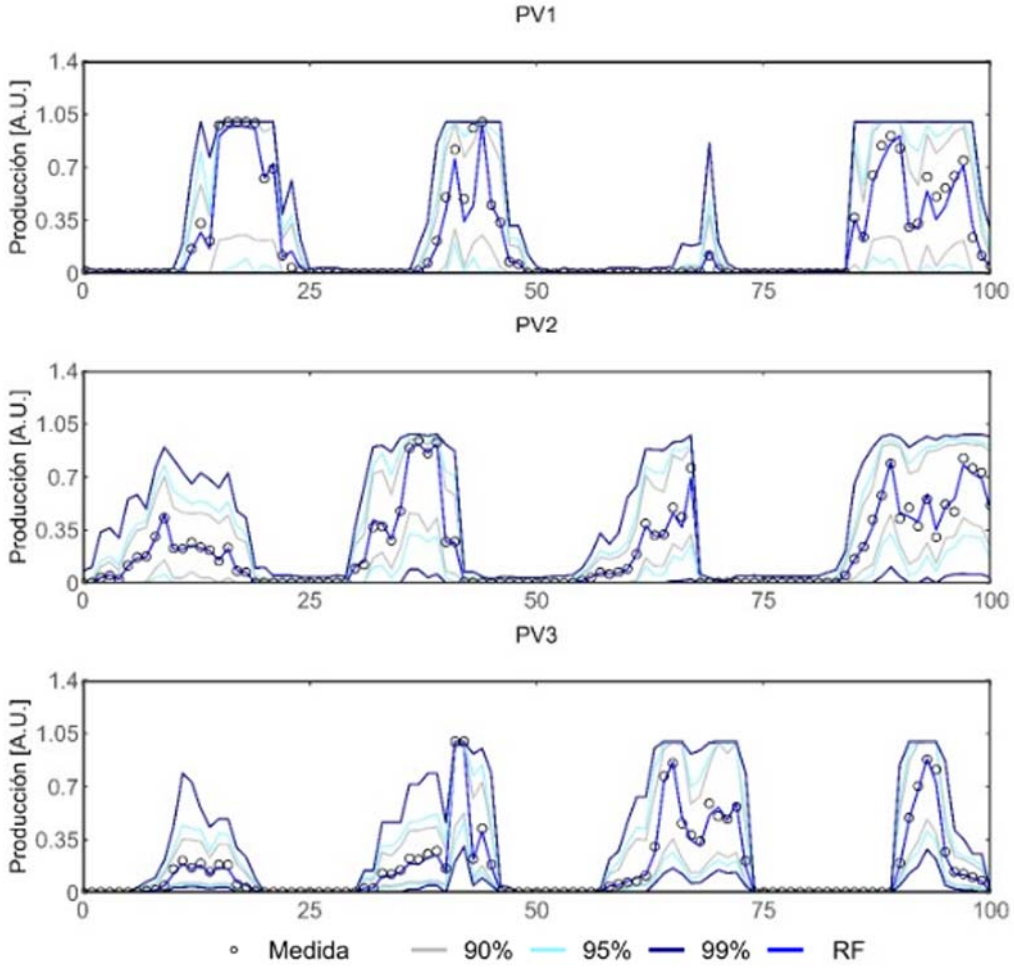


Figura 33. Muestra del intervalo de confianza del modelo QRF. Aplicación generación PV

En la Figura 34 se representan las líneas de tendencia de la relación de la producción medida con la predicción del modelo y con los límites de ese modelo para cada planta. Los limites representados corresponden a los inferiores y superiores de los niveles de confianza de 90%, 95% y 99%. Se incluye también la nube de dispersión de puntos de la relación medida y predicción con el modelo RF de cada planta.

Del análisis de la Figura 34, se observa asimetría en los intervalos, siendo los límites superiores más estrechos respecto al modelo que los inferiores. En la planta PV1, el límite inferior con un 99

% de nivel de confianza presenta una tendencia que es igual al valor mínimo admisible, la producción nula. Esta planta presenta un intervalo de confianza mayor a las otras dos plantas en todos los niveles calculados. Con este modelo la planta PV3 tiene el intervalo de confianza del nivel 90% muy similar en amplitud al de la planta PV2, pero tiene una mayor amplitud en sus intervalos de confianza del 95% y 99%.

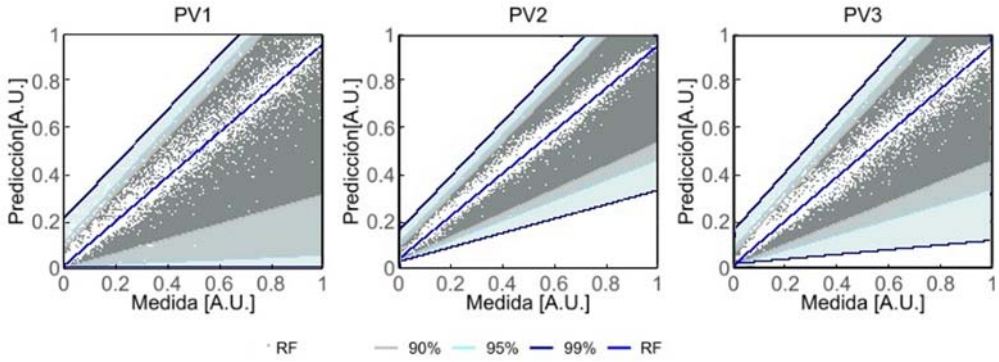


Figura 34. Resultado del intervalo de confianza del modelo QRF. Aplicación generación PV

"Gradient Boosting"

El algoritmo "Gradient Boosting" [220], al ser un método también basado en árboles de decisión tiene hiperparámetros similares a RF que deben seleccionarse antes de ser aplicados en la predicción de la producción de una planta fotovoltaica. La diferencia entre los hiperparámetros de un algoritmo RF y un algoritmo GB es que este último no necesita de-correlacionar los árboles por lo que siempre se trabaja con todas las variables, y por tanto sus hiperparámetros son solo la profundidad (Profundidad) del árbol que define su tamaño, el número de árboles (Nº Árboles) y la tasa de muestreo (T. Muestreo).

Al trabajar solo con tres hiperparámetros puede aplicarse el método de búsqueda "Grid Search", en una única iteración cubriendo así toda la casuística de combinaciones con un mallado amplio y con un paso pequeño. El mallado propuesto, Tabla 10, es idéntico en todas las plantas y genera 90000 modelos en cada una de ellas.

Tabla 10. Rango de valores de los hiperparámetros y número de modelos (GB). Aplicación generación PV

	PV1	PV2	PV3
T. Muestreo	0.1-1, delta 0.1	0.1-1, delta 0.1	0.1-1, delta 0.1
Profundidad	1-15, delta 1	1-15, delta 1	1-15, delta 1
Nº Árboles	5-3000, delta 5	5-3000, delta 5	5-3000, delta 5
Nº Modelos	90000	90000	90000

La representación gráfica de los errores nRMSE de los 270000 modelos obtenidos de todas las combinaciones de hiperparámetros se presentan en la Figura 35 distribuidos en seis gráficas diferentes. Este método es sensible al sobre-entrenamiento, por lo que se verifica si existen combinaciones de hiperparámetros que producen un sobreajuste. Para ello se comparan en la misma figura los errores de todos los modelos generados para el periodo de entrenamiento y el periodo de validación. En la primera fila de la figura se presentan en tres gráficas, una para cada planta, los errores de los modelos obtenidos con los datos de entrenamiento mientras que las tres

gráficas de la segunda fila corresponden a los errores de los modelos al aplicarlos al periodo de validación. En cada gráfica, se representa el error de los modelos, nRMSE, en función del número de árboles indicando con diferente color la variación de la profundidad y con diferente símbolo la variación de la tasa de muestreo.

En la Figura 35 se observa que existe sobre-entrenamiento cuando se incrementa la profundidad del árbol. El error de los modelos en el entrenamiento disminuye de manera más significativa que cuando se aplican los mismos modelos al periodo de validación. Al aumentar el tamaño del árbol, el modelo es más complejo teniendo capacidad de aprender patrones más detallados, pero adquiriendo una sensibilidad excesiva a las características específicas de los datos y, por lo tanto, el modelo se sobre-ajusta. La Figura 35 muestra también la importancia de optimizar los hiperparámetros en este algoritmo, ya que según la elección de estos el error del modelo es muy superior. También en este caso existen modelos superpuestos, que presentan un error similar, por lo que la selección de valores de los hiperparámetros se llevará a cabo con un triple criterio: minimizando el error, evitando sobre-entrenamiento y además buscando las combinaciones de hiperparámetros más sencillas.

En la Figura 35 se observa que existe sobre-entrenamiento cuando se incrementa la profundidad del árbol. El error de los modelos en el entrenamiento disminuye de manera más significativa que cuando se aplican los mismos modelos al periodo de validación. Al aumentar el tamaño del árbol, el modelo es más complejo teniendo capacidad de aprender patrones más detallados, pero adquiriendo una sensibilidad excesiva a las características específicas de los datos y, por lo tanto, el modelo se sobre-ajusta. La Figura 35 muestra también la importancia de optimizar los hiperparámetros en este algoritmo, ya que según la elección de estos el error del modelo es muy superior. También en este caso existen modelos superpuestos, que presentan un error similar, por lo que la selección de valores de los hiperparámetros se llevará a cabo con un triple criterio: minimizando el error, evitando sobre-entrenamiento y además buscando las combinaciones de hiperparámetros más sencillas.

Para observar el comportamiento del error según varía la profundidad del árbol, se presenta la Figura 36, formada por tres gráficas una para cada planta. En esta figura se representan el error de los resultados tanto con los datos de entrenamiento como de su aplicación al periodo de validación. Para que la visualización sea mejor se representan sólo el conjunto de modelos con valor de tasa de muestreo de 0.5 y valores de profundidad entre 1 y 5. No se toman valores más altos de profundidad, porque la Figura 35 indicaba, que si se superaban estos valores había sobre-entrenamiento del modelo. En cada una de las gráficas se muestra el error de los modelos frente al número de árboles, con diferente color se indican los valores de profundidad considerados en cada modelo y con diferente símbolo si se trata de un error de los modelos con los datos de entrenamiento o de la aplicación al periodo de validación.

En la Figura 36, se puede observar en detalle el sobre-entrenamiento de los modelos. El criterio adoptado para evitar el sobreajuste es que la diferencia entre el error nRMSE del periodo de entrenamiento y del periodo de validación sea inferior a $5 \cdot 10^{-2}$. Con este criterio y buscando minimizar el error, la profundidad recomendable es 2 para la planta PV1, 4 para la planta PV2, y 3 para la planta PV3.

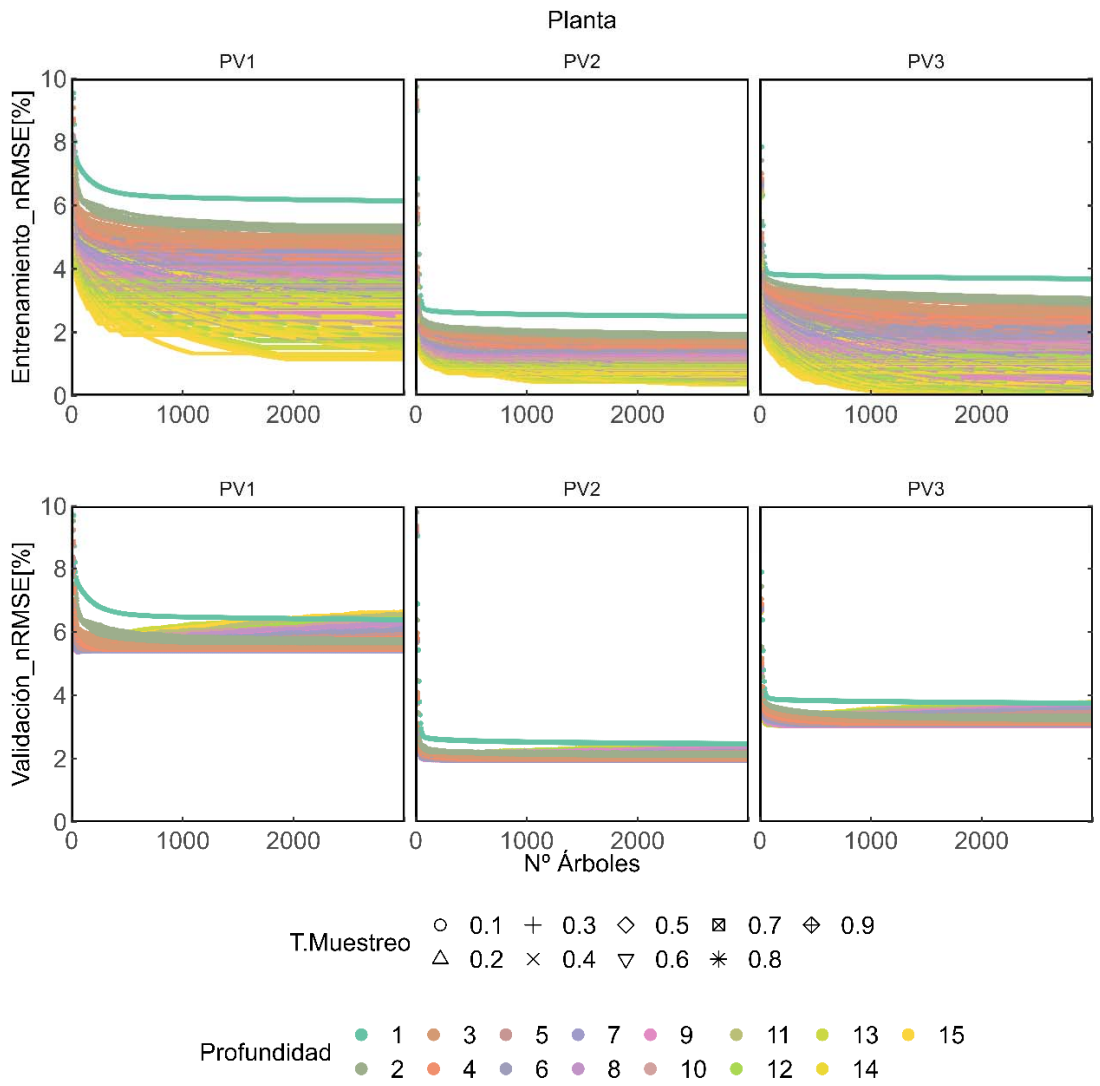


Figura 35. Representación del error de los modelos generados en la búsqueda. (GB). Aplicación generación PV

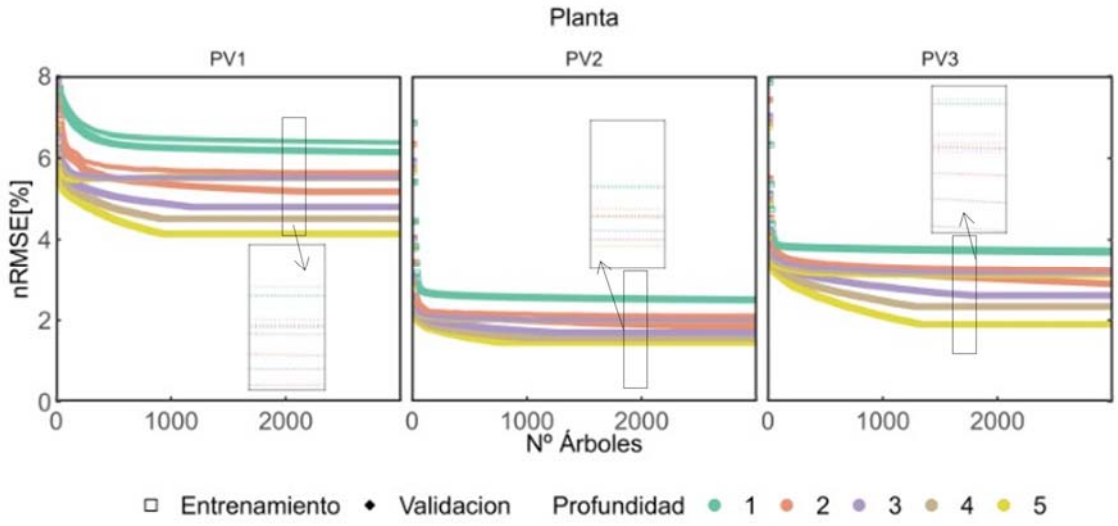


Figura 36. Detalle del error de los modelos generados en la búsqueda. (GB, T. Muestreo=0.5, Profundidad entre 1 y 5). Aplicación generación PV

Para observar la influencia de la tasa de muestreo sobre el error se seleccionan un conjunto de

modelos para cada planta, cada uno con su profundidad correspondiente ya seleccionada, y se varía la tasa de muestreo entre los valores 0.1, 0.5, y 0.9, cubriendo el rango completo de este hiperparámetro. Los errores de los resultados de estos modelos se representan en la Figura 37, donde una vez más se representa para cada planta una gráfica en la cual en el eje de abscisas se muestran el número de árboles, en el eje de ordenadas el error de los modelos, y con color se indican los valores de la tasa de muestreo.

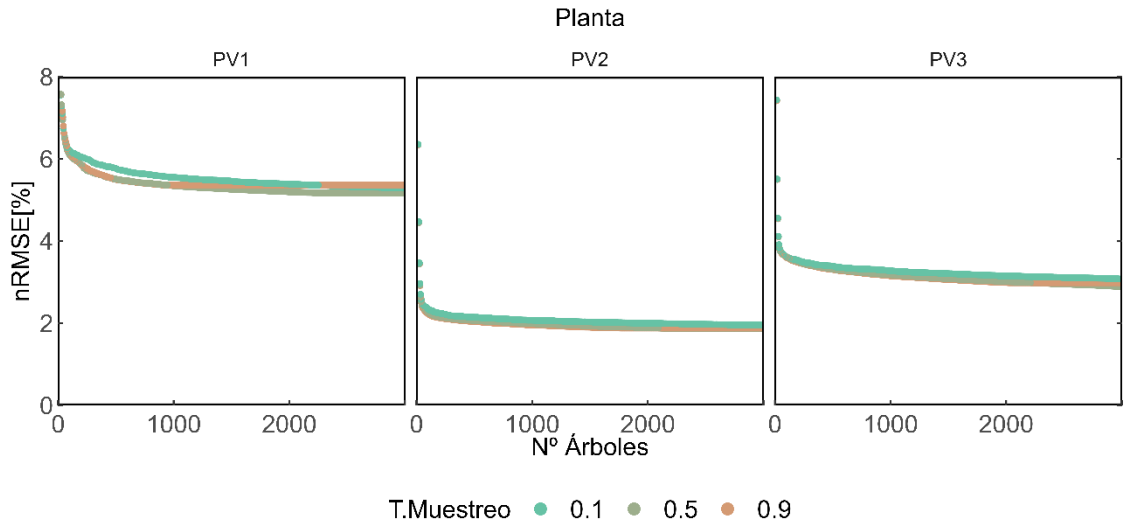


Figura 37. Detalle del error de los modelos generados en la búsqueda. (GB, Profundidad= 2, 4 y 3 para las plantas PV1, PV2 y PV3 respectivamente, muestra de T. Muestreo). Aplicación generación PV

La Figura 37, corrobora que la tasa de muestreo influye muy poco en el error de los modelos, ya que este está prácticamente superpuesto en todas las opciones.

El aumento del número de árboles presenta un comportamiento asintótico en el error, es decir, un aumento de los mismos consigue una disminución del error, sin embargo, a partir de un determinado valor, el error se estabiliza. El número de árboles óptimo para unos mismos valores de profundidad será el menor número de árboles tal que al aumentar estos, el error no disminuya significativamente, considerando como criterio de estabilidad, el mismo que el del modelo RF, ecuación (24).

En las tablas, Tabla 11 a Tabla 13, se presentan una muestra de 279 combinaciones de hiperparámetros, con una profundidad predefinida que evita el sobre-entrenamiento, variando la tasa de muestreo en todo su rango de estudio y el número de árboles entre los valores de 1000 y 2500 con un incremento constante de 50 en 50. Las tablas se han coloreado con el rango del error, de manera que los mejores modelos están en verde. La elección se llevará acabo según el criterio de mayor simplicidad.

Así para la planta PV1, Tabla 11, se observa una convergencia hacia las combinaciones con mayor número de árboles y valores de tasa de muestreo intermedios. Aplicando el criterio de estabilidad y sencillez, la tasa seleccionada es 0.4, y el número de árboles, en torno a los 2400, y si se busca el dato concreto, no representado en la tabla, 2430 árboles.

Tabla 11. Error nRMSE. Selección de hiperparámetros de los modelos GB en PV1. Aplicación generación PV

PV1	T. Muestreo								
Nº Árboles	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1000	5.861	5.767	5.721	5.716	5.693	5.687	5.689	5.670	5.699
1050	5.852	5.768	5.707	5.717	5.692	5.685	5.681	5.667	5.699
1100	5.854	5.762	5.704	5.709	5.685	5.679	5.677	5.665	5.699
1150	5.839	5.756	5.699	5.712	5.679	5.679	5.679	5.666	5.699
1200	5.843	5.750	5.701	5.707	5.682	5.674	5.679	5.666	5.699
1250	5.837	5.747	5.701	5.701	5.679	5.673	5.678	5.666	5.699
1300	5.837	5.740	5.696	5.698	5.673	5.674	5.673	5.666	5.699
1350	5.835	5.738	5.692	5.694	5.677	5.674	5.670	5.666	5.699
1400	5.836	5.731	5.692	5.693	5.676	5.673	5.670	5.666	5.699
1450	5.827	5.728	5.688	5.688	5.672	5.673	5.670	5.666	5.699
1500	5.821	5.732	5.686	5.694	5.667	5.673	5.670	5.666	5.699
1550	5.820	5.724	5.690	5.696	5.665	5.673	5.670	5.666	5.699
1600	5.821	5.714	5.687	5.694	5.666	5.673	5.670	5.666	5.699
1650	5.810	5.718	5.682	5.687	5.663	5.673	5.670	5.666	5.699
1700	5.804	5.713	5.680	5.685	5.662	5.673	5.670	5.666	5.699
1750	5.802	5.711	5.680	5.689	5.658	5.673	5.670	5.666	5.699
1800	5.806	5.707	5.673	5.687	5.658	5.673	5.670	5.666	5.699
1850	5.803	5.710	5.673	5.680	5.654	5.673	5.670	5.666	5.699
1900	5.803	5.715	5.676	5.680	5.651	5.673	5.670	5.666	5.699
1950	5.799	5.712	5.674	5.679	5.650	5.673	5.670	5.666	5.699
2000	5.801	5.712	5.666	5.676	5.653	5.673	5.670	5.666	5.699
2050	5.793	5.707	5.667	5.671	5.649	5.673	5.670	5.666	5.699
2100	5.784	5.704	5.666	5.670	5.650	5.673	5.670	5.666	5.699
2150	5.790	5.698	5.668	5.663	5.649	5.673	5.670	5.666	5.699
2200	5.797	5.699	5.663	5.660	5.649	5.673	5.670	5.666	5.699
2250	5.795	5.699	5.664	5.660	5.650	5.673	5.670	5.666	5.699
2300	5.792	5.704	5.668	5.651	5.650	5.673	5.670	5.666	5.699
2350	5.785	5.699	5.662	5.654	5.650	5.673	5.670	5.666	5.699
2400	5.775	5.694	5.655	5.653	5.650	5.673	5.670	5.666	5.699
2450	5.778	5.693	5.657	5.653	5.650	5.673	5.670	5.666	5.699
2500	5.771	5.692	5.652	5.648	5.650	5.673	5.670	5.666	5.699

Tabla 12. Error nRMSE. Selección de hiperparámetros de los modelos GB en PV2. Aplicación generación PV

PV2	T. Muestreo								
Nº Árboles	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1000	2.221	2.161	2.140	2.140	2.150	2.148	2.142	2.140	2.145
1050	2.199	2.142	2.136	2.134	2.143	2.145	2.142	2.137	2.140
1100	2.201	2.148	2.137	2.141	2.135	2.145	2.139	2.133	2.139
1150	2.202	2.144	2.132	2.138	2.134	2.142	2.135	2.131	2.137
1200	2.208	2.146	2.132	2.135	2.132	2.140	2.134	2.129	2.132
1250	2.206	2.149	2.127	2.124	2.130	2.135	2.133	2.130	2.130
1300	2.192	2.144	2.121	2.117	2.129	2.128	2.130	2.128	2.127
1350	2.208	2.147	2.122	2.122	2.129	2.128	2.129	2.125	2.127
1400	2.203	2.143	2.124	2.118	2.129	2.126	2.130	2.125	2.127
1450	2.203	2.147	2.122	2.118	2.129	2.125	2.124	2.125	2.127
1500	2.193	2.151	2.125	2.115	2.124	2.124	2.120	2.125	2.127
1550	2.193	2.152	2.120	2.119	2.123	2.121	2.128	2.125	2.127
1600	2.198	2.158	2.124	2.127	2.120	2.121	2.125	2.125	2.127
1650	2.199	2.142	2.130	2.127	2.120	2.121	2.124	2.125	2.127
1700	2.192	2.152	2.124	2.126	2.120	2.121	2.124	2.125	2.127
1750	2.192	2.142	2.123	2.118	2.120	2.121	2.124	2.125	2.127
1800	2.209	2.143	2.122	2.117	2.120	2.121	2.124	2.125	2.127
1850	2.191	2.144	2.124	2.116	2.120	2.121	2.124	2.125	2.127
1900	2.191	2.144	2.123	2.117	2.120	2.121	2.124	2.125	2.127
1950	2.189	2.143	2.118	2.115	2.120	2.121	2.124	2.125	2.127
2000	2.180	2.140	2.115	2.116	2.120	2.121	2.124	2.125	2.127
2050	2.179	2.136	2.112	2.114	2.120	2.121	2.124	2.125	2.127
2100	2.188	2.137	2.112	2.120	2.120	2.121	2.124	2.125	2.127
2150	2.186	2.135	2.111	2.116	2.120	2.121	2.124	2.125	2.127
2200	2.194	2.139	2.113	2.120	2.120	2.121	2.124	2.125	2.127
2250	2.181	2.130	2.106	2.117	2.120	2.121	2.124	2.125	2.127
2300	2.188	2.135	2.112	2.113	2.120	2.121	2.124	2.125	2.127
2350	2.192	2.136	2.110	2.114	2.120	2.121	2.124	2.125	2.127
2400	2.181	2.122	2.105	2.114	2.120	2.121	2.124	2.125	2.127
2450	2.177	2.117	2.103	2.114	2.120	2.121	2.124	2.125	2.127
2500	2.181	2.120	2.103	2.113	2.120	2.121	2.124	2.125	2.127

En el caso de la planta PV2, Tabla 12, se repite el mismo patrón, número de valores de árboles

altos con tasas de muestreo intermedias, siendo también el valor seleccionado para tasa de muestreo 0.4. El número de árboles en esta planta es algo menor, en torno a los 1300 y en concreto se seleccionan 1280 árboles.

Tabla 13. Error nRMSE. Selección de hiperparámetros de los modelos GB en PV3. Aplicación generación PV

PV3	T. Muestreo								
Nº Árboles	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1000	3.363	3.291	3.262	3.226	3.227	3.200	3.204	3.203	3.192
1050	3.359	3.286	3.261	3.222	3.221	3.194	3.194	3.203	3.191
1100	3.362	3.280	3.252	3.216	3.219	3.185	3.192	3.200	3.191
1150	3.354	3.278	3.249	3.214	3.217	3.183	3.184	3.200	3.191
1200	3.346	3.275	3.247	3.208	3.214	3.178	3.175	3.200	3.191
1250	3.340	3.269	3.243	3.203	3.217	3.173	3.173	3.200	3.191
1300	3.342	3.266	3.242	3.195	3.209	3.173	3.171	3.200	3.191
1350	3.341	3.265	3.237	3.188	3.206	3.172	3.171	3.200	3.191
1400	3.343	3.271	3.234	3.186	3.205	3.172	3.171	3.200	3.191
1450	3.344	3.273	3.231	3.180	3.201	3.172	3.171	3.200	3.191
1500	3.352	3.271	3.234	3.176	3.203	3.172	3.171	3.200	3.191
1550	3.352	3.267	3.234	3.176	3.205	3.172	3.171	3.200	3.191
1600	3.345	3.265	3.236	3.174	3.204	3.172	3.171	3.200	3.191
1650	3.352	3.262	3.235	3.176	3.203	3.172	3.171	3.200	3.191
1700	3.345	3.266	3.233	3.176	3.199	3.172	3.171	3.200	3.191
1750	3.351	3.265	3.232	3.175	3.195	3.172	3.171	3.200	3.191
1800	3.347	3.264	3.232	3.177	3.194	3.172	3.171	3.200	3.191
1850	3.345	3.263	3.231	3.177	3.194	3.172	3.171	3.200	3.191
1900	3.340	3.262	3.229	3.177	3.194	3.172	3.171	3.200	3.191
1950	3.337	3.260	3.229	3.177	3.194	3.172	3.171	3.200	3.191
2000	3.347	3.256	3.229	3.176	3.194	3.172	3.171	3.200	3.191
2050	3.347	3.255	3.231	3.173	3.194	3.172	3.171	3.200	3.191
2100	3.348	3.254	3.230	3.175	3.194	3.172	3.171	3.200	3.191
2150	3.349	3.252	3.230	3.176	3.194	3.172	3.171	3.200	3.191
2200	3.350	3.251	3.228	3.178	3.194	3.172	3.171	3.200	3.191
2250	3.351	3.247	3.230	3.179	3.194	3.172	3.171	3.200	3.191
2300	3.352	3.249	3.230	3.179	3.194	3.172	3.171	3.200	3.191
2350	3.353	3.250	3.228	3.178	3.194	3.172	3.171	3.200	3.191
2400	3.352	3.249	3.227	3.178	3.194	3.172	3.171	3.200	3.191
2450	3.350	3.248	3.228	3.178	3.194	3.172	3.171	3.200	3.191
2500	3.349	3.246	3.224	3.176	3.194	3.172	3.171	3.200	3.191

En el caso de la planta PV3, Tabla 13, también se consigue minimizar el error con modelos formados por muchos árboles, pero en este caso el valor de tasa de muestreo es algo mayor 0.7. El número de árboles también en torno a los 1300, y en concreto es el mismo valor que para la planta PV2, 1280.

El resumen de los valores óptimos de los hiperparámetros para el modelo GB, junto con el número de modelos que intervienen en la selección se presenta en la Tabla 14.

Tabla 14. Hiperparámetros óptimos de cada planta y número de modelos para su obtención (GB). Aplicación generación PV

Hiperparámetros	PV1	PV2	PV3
T. Muestreo	0.4	0.4	0.7
Profundidad	2	4	3
Nº Árboles	2430	1280	1280
Nº Modelos	90000	90000	90000

Los errores de los modelos de los algoritmos de GB son similares a los errores de los algoritmos con RF. La diferencia entre ambas técnicas es que GB utiliza una mayor cantidad de árboles, pero cada uno de ellos muy pequeño. Este hecho, da agilidad y demanda menos carga computacional en la aplicación del modelo.

Una vez definidos los hiperparámetros para cada planta, se entrenan los modelos GB en cada una de ellas para calcular sus parámetros internos. Posteriormente se calcula la producción con los datos de prueba de cada una de las plantas. El resultado se presenta en la muestra de la Figura 38 y la Figura 39.

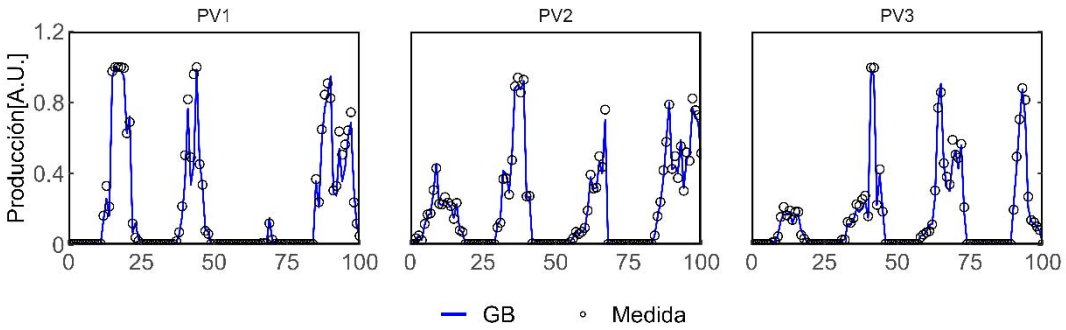


Figura 38. Muestra del resultado del modelo GB. Aplicación generación PV

En la Figura 38 se observa que el modelo GB representa muy bien la realidad tanto a valores bajos de producción como a valores altos, ya que el modelo sigue fielmente las mediciones. En la Figura 39 se presentan todas las medidas del periodo de prueba y la estimación de la producción realizada con el modelo, junto con la función identidad y la tendencia del resultado del modelo.

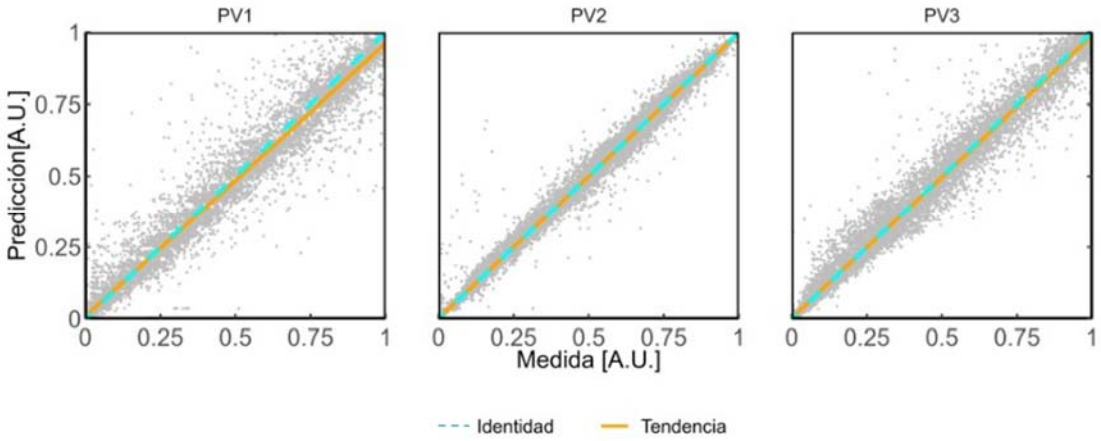


Figura 39. Gráfica de dispersión del resultado del modelo GB. Aplicación generación PV

Como se puede ver en la Figura 39, el modelo GB es adecuado para estimar la producción de una planta fotovoltaica, las tendencias y la función identidad prácticamente son coincidentes en las tres plantas y la dispersión al igual que sucedía con los modelos RF y MLR es menor en la planta PV2, intermedia en la planta PV3 y más amplia que en las anteriores en la planta PV1.

De manera análoga a los modelos RF y MLR, existe el modelo “Quantil Gradient Boosting” (QGB) que permite calcular los intervalos de confianza del modelo. Una muestra de estos intervalos se presenta en la Figura 40, en la cual, con líneas de diferentes colores, se indican los intervalos según el nivel de confianza y con puntos la medida de producción.

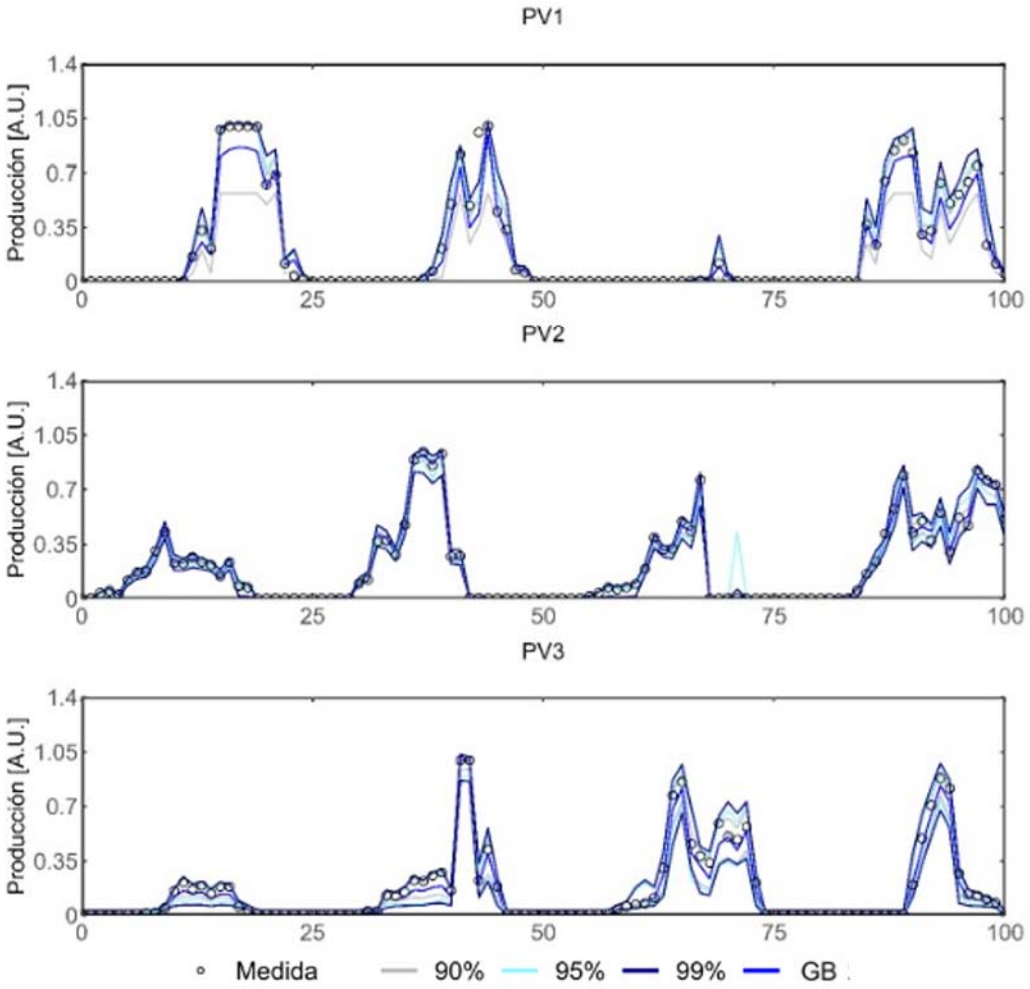


Figura 40. Muestra del intervalo de confianza del modelo QGB. Aplicación generación PV

La representación de la muestra evaluada completa se expone en la Figura 41. En esta figura se muestran las líneas de tendencia de la relación de la medida con el resultado del modelo GB y con el resultado de sus límites de confianza para el 90%, 95% y 99% de probabilidad. También se representan la nube de dispersión de los datos que representa para cada medida la predicción del modelo GB. Comparando los intervalos de QGB con respecto a los intervalos obtenidos en los modelos RF y MLR, se observa que la amplitud es bastante menor

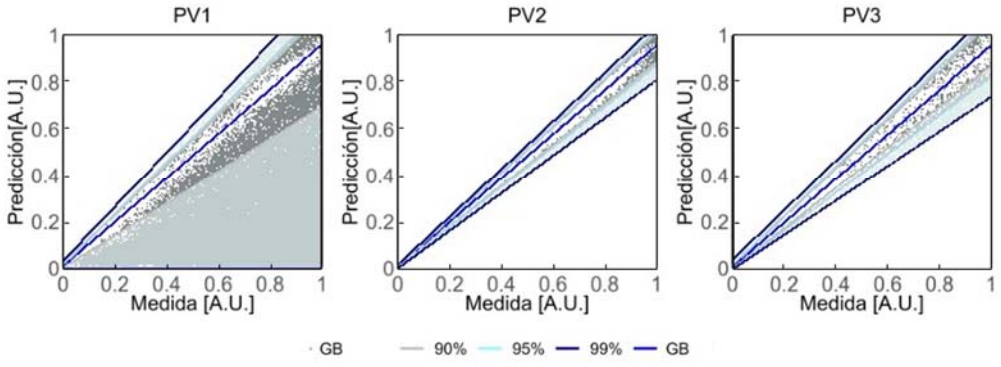


Figura 41. Resultado del intervalo de confianza del modelo QGB. Aplicación generación PV

.

En la Figura 41 puede observarse la asimetría de los intervalos. En la planta PV1 los intervalos son más amplios y para los intervalos de confianza superiores al 90% el límite inferior se mantiene constante en cero. En las otras dos plantas la amplitud de los límites de confianza es inferior siendo la menor amplitud en la planta PV2.

Los errores del modelo GB, son del mismo orden que los encontrados con RF, sin embargo, la amplitud de los intervalos es inferior, y la demanda computacional menor al trabajar con árboles más pequeños, por lo que puede ser una buena opción para modelizar la producción de una planta fotovoltaica.

Redes neuronales

Los algoritmos basados en redes neuronales [66] y en particular las MPNN tienen un amplio rango de hiperparámetros que ajustar para optimizar el modelo. En este estudio se han seleccionado como hiperparámetros a optimizar el número de capas ocultas (Nº Capas), el número de neuronas en cada capa (Nº Neuronas), la tasa de aprendizaje "Learning Rate" (T. Aprendizaje), y las tasas de descarte de datos empleados en el aprendizaje en cada capa oculta que se conocen con la denominación anglosajona "Dropout" (T. Descarte) y de la capa de entrada "Input dropout" (T. Descarte inicial).

El número de capas ocultas y de neuronas definen la forma de la red del modelo. Cuantas más capas tiene un modelo y más neuronas existan en cada capa más complejo es, más capacidad para encontrar relaciones más precisas y, por tanto, se espera que disminuya el error. Pero esto no es siempre así, a veces un exceso de capas y neuronas confunden al modelo y causan sobre-entrenamientos, además de que el tiempo y, por tanto, la carga de computación aumenta haciendo inviable su procesado.

La tasa de aprendizaje define la velocidad a la que aprende el modelo, un modelo que aprende muy deprisa puede no llegar a conseguir los errores mínimos, y viceversa un modelo que aprende muy lento puede no llegar a converger.

La tasa de descarte es el porcentaje de datos de la muestra que se excluyen en una iteración del modelo. Se puede definir una tasa de descarte por cada capa oculta y para la capa de entrada Este hiperparámetro ayuda a agilizar el entrenamiento del modelo y previene el sobre-entrenamiento al conseguir diversidad en las iteraciones de la red.

El número de hiperparámetros que se consideran y el amplio rango de valores que pueden tener alguno de ellos, impiden aplicar una búsqueda de sus valores en malla. De todas las consideradas en este trabajo, la técnica más adecuada para seleccionar los valores de hiperparámetros en las redes neuronales es el método de búsqueda de hiper-banda [186].

En el algoritmo de búsqueda, el criterio de parada impuesto está determinado por la cantidad máxima de recursos que puede asignarse a una única configuración, y η, que es la entrada que controla la proporción de configuraciones descartadas en cada ronda de divisiones sucesivas. Los recursos asignados coinciden con el número de pasadas del modelo en el entrenamiento "epoch" que en este caso de estudio es de 81, y el valor de η es 3, con estos parámetros el número de configuraciones generadas es 10550.

Además, para tener una muestra con un volumen alto de combinaciones se ejecuta el algoritmo de optimización varias veces generando diferentes baterías de pruebas. Cada prueba corresponderá a aplicar el algoritmo a una planta concreta y a un modelo con un número determinado de capas ocultas concreto, lo que implica definir a priori también el número máximo de capas ocultas que se va a comprobar. En una primera aproximación se considera que el número máximo de capas puede ser tres, valor que se verificará y en caso de no ser correcto se aumentará repitiendo el proceso tantas veces como sea necesario.

Trabajar con tres plantas y modelos de una, dos y tres capas genera nueve baterías de pruebas en las que en cada una se habrán seleccionado 10550 modelos. Esto implica un total de 94.950 modelos y de los cuales cada planta tendrá un total 31650 modelos seleccionados. Los hiperparámetros de ajuste del método de hiper-banda se muestran en la Tabla 15. La Figura 42 representa el error nRMSE de todos modelos

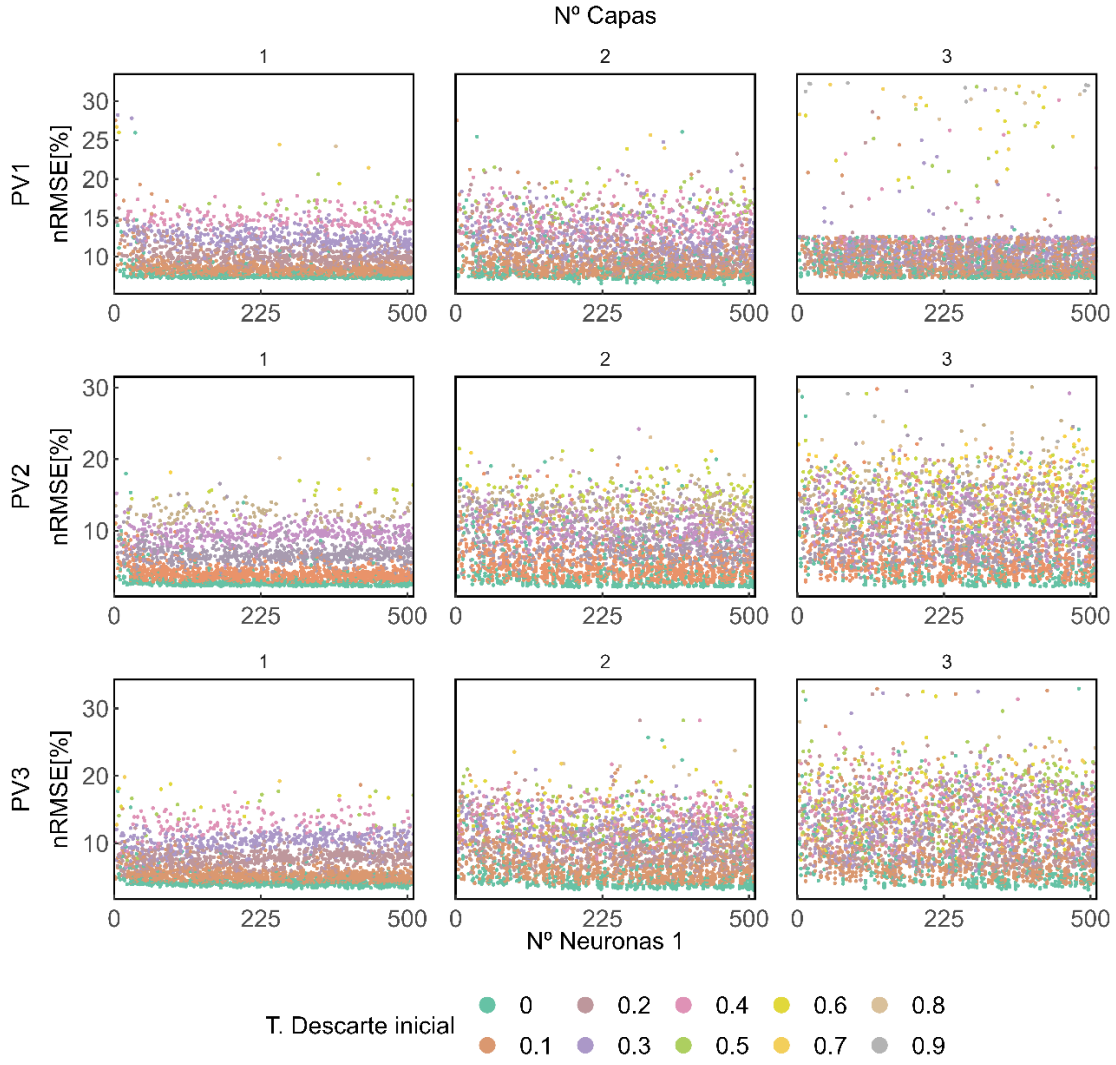


Figura 42. Representación del error de todos los modelos generados en la búsqueda. Método de hiper-banda (ANN). Aplicación generación PV

Tabla 15. Rango de valores de los hiperparámetros y número de modelos. Método de Hiper-banda (ANN). Aplicación generación PV

	PV1			PV2			PV3		
Prueba	1	2	3	4	5	6	7	8	9
Nº Capas	1	2	3	1	2	3	1	2	3
T. Aprendizaje	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001
T. Descarte Inicial	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1
T. Descarte Capa 1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1
T. Descarte Capa 2	-	0-0.9 delta 0.1	0-0.9 delta 0.1	-	0-0.9 delta 0.1	0-0.9 delta 0.1	-	0-0.9 delta 0.1	0-0.9 delta 0.1
T. Descarte Capa 3	-	-	0-0.9 delta 0.1	-	-	0-0.9 delta 0.1	-	-	0-0.9 delta 0.1
Nº Neuronas Capa1	2-512 delta 2	2-512 delta 2	2-512 delta 2	2-512 delta 2	2-512 delta 2	2-512 delta 2	2-512 delta 2	2-512 delta 2	2-512 delta 2
Nº Neuronas Capa2	-	2-512 delta 2	2-512 delta 2	-	2-512 delta 2	2-512 delta 2	-	2-512 delta 2	2-512 delta 2
Nº Neuronas Capa 3	-	-	2-512 delta 2	-	-	2-512 delta 2	-	-	2-512 delta 2
Nº modelos	10550	10550	10550	10550	10550	10550	10550	10550	10550

La Figura 42 representa el error nRMSE de los modelos, organizados en nueve gráficos distintos, cada uno correspondiente a una prueba específica. Los gráficos están dispuestos de tal manera que cada fila representa una planta, mientras que cada columna muestra los gráficos con el mismo número de capas ocultas en sus modelos. En cada gráfico, se han seleccionado únicamente los 3400 modelos con el error más bajo, excluyendo aquellos con errores significativamente altos. En ellas se representa el error de los modelos en función del número de neuronas en la primera capa oculta. Además, se utilizan diferentes colores para indicar la tasa de descarte inicial.

El análisis de la Figura 42 proporciona información valiosa sobre la optimización de los hiperparámetros en los modelos. Al ajustar los hiperparámetros adecuadamente, se logra reducir el error en los modelos, a veces en un factor de casi 10 en comparación con los valores iniciales.

Además, se observa una superposición visible de modelos, lo que sugiere que existen múltiples combinaciones de hiperparámetros que logran resultados similares en términos de error.

La superposición de los errores de los modelos indica que no hay una única configuración de hiperparámetros que sea la mejor, sino que existen diferentes combinaciones válidas que pueden producir resultados similares y por tanto la elección de los mismos seguirá el principio de máxima simplicidad. También permite verificar, que visualizar solo en el análisis los 3400 mejores modelos de cada prueba es suficiente ya que la muestra representada cubre un amplio rango de valores de errores concentrando además modelos con errores mínimos con un comportamiento asintótico.

Los modelos de 3 capas, más complicados, presentan un error similar o incluso mayor a los modelos de menos capas. Este hecho verifica que no es necesario ampliar la búsqueda a modelos de más capas. Además, observando la Figura 42, se puede determinar que el número menor y suficiente de capas ocultas para las plantas PV2 y PV3 es 1, mientras que para la planta PV1, podría ser tanto 1 como 2 capas ocultas las óptimas para el modelo.

Por último, se observa que en todas las pruebas realizadas la tasa de descarte inicial es nula para minimizar el error, es decir es mejor trabajar con todas las observaciones en la capa de entrada.

Considerando las selecciones previas, se representan de nuevo solo los modelos de una y dos capas ocultas para la planta PV1, y los modelos de una capa oculta para las plantas PV2 y PV3, y que cumplen que no tienen tasa de descarte inicial, Figura 43. En esta figura se muestran tres gráficas, una para cada planta, en cada gráfica se representa el error en función de las neuronas de la capa 1, y se muestra la tasa de aprendizaje con diferente color.

En la Figura 43, se observa que las elecciones realizadas hasta el momento conservan los valores mínimos de errores y que además estos son similares a los que se alcanzaban con los modelos anteriores. Además, la tasa de aprendizaje no tiene una gran influencia sobre el error. Con cualquiera de los tres valores se puede llegar a alcanzar el mínimo error, por lo que se va a considerar el más pequeño, 10^{-4}, para estar del lado de la seguridad al minimizar la velocidad del aprendizaje.

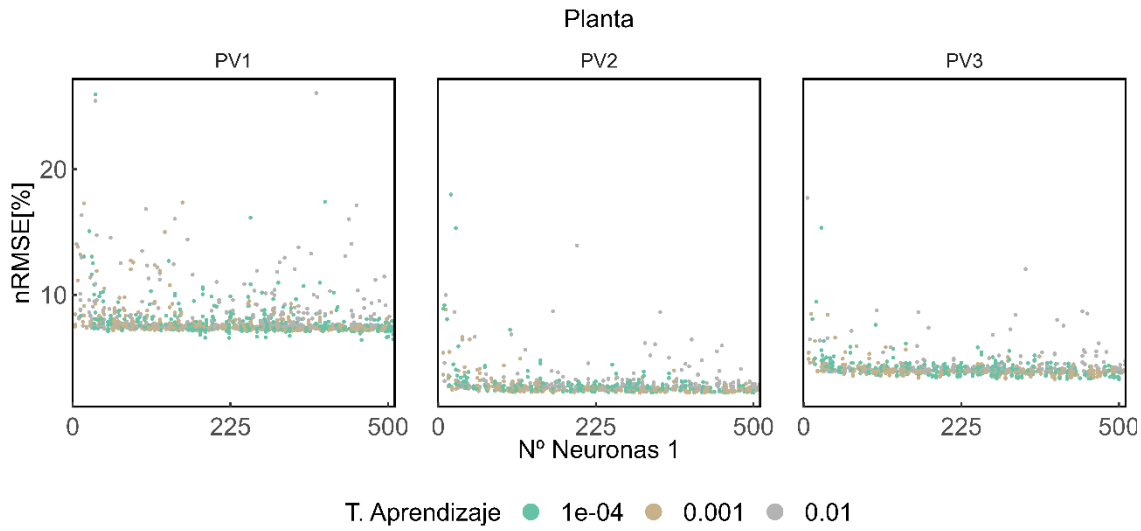


Figura 43. Detalle del error de los modelos generados en la búsqueda. Barrido1 (ANN, T. Descarte inicial nula, Nº Capas: la planta PV1, 1-2, y para las plantas PV2 y PV3, 1 capa). Aplicación generación PV

Aplicando este nuevo criterio discriminatorio el volumen de modelos deja de ser representativo para identificar los hiperparámetros que faltan por estudiar, como las tasas de descarte y las neuronas de las capas ocultas. Por ese motivo se complementa el estudio con una búsqueda aleatoria "Random Search", que permitirá buscar modelos con valores de hiperparámetros aleatorios, en este caso dentro del rango de valores que se ha seleccionado hasta el momento, poblando así de nuevos valores el rango de hiperparámetros de interés.

Se realizan en este caso una batería de 4 pruebas, una por planta y en el caso de la planta PV1 se realizan dos porque se prueban por separado los modelos de 1 y 2 capas. El criterio de parada es de un máximo de 4000 modelos. Los rangos de hiperparámetros se resumen en la Tabla 16, donde los hiperparámetros definidos por la búsqueda anterior se mantienen fijos.

Tabla 16. Rango de valores de los hiperparámetros y números de modelos generados en la búsqueda. Método búsqueda aleatoria (ANN). Aplicación generación PV

	PV1		PV2	PV3
Prueba	1	2	3	4
Nº Capas	1	2	1	1
T. Aprendizaje	0.0001	0.0001	0.0001	0.0001
T. Descarte inicial	0	0	0	0
T. Descarte capa 1	0-0.3 delta 0.1	0-0.3 delta 0.1	0-0.3 delta 0.1	0-0.3 delta 0.1
T. Descarte capa 2	-	0-0.3 delta 0.1	-	-
Units capa 1	2-512 delta 1	350-512 delta 1	2-512 delta 1	2-512 delta 1
Units capa 2	-	350-512 delta 1	-	-
Nº Modelos	4000	4000	4000	4000

La representación gráfica de todos los modelos seleccionados de la primera búsqueda con método de hiper-banda junto con los modelos de la búsqueda aleatoria en cada planta se presenta en la

Figura 44. La figura está formada por cinco gráficas diferentes en las que se presentan los errores de los modelos en función del número de neuronas de la primera capa o de la segunda capa. Las gráficas están ordenadas por filas, en cada fila se representa una prueba diferente. Así en la primera fila hay solo una gráfica que corresponde a la representación de los modelos de una capa de la planta PV1. Por lo tanto, el error está solo en función de las neuronas de la primera capa y con color se representa el valor de la tasa de descarte de la primera capa. En la segunda fila hay dos gráficas porque el modelo tiene dos capas y se ha discriminado la representación del error en dos gráficas cada una de ellas en función de las neuronas de la primera o de la segunda capa. En cada una de las gráficas se indica la tasa de descarte de la primera capa con color y el de la segunda capa con símbolo. En la tercera y cuarta fila se representan las mismas gráficas que en la primera fila, pero para las plantas PV2 y PV3.

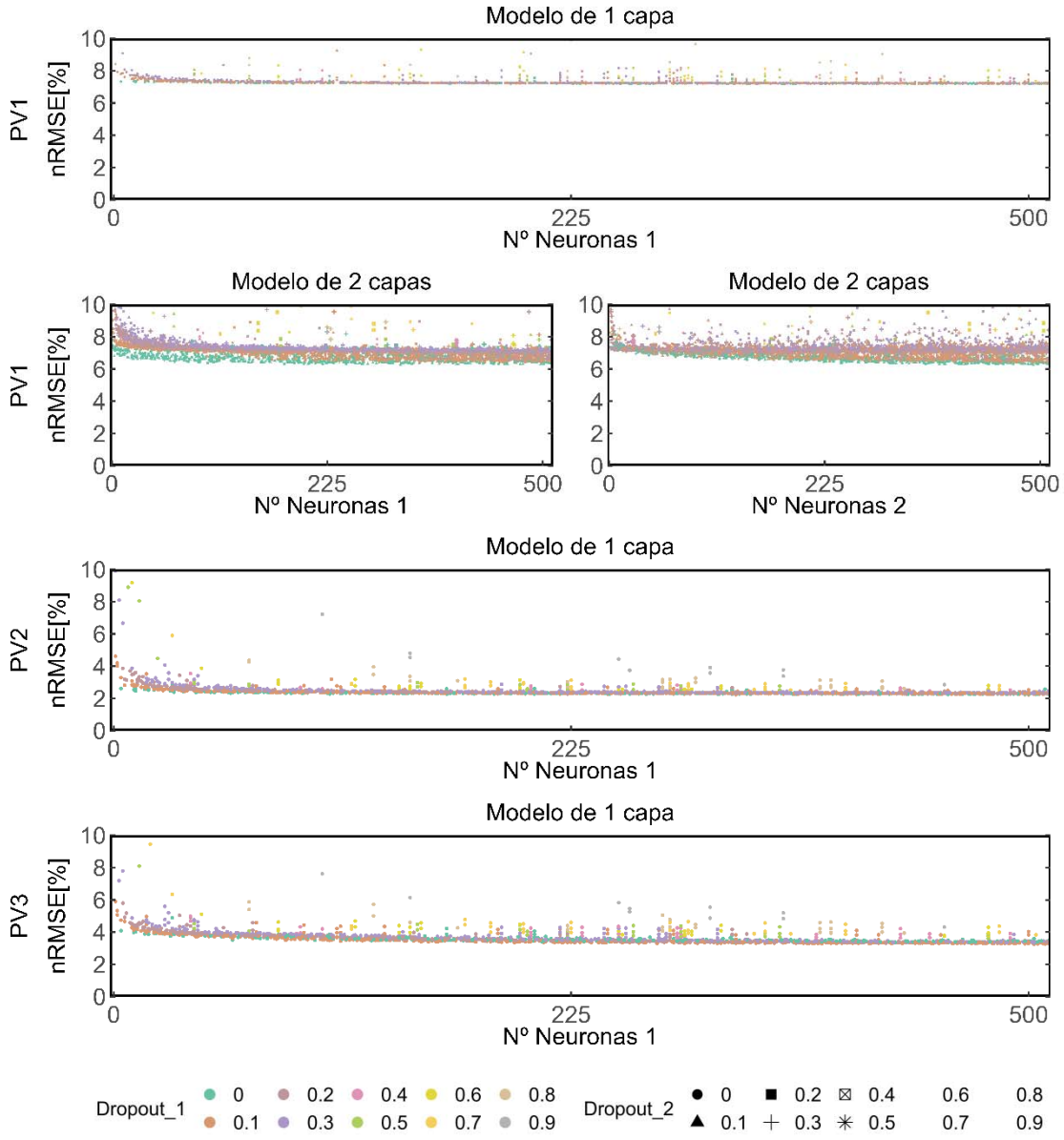


Figura 44. Representación del error de los modelos generados en la búsqueda. Barrido1&2 (ANN, T. Descarte inicial nula, Nº Capas: la planta PV1, 1-2, y para las plantas PV2 y PV3, 1 capa, T. Aprendizaje=10-4). Aplicación generación PV

En la Figura 44, se observa que en la planta PV1, se consigue minimizar el error cuando se trabaja con modelos de dos capas con tasas de descarte para las capas ocultas nulas. Los modelos de las plantas PV2 y PV3 muestran que las tasas de descarte que minimizan el error son respectivamente las de valor 0 y 0.1.

El comportamiento de las neuronas en todos los casos es asintótico, por lo que el criterio de búsqueda del número de neuronas es en base a la estabilidad del modelo y el mínimo error. En el caso de las plantas PV2 y PV3 las neuronas de la capa 1 que estabilizan el error en una diferencia de 10^{-3}, son los valores de 204 y 114 respectivamente. En el caso de la planta PV1, los valores de las neuronas de cada capa están relacionados, por lo que se realiza un análisis más detallado, (ver Tabla 17).

En la Tabla 17 se muestra la relación entre las neuronas de la capa 1 y 2. Escogiendo los modelos más sencillos que mantengan el error nRMSE en un rango bajo, la mejor opción es en torno a 240 neuronas para la capa 1 y 470 neuronas para la capa dos, y en concreto 239 y 475 respectivamente en las capas ocultas 1 y 2.

Según estos criterios y análisis aplicados para definir el número de neuronas, los hiperparámetros para las redes neuronales, junto con el número de modelos que han intervenido en la selección, se adjuntan en la Tabla 18.

El ajuste de hiperparámetros que se obtiene con las redes neuronales indica que los errores que se alcanzarán serán ligeramente más altos que los de los otros modelos. También se observa que la complejidad de la red va en función de la dificultad para modelizar las relaciones entre las variables de las plantas, siendo la red más compleja la de la planta PV1. También se observa como la planta PV3 que tiene más variables se optimiza con una tasa de descarte de la primera capa oculta ligeramente más alta a los de los otros modelos.

Tabla 17. Error nRMSE. Selección de hiperparámetros de los modelos ANN en PV1. Aplicación generación PV

PV1	Nº Neuronas 2										
Nº Neuronas 1	400	410	420	430	440	450	460	470	480	490	500
90				6.485	6.615					6.446	
100	6.576	6.493	6.587	6.748	6.500	6.510	6.508		6.578		6.775
110				6.524							6.384
120	6.608				6.458			6.390	6.519		6.471
140		6.594	6.660				6.551			6.432	
150			6.488							6.465	
160						6.461		6.480		6.472	
170			6.817				6.425				
180				6.592	6.410	6.399	6.602				
190						6.521					
200						6.441		6.433	6.534	6.639	
210	6.529	6.528				6.393	6.379			6.478	6.539
220			6.518		6.357		6.480	6.383			6.438
230		6.506		6.462				6.371		6.450	
240		6.447				6.379		6.303	6.406	6.457	6.327
250			6.370				6.473	6.402			
260	6.659		6.738			6.473				6.361	
270				6.557				6.423		6.306	
280							6.527				
290	6.352						6.799			6.361	6.317
300				6.388							

Tabla 18. Hiperparámetros óptimos de cada planta y números de modelos para su obtención (ANN). Aplicación generación PV

	PV1	PV2	PV3
Nº Capas	2	1	1
T. Aprendizaje	0.0001	0.0001	0.0001
T. Descarte inicial	0	0	0
T. Descarte 1	0	0	0.1
T. Descarte 2	0	-	-
Nº Neuronas 1	239	204	114
Nº Neuronas 2	475	-	-
Nº Modelos	39650	35650	35650

A continuación, se entrenan los modelos para calcular los parámetros en cada planta y se estiman los resultados para el periodo de prueba. Una muestra de estos resultados se presenta en la Figura 45. En el ejemplo representado se observa que este modelo mejora el resultado del MLR, pero no reproduce las medidas como los modelos basados en árboles de decisión en el rango de las producciones altas.

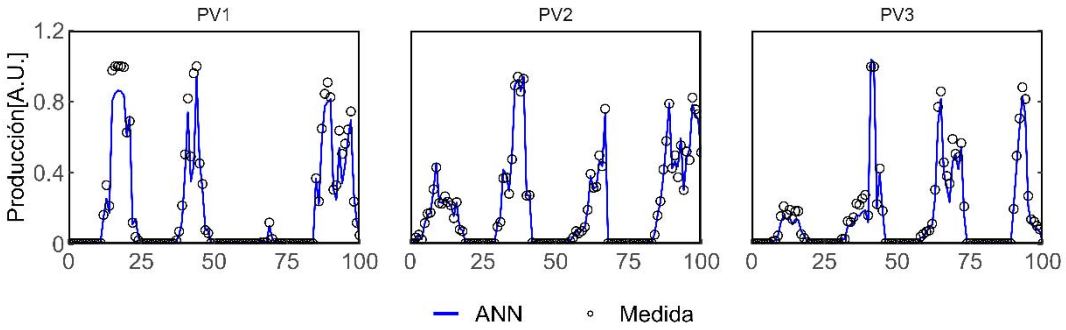


Figura 45. Muestra del resultado del modelo ANN. Aplicación generación PV

La representación del periodo completo en la gráfica de dispersión muestra la relación de la medida de generación con la estimación del modelo junto con su tendencia y la función identidad, (ver Figura 46).

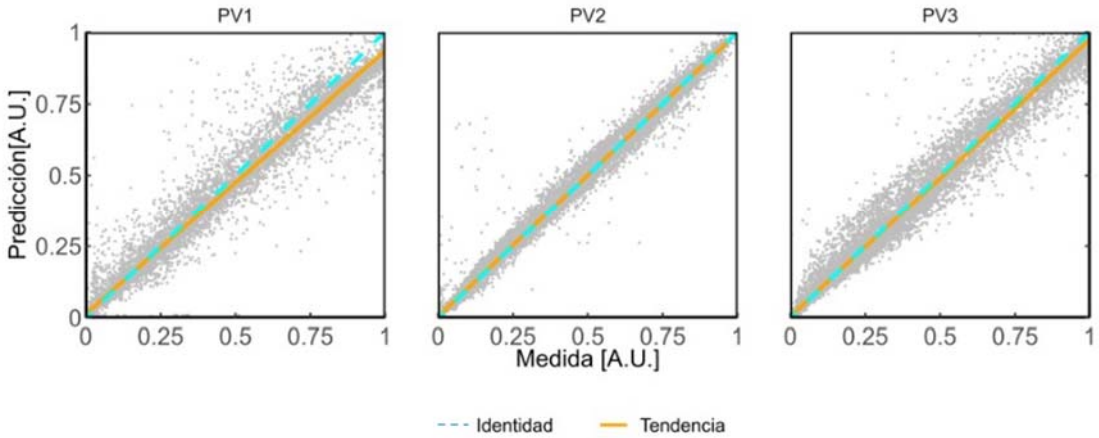


Figura 46. Gráfica de dispersión del resultado del modelo ANN. Aplicación generación PV

Estudiando el periodo de prueba completo representado en la Figura 46, se observa que en la planta PV1 la tendencia del modelo es ligeramente diferente a la función identidad. En el caso de las plantas PV2 y PV3 las líneas de tendencia y de identidad son coincidentes y por tanto la precisión de los modelos de ANN parece similar a los modelos basados en árboles de decisión.

En cuanto a la dispersión de los datos se mantiene la tendencia observada con los anteriores modelos, la menor dispersión la presenta la planta PV2, seguida de la planta PV3 y de la planta PV1.

Los intervalos de confianza del modelo se calculan modificando las capas de salida de la red para tener una distribución Normal de soluciones sobre las que se puede aplicar un intervalo con nivel de confianza del 90%, 95% y 99%. Esta modificación se lleva a cabo mediante la incorporación de la capa "layer_distribution_lambda", como capa de salida, donde el parámetro lambda de la capa indica el tipo de distribución en el que se presenta la salida que en este caso es la distribución Normal.

Para automatizar este proceso, se incorpora una capa adicional en el modelo de la red neuronal llamada "layer-variable".[208]. Esta capa genera estimaciones puntuales de manera iterativa, donde en cada iteración se introducen variaciones en los parámetros del modelo.

El resultado se muestra en la Figura 47, en la cual los intervalos se trazan mediante líneas de colores, cada color representa un nivel de confianza, también se muestra en el resultado del modelo ANN y las medidas reales. En esta figura se aprecia la simetría de los intervalos.

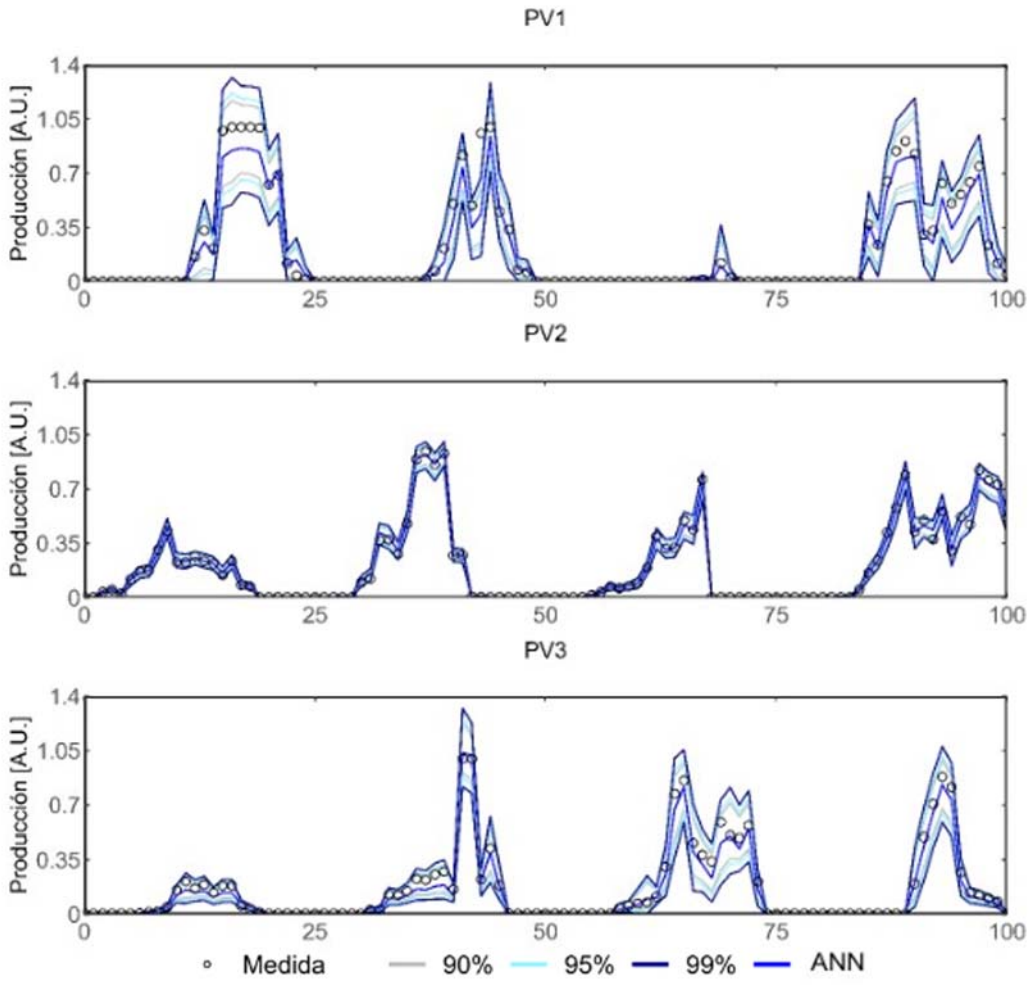


Figura 47. Muestra del intervalo de confianza del modelo ANN. Aplicación generación PV

La Figura 48, muestra las líneas de tendencias de la relación de la medida con el resultado del modelo y de la medida con los diferentes intervalos. El espacio entre estas líneas se sombrea con colores distintos para indicar la probabilidad correspondiente a cada intervalo. Además, se incluye la nube de dispersión de puntos correspondiente a la medida y la predicción del modelo ANN.

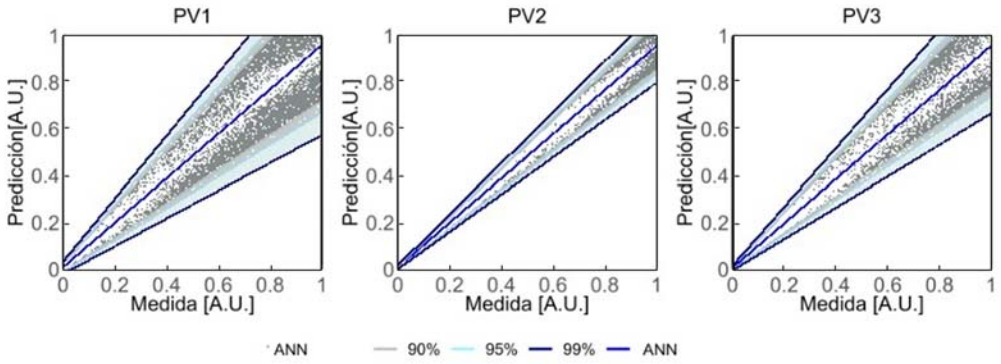


Figura 48. Resultado del intervalo de confianza del modelo ANN. Aplicación generación PV

Observando la Figura 48, se ve que los intervalos son muy ajustados, presentando cierta simetría, que no es habitual en la producción de las plantas fotovoltaicas, cuyo máximo de producción está limitado por la capacidad del inversor y mientras que el mínimo puede llegar al valor nulo.

3.4. Discusión de los resultados

Para valorar y comparar los modelos se calculan las métricas definidas en la Sección 2.7, nRMSE, nMAE, nBIAS, SS [227] y la mediana de la amplitud del intervalo en tanto por ciento con los tres niveles de confianza estimados en todos los escenarios de trabajo definidos con los cuatro modelos aplicados a tres plantas de estudio.

Las métricas comparan los resultados de los modelos con la realidad atendiendo a diferentes relaciones matemáticas que cubren todos los aspectos de su comportamiento. Los resultados serán además las magnitudes que servirán para comparar los modelos entre ellos. El resumen de los cálculos de las métricas se presenta en la Tabla 19.

Tal y como se observa en la Tabla 19, en los dos errores calculados, nRMSE y nMAE, en cualquiera de las tres plantas, los modelos de ML presentan menor valor respecto a MLR, lo que significa que los modelos más complejos son capaces de captar mejor el comportamiento de las plantas. Respecto al sesgo de los modelos su valor es siempre muy bajo independientemente del modelo del que se trate. Si se analiza el valor de SS, que indica la mejora respecto a la referencia, MLR, se observa como todos los modelos más avanzados mejoran al base, siendo los modelos GB y RF, los que muestran una mejora significativa. Y de estos dos, en términos de estimación puntual, el modelo RF resulta ligeramente el más preciso de todos.

Por otro lado, al analizar la incertidumbre de los modelos, se observa que los intervalos de confianza obtenidos para el modelo RF son demasiado amplios. En contraste, los modelos GB y ANN presentan intervalos más estrechos, lo que indica una mayor precisión en la estimación de la incertidumbre. Entre ellos se diferencian sustancialmente en el nivel de confianza del 99% ya que las redes neuronales tienden a hacer simétrico el intervalo de confianza.

Tabla 19.Resumen de las métricas de los diferentes modelos. Aplicación generación PV

	PV1				PV2				PV3			
	MLR	RF	GB	ANN	MLR	RF	GB	ANN	MLR	RF	GB	ANN
nRMSE	7.9%	5.4%	5.9%	6.6%	2.6%	1.9%	2%	2.2%	4.1%	3.1%	3.2%	3.5%
nMAE	2.8%	1.9%	2.2%	2.4%	1.2%	0.8%	0.8%	1.0%%	2.1%	1.3%	1.4%	1.7%
nBIAS	0.0%	0.1%	0%	0%	0%	0%	0%	-0.3%	0%	0%	0%	0%
SS	-	*30.8%*	*25.5%*	*15.5%*	-	*25.5%*	*22.7%*	*14.2%*	-	*24.9%*	*22.1%*	*13.7%*
Mediana de la amplitud del intervalo 90%	1.8%	6.0%	0.6%	3.2%	1.6%	2.4%	0.2%	0.6%	1.2%	4.3%	0.3%	4.0%
Mediana de la amplitud del intervalo 95	5.6%	9.3%	1.1%	3.8%	2.0%	4.2%	0.2%	0.7%	1.7%	6.6%	0.5%	0.5%
Mediana de la amplitud del intervalo 99%	23.6%	22.7%	2.8%	5.0%	4.2%	10.0%	1.0%	0.9%	6.3%	17.9%	3.1%	0.6%

Además de la precisión y la incertidumbre, también se debe considerar la demanda computacional y la complejidad de los modelos. Tanto el RF como las ANN requieren una mayor capacidad computacional, y las ANN son más complejas en comparación con los otros modelos. Teniendo en cuenta estos factores y los resultados compensados obtenidos por el modelo GB, se concluye que el modelo “Gradient Boosting (GB)” es el más adecuado para estimar la producción en una planta de gran tamaño con falta de uniformidad en la irradiancia de los equipos de medida y la generación.

4. Aplicación de los modelos ML al cálculo de la irradiancia

4.1. Introducción

La irradiancia es la variable que más afecta a la precisión de la estimación de la producción de un sistema fotovoltaico. Para garantizar resultados precisos en dicha estimación, es necesario asegurar la calidad de los datos de la medida de la irradiancia. Sin embargo, la verificación de esta variable no es una tarea sencilla, ya que tiene un comportamiento periódico interrumpido por cambios rápidos e impredecibles. Por ejemplo, la medida de irradiancia puede pasar, en las horas centrales del día, de valores máximos a mínimos debido a la aparición de sombras o nubosidad, sin que ello signifique un error en la medición.

Este hecho, hace que los métodos estadísticos de filtrado y detección de cambios de comportamiento, que suelen ser efectivos para identificar medidas anómalas en otras variables, no resulten útiles en el caso de la irradiancia [229]. Para garantizar la calidad de esta variable, se aplica un proceso similar al descrito en el análisis de la producción de las instalaciones fotovoltaicas. Este proceso implica la modelización de una referencia de la variable irradiancia, con la que se pueda comparar la señal medida e identificar posibles anomalías en la medición de la misma.

En este capítulo se lleva a cabo la estimación de la irradiancia a partir de variables medidas en las estaciones meteorológicas, y utilizando algoritmos basados en metodologías de RF, GB y ANN. La analogía en la metodología con el estudio de la producción permite además analizar las diferencias en el comportamiento de los algoritmos seleccionados al trabajar con otras variables y propósito.

Los resultados obtenidos de los diferentes algoritmos se comparan con los obtenidos con el algoritmo MLR. Este algoritmo se utiliza como referencia para determinar cuál de los métodos es más apropiado para estimar la irradiación, a partir de variables meteorológicas, medidas en la estación meteorológica, sin incluir sensores o cámaras adicionales.

Disponer de un algoritmo, capaz de calcular la irradiancia, cuando esta no se mide directamente en las plantas, es un valor añadido ya que su resultado además de ser utilizado para identificar anomalías en la medida puede ayudar a conocer la eficiencia de las plantas y contrastar el recurso existente en los emplazamientos con las bases de datos disponibles.

Para llevar a cabo el estudio se trabaja con una estación de medición y en base a las variables meteorológicas medidas de la estación se determinará la Irradiancia Global Horizontal con las cuatro técnicas mencionadas, MLR, RF, GB y ANN. Para definir los modelos de los algoritmos de ML se buscan los hiperparámetros adecuados, se entrenan para determinar los parámetros y se calcula la irrandiancia. El resultado se completa calculando el intervalo de confianza para la estimación realizada.

El capítulo se organiza exponiendo primero la metodología utilizada, y, posteriormente, se presentarán y discutirán los resultados obtenidos.

4.2. Metodología

La metodología de trabajo es similar a la utilizada en el Capítulo 3. La diferencia esencial reside en que para los modelos que estiman la irradiancia se utiliza un modelo paramétrico previo para calcular la irradiancia teórica. Este dato, junto con las medidas de la torre meteorológica, se utilizan para obtener los modelos para cada algoritmo seleccionado. A continuación, se describe la estación meteorológica utilizada en el estudio.

Descripción de la estación de medición

Para llevar a cabo este estudio, se trabaja con la estación de medición ubicada en la planta PV1.

Figura 49. Estación meteorológica

Dicha estación cuenta con diversos instrumentos para medir variables meteorológicas, tales como un piranómetro para medir la irradiancia global horizontal, un termo-higrómetro para medir la temperatura y humedad ambiente, un anemómetro para medir la velocidad del viento, una veleta para medir la dirección del viento y un barómetro para medir la presión atmosférica. La disposición de estos sensores se puede observar en la Figura 49.

Las mediciones meteorológicas se realizan cada tres segundos y se registran las medias de los datos obtenidos en intervalos de diez minutos. Por lo tanto, los datos utilizados en este estudio serán medias de diez minutos.

Además de los sensores mencionados, la estación meteorológica cuenta con un registrador capaz de almacenar la información y grabar los datos promediados cada intervalo de diez minutos registrando la fecha y hora además del valor de las distintas variables. Para alimentar al registrador, se dispone de una placa fotovoltaica, una batería y un regulador. La información se transfiere al SCADA de la planta mediante un sistema de fibra óptica, donde los datos se integran con el resto de las señales de la planta.

Las variables climatológicas que se miden en la estación y que pueden influir en la irradiancia (IR) son la temperatura ambiente (T_{amb}), la humedad (H), la presión (B) y la velocidad del viento (Vel). Estas variables determinan las condiciones atmosféricas y, por lo tanto, la claridad del cielo, junto con otras variables no medidas como la contaminación o el nivel de partículas. Se utilizan, además, las etiquetas temporales fecha y hora, porque proporcionan información sobre la posición del sol, necesaria para obtener la irradiancia teórica.

Metodología de trabajo

La descripción del procedimiento empleado para estimar la irradiancia a partir de datos meteorológicos medidos en la estación, la fecha, y hora, la ubicación geográfica y la altitud del terreno, se presenta en la Figura 50.

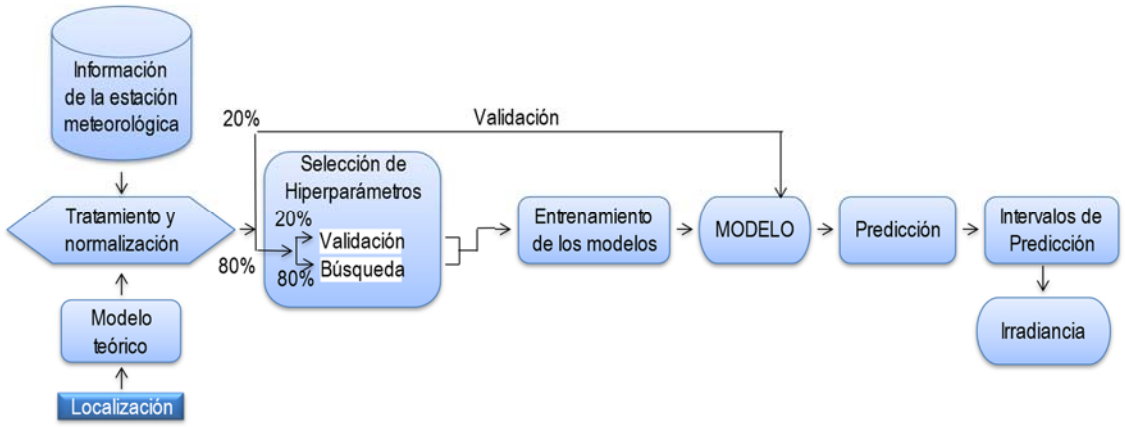


Figura 50. Flujo de trabajo para estimar la irradiancia

Selección de Información

En el algoritmo presentado, los datos provienen de dos fuentes distintas: la información registrada en la estación de medición y la irradiancia teórica que se recibe del Sol en el punto de la estación de medición.

Las variables meteorológicas de la estación de medición, como la temperatura ambiente (T_{amb}), la humedad (H), la presión (B) y la velocidad del viento (Vel), no presentan una correlación clara con la irradiancia (IR), como se muestra en los diferentes gráficos de la Figura 51.

Según la Figura 51, las correlaciones entre la irradiancia con la temperatura y la humedad son más altas que con la presion y la velocidad. A medida que aumenta la temperatura, también lo hace la irradiancia, lo cual es previsible ya que las temperaturas más altas se encuentran en días soleados. Lo mismo ocurre con la humedad que presenta una correlación negativa: los días de lluvia, generalmente nublados, tienen una humedad superior a los días secos y despejados. Sin embargo, la presión y la velocidad del viento tienen un comportamiento indefinido. La presencia de viento puede despejar las nubes, pero también puede ser un síntoma de nubosidad. En cuanto a la presión, no hay tendencia.

La relación combinada de estas variables induce a comportamientos confusos para los modelos y a errores e incertidumbres muy grandes. Esto se debe a que combinaciones de los mismos valores según la época del año y la ubicación son indicativas de irradiancias completamente diferentes. Por ejemplo, en latitudes Norte, tener 15ºC de temperatura a medio día en el mes de febrero es un indicativo de un día totalmente despejado. Sin embargo, ese mismo valor de temperatura a la misma hora en agosto puede ser indicativo de un cielo totalmente nublado. De la misma forma, en otra latitud también podría tener un significado diferente. Para solventar este problema, se incluyen las variables temporales, la hora (h) y el mes (m), que se obtienen de la misma estación de medición y la irradiancia teórica.

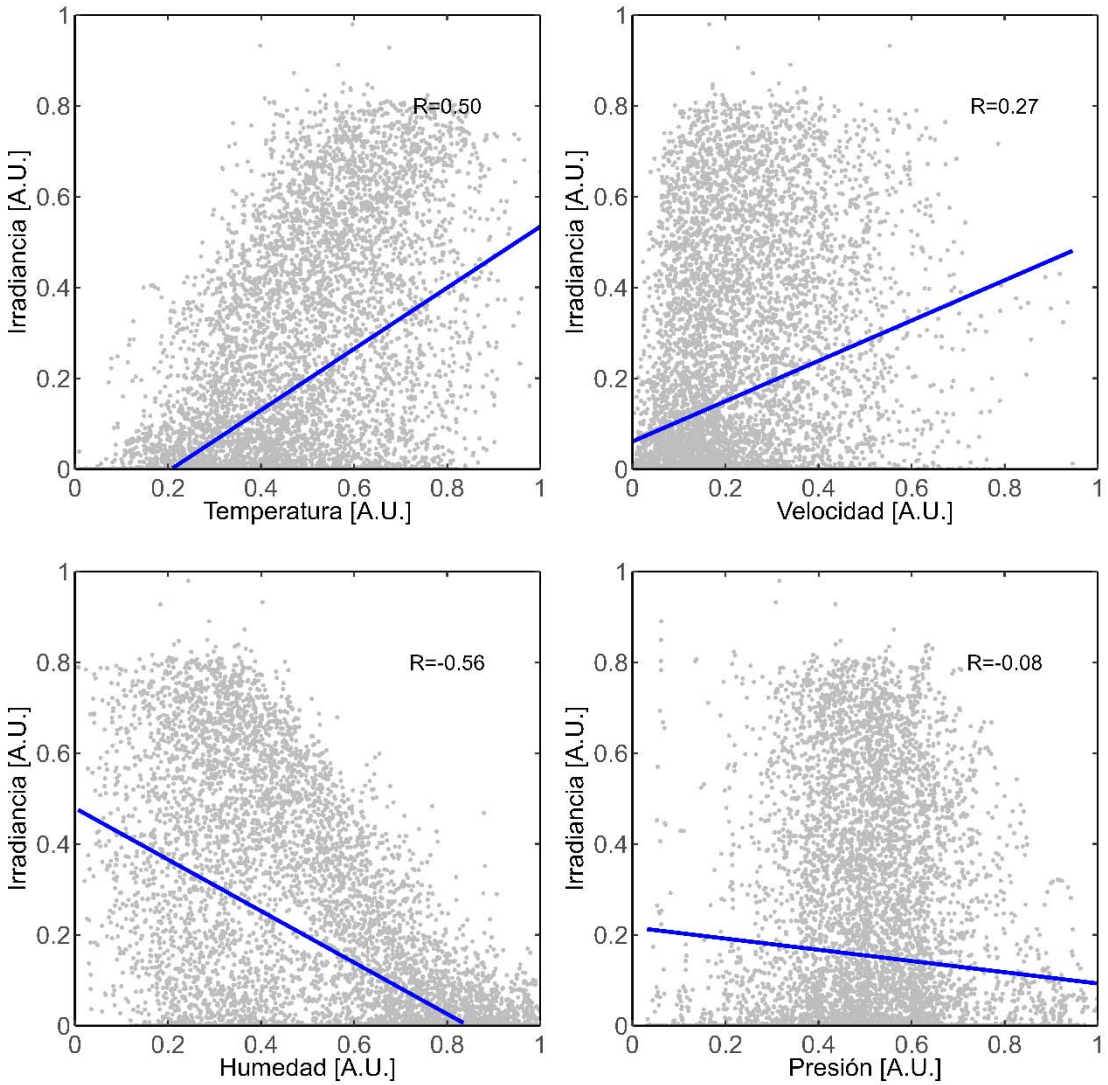


Figura 51. Correlación entre las variables medidas en la estación meteorológica y la irradiancia

A partir de la fecha, hora, minuto y de la ubicación (latitud) de las planta se calcula la irradiancia teórica ($IR_{teorica}$) que llega a ese punto concreto de la tierra y que se obtiene con un modelo paramétrico basado en geometría. Este modelo proporciona estimaciones de la irradiación solar que llega a la superficie en ausencia de nubes. En concreto se utiliza el algoritmo de cielo despejado, "European Solar Radiation Atlas" (ESRA) que es un modelo desarrollado en el marco del Atlas Digital Europeo de Radiación Solar [145,230].

ESRA requiere del conocimiento de la ubicación geográfica, de la elevación del terreno, información conocida de antemano y del factor de turbidez Linke "LinkeTurbidity Factor" [231,232]. El factor de turbidez Linke es una aproximación para simular la absorción y dispersión atmosférica de la radiación solar en cielos despejados. Describe el espesor óptico de la atmósfera debido tanto a la absorción por el vapor de agua como a la absorción y dispersión por las partículas de aerosol en relación con una atmósfera seca y limpia. Resume la turbidez de la atmósfera y, por lo tanto, la atenuación del haz directo de radiación solar. Cuanto mayor sea su valor, mayor será la atenuación de la radiación por la atmósfera; es una medida de la transparencia de esta en ausencia de nubes. Si el cielo estuviera perfectamente seco y limpio (azul profundo), el factor tendría un valor igual a

la unidad. Cuando hay una gran cantidad de vapor de agua y el cielo toma una tonalidad casi blanca, el factor tiene un valor mayor que 3; mientras que en ambientes con alta contaminación (con color grisáceo), puede llegar a valores entre 6 o 7. Estos valores se tabulan para cada emplazamiento en forma de 12 valores mensuales. En este caso de estudio el conjunto de valores son: {2.4, 2.7, 2.6, 3.1, 3.2, 3.5, 3.7, 3.9, 3.9, 2.9, 2.7, 2.3} [231].

La irradiancia teórica calculada con el modelo ESRA presenta un excelente resultado en los días soleados, como se muestra en la Figura 52. Sin embargo, este modelo no está diseñado para predecir situaciones de nubosidad, en los que no funciona bien.

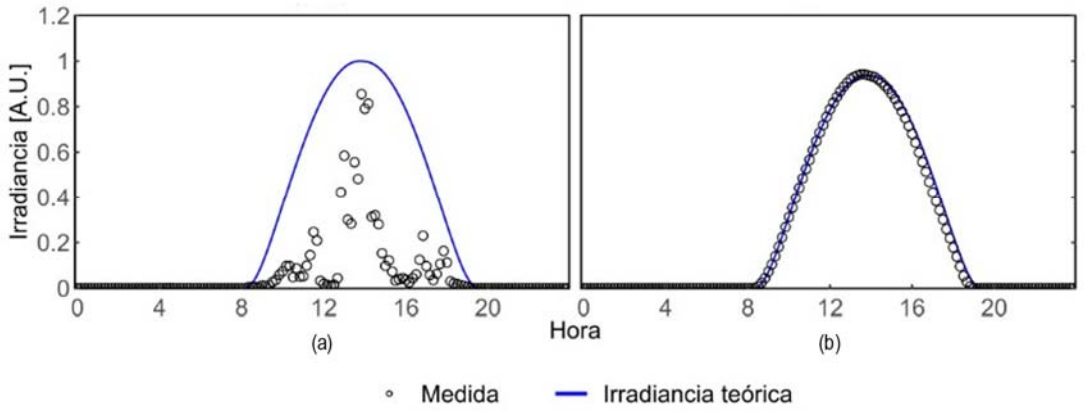


Figura 52. Comparación entre irradiancia teórica y medida en un día nublado (a) y soleado (b)

Tratamiento y Normalización

Antes de incorporar los datos meteorológicos en el modelo, se lleva a cabo un proceso de tratamiento con el propósito de garantizar la calidad y la coherencia de la información. Este proceso abarca diversas etapas, incluyendo la recolección de datos proveniente de la estación de medición y su sincronización con los datos del modelo ESRA, y la limpieza de posibles valores atípicos o inconsistentes. Este proceso es fundamental para asegurar que los modelos aprendan de manera efectiva y posteriormente puedan generar pronósticos adecuados en base a la información proporcionada.

Una vez definidas las variables de trabajo, se normalizan para que sus valores estén en el rango de 0 a 1, mediante la ecuación (22) de la Sección 3.2. A excepción de las variables hora y mes que requieren de un proceso especial. La distribución de los valores númericos de estas variables a lo largo del tiempo es un diente de sierra, que no tiene sentido para los modelos, ya que diciembre y enero, cuyas caracteristicas solares son similares presentan el valor más alto y más bajo respectivamente, y lo mismo sucede en las horas con el cambio de día. Aplicando una transformación adecuada, es posible conseguir un efecto en estas variables que se corresponda con el ciclo solar (ver Figura 53). Este proceso consta de dos pasos: primero se realiza un escalado de los valores de la variable en el rango de 0 a π; y segundo, se calcula el seno de estos valores siguiendo las ecuaciones correspondientes, y obteniendo así un comportamiento acorde con el comportamiento de la irradiancia (ver ecuación (25)).

$$hora_n = sen\left(\frac{hora \cdot \pi}{23}\right);\ mes_n = sen\left((mes-1)\cdot\frac{\pi}{11}\right) \quad (25)$$

En la figura Figura 53 se presentan los valores de las variables mes y hora, normalizados y sin normalizar.

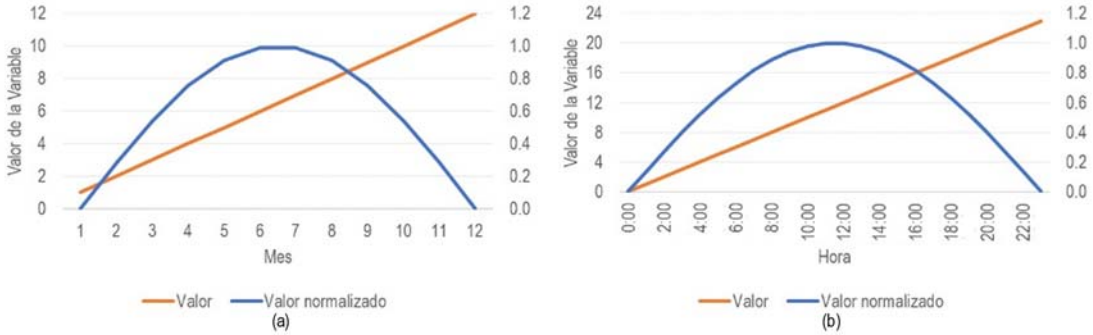


Figura 53. Normalización (a) Mes y (b) Hora

Selección de hiperparámetros y entrenamiento

Una vez normalizados los datos, se dividen de manera aleatoria en dos partes siguiendo un ratio 80/20 para tener un periodo de entrenamiento con el 80% de los datos y un periodo de prueba con el 20% de los datos. El periodo de entrenamiento en el cálculo de hiperparámetros es a su vez dividido de nuevo, en un ratio 80/20, para tener un periodo de validación en los métodos de búsqueda de hiperparámetros.

La búsqueda de hiperparámetros se realiza de manera análoga a la estimación de la producción, presentada en el Sección 3.3. Una vez definidos los valores de hiperparámetros, se entrenan los modelos para determinar sus parámetros internos. Con ello se dispone de cuatro modelos, uno para cada metodología, que permiten estimar la irradiancia.

Predicción de la irradiancia e intervalos de confianza

Con los modelos definidos es posible estimar la irradiancia. Los datos de prueba alimentan a los modelos, que generan predicciones basadas en los patrones aprendidos durante el entrenamiento.

La muestra de prueba se evalúa con cada uno de los modelos, calculando la estimación puntual y su intervalo de confianza. Para calcular el intervalo de confianza se verifica si la irradiancia tiene una distribución normal, tal y como ya se hizo para la producción de un inversor. Para ello se utilizan los métodos gráficos, como la representación de su distribución y la gráfica de cuantiles (ver Figura 54).

Al igual que sucedía con la producción de los inversores, la irradiancia tampoco tiene una distribución normal. Los intervalos de confianza se calcularán con las mismas técnicas utilizadas para el cálculo de los intervalos de confianza de la producción del inversor y presentados en la sección 2.6. Los resultados de todos los modelos se analizarán para evaluar cual de todos es más adecuado para esta aplicación. Todos los algoritmos también se han programado en R, “R programming language” [228].

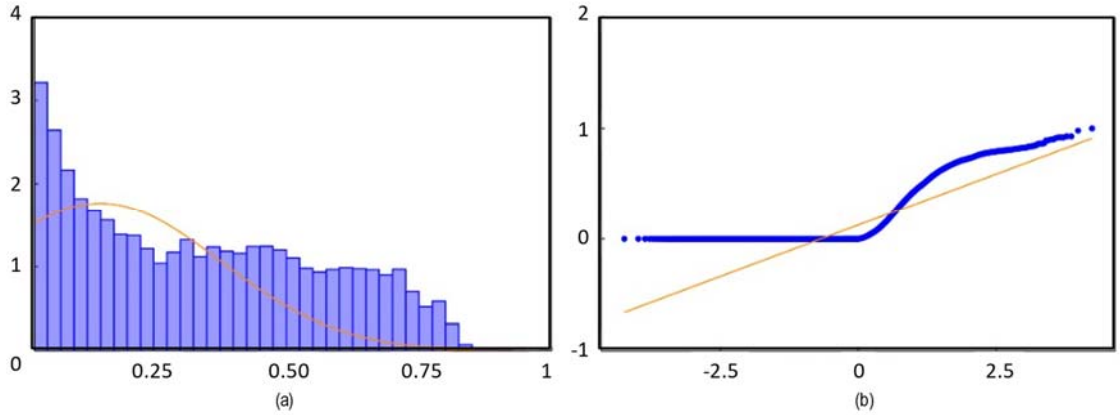


Figura 54. Histograma de la irradiancia (a) y grafica de cuantiles (b)

4.3. Aplicación de los algoritmos a la predicción de la irradiancia

Regresión Lineal Multivariable

El modelo "Multiple Linear Regression" [214], necesita definir los parámetros que son los coeficientes o pesos β_i que relacionan la irradiancia buscada con el resto de las variables que intervienen en el modelo, de acuerdo a la ecuación (26) que es una particularización de la ecuación (1) de la Sección 2.2 en la cual se tienen en cuenta todas las variables. Estos coeficientes se calculan durante el entrenamiento, minimizando el error mediante el método de mínimos cuadrados (ver Tabla 20) .

$$y_i = \beta_0+\beta_1\cdot IR_{teorica}+\beta_2\cdot T_{amb}+\beta_3\cdot Vel+\beta_4\cdot H+\beta_5\cdot B+\beta_6\cdot m+\beta_7\cdot h \quad (26)$$

Tabla 20. Parámetros del modelo MLR. Aplicación estimación de irradiancia

Término Independiente (β_0)	0.0673
Coeficiente de la Irradiancia teórica (β_1)	0.6050
Coeficiente de la Temperatura ambiente (β_2)	0.0414
Coeficiente de la Velocidad (β_3)	0.0304
Coeficiente de la Humedad (β_4)	-0.1972
Coeficiente de la Presión (β_5)	0.0839
Coeficiente del Mes (β_6)	0.0006
Coeficiente de la Hora (β_7)	-0.0013

La Tabla 20 presenta los coeficientes obtenidos tras el ajuste del modelo. Estos revelan que la irradiancia teórica es la variable de mayor relevancia para el modelo, como era de esperar, aun así, el bajo valor obtenido se debe fundamentalmente a la presencia de nubes y sombras en parte del periodo analizado. La humedad, con un coeficiente negativo que representa su relación con la irradiancia, es el segundo factor más influyente. Le siguen en importancia la temperatura y la velocidad del viento. La presión atmosférica ejerce la influencia más débil en la radiación, acorde con las observaciones previas del análisis de correlación en la Figura 51. En cuanto a las variables

temporales, hora y mes, su impacto es menor comparado con el de las variables climáticas

Los resultados de la estimación de la irradiancia con este modelo se presentan en la Figura 55. Esta figura está formada por dos gráficas, la gráfica (a) muestra, para un conjunto de observaciones aleatorias, la predicción del modelo junto con los valores de irradiancia medidos. En la gráfica de dispersión (b), se presenta la relación entre la irradiancia medida y la predicción en toda la muestra de prueba. Además, se incluye la función identidad y la línea de tendencia de la medida-predicción. La función identidad representa la precisión del modelo, cuanto más cerca estén los pares de puntos de la línea de identidad mejor será el modelo y menor su error. En cuanto a la línea de tendencia, representa la incertidumbre, cuanto más próximos a la línea están los pares de puntos menor es la incertidumbre.

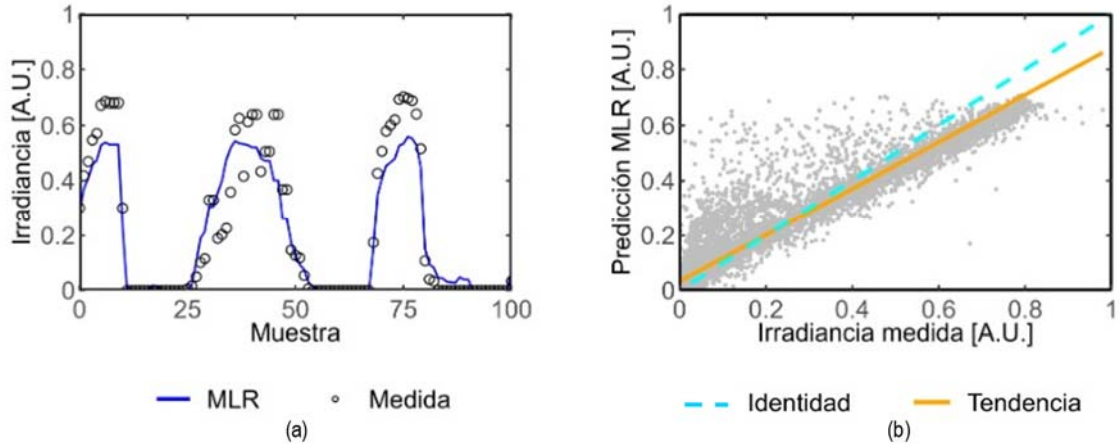


Figura 55. Muestra del resultado del MLR (a) y gráfica de dispersión del resultado del modelo MLR (b) . Aplicación estimación de irradiancia

Observando la Figura 55, se ve que el modelo no es capaz de simular la medida, en varias de las observaciones, siendo más acusado este efecto con irradiancias altas. La gráfica de dispersión nos indica que el resultado del MLR sigue la tendencia de la irradiancia, pero sus resultados son conservadores. El resultado es impreciso por su divergencia con la función identidad e incierto, por la dispersión de puntos respecto a la línea de tendencia.

El intervalo de confianza del resultado calculado con el algoritmo QR a los niveles de confianza de 90%, 95% y 99% se muestra en la Figura 56.

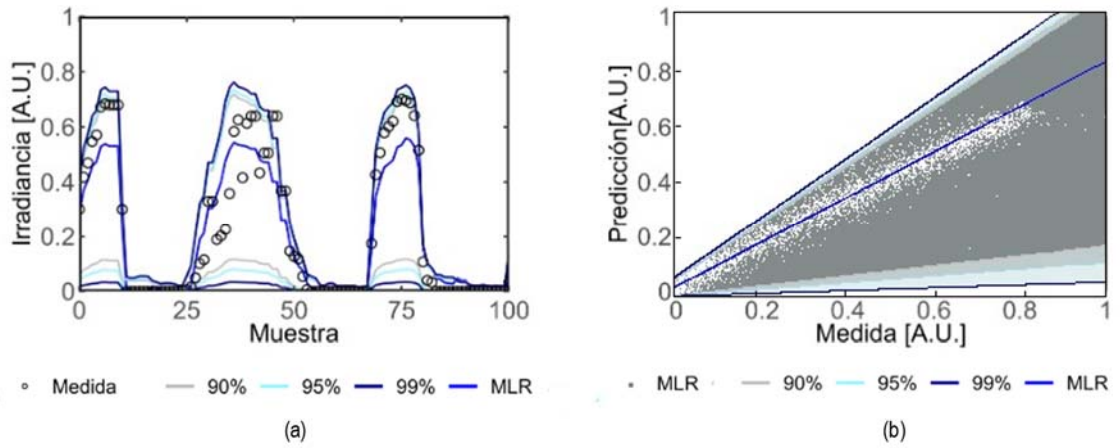


Figura 56. Muestra del intervalo de confianza del modelo MLR (a) y gráfica de dispersión con intervalo de confianza del resultado del modelo MLR (b) . Aplicación estimación de irradiancia

En la gráfica situada a la izquierda de la Figura 56, se muestra el mismo conjunto aleatorio escogido en la figura anterior incluyendo los intervalos de confianza y en la gráfica de la derecha se presenta el gráfico dispersión con todas las observaciones y las tendencias de los pares de datos medida y predicciones con el modelo y con los diferentes límites de confianza.

En ella se observa como los límites inferiores son bajos en cualquiera de los tres niveles de confianza.

"Random Forest"

El algoritmo RF [219], necesita ajustar los hiperparámetros que determinen el número de árboles (Nº Árboles) su tamaño (Profundidad) y como se conforman estos, número de variables (Nº Variables), y la tasa de muestreo (T. Muestreo) en cada árbol. El número de variables que intervienen en cada árbol se define con la regla de que el valor óptimo es un tercio del número de variables totales, y en este caso particular es 2. Este valor se deja fijo a partir de los resultados del Capítulo 3. El resto de hiperparámetros se definen mediante una búsqueda sistemática en red variando cada uno de ellos en un rango de valores predefinido según se muestra en la Tabla 21. Al prefijar el número de variables de cada árbol, en principio un único mallado puede cubrir todo el rango razonable de valores del resto de hiperparámetros, sin necesidad de realizar varias iteraciones.

Tabla 21. Rango de valores de los hiperparámetros y número de modelos. Barrido 1 (RF). Aplicación estimación de irradiancia

T. Muestreo	0.1-0.9 delta 0.1 y 0.97,0.98,0.99
Profundidad	1-32 delta 1
Nº Variables	2
Nº Árboles	10-1500 delta 10
Nº Modelos	57600

La representación gráfica del error de los resultados de todos los modelos se muestra en la Figura 57. En esta figura se representan 12 gráficas, cada gráfica corresponde a uno de los valores del hiperparámetro T. Muestreo. En cada gráfica se representa el error de los resultados de los modelos, eje de ordenadas, en función del número de árboles, eje de abscisas, indicando con diferente color la profundidad del árbol.

En la Figura 57, se observa como la adecuada combinación de hiperparámetros, puede disminuir el error de valores superiores al 15% a inferiores al 5%. También se puede ver que es posible alcanzar un error mínimo similar, con diferentes combinaciones de hiperparámetros. Por último, de la gráfica se deduce que la T. Muestreo que minimiza el error está entre los valores 0.6-0.9, combinado con una profundidad superior a 20.

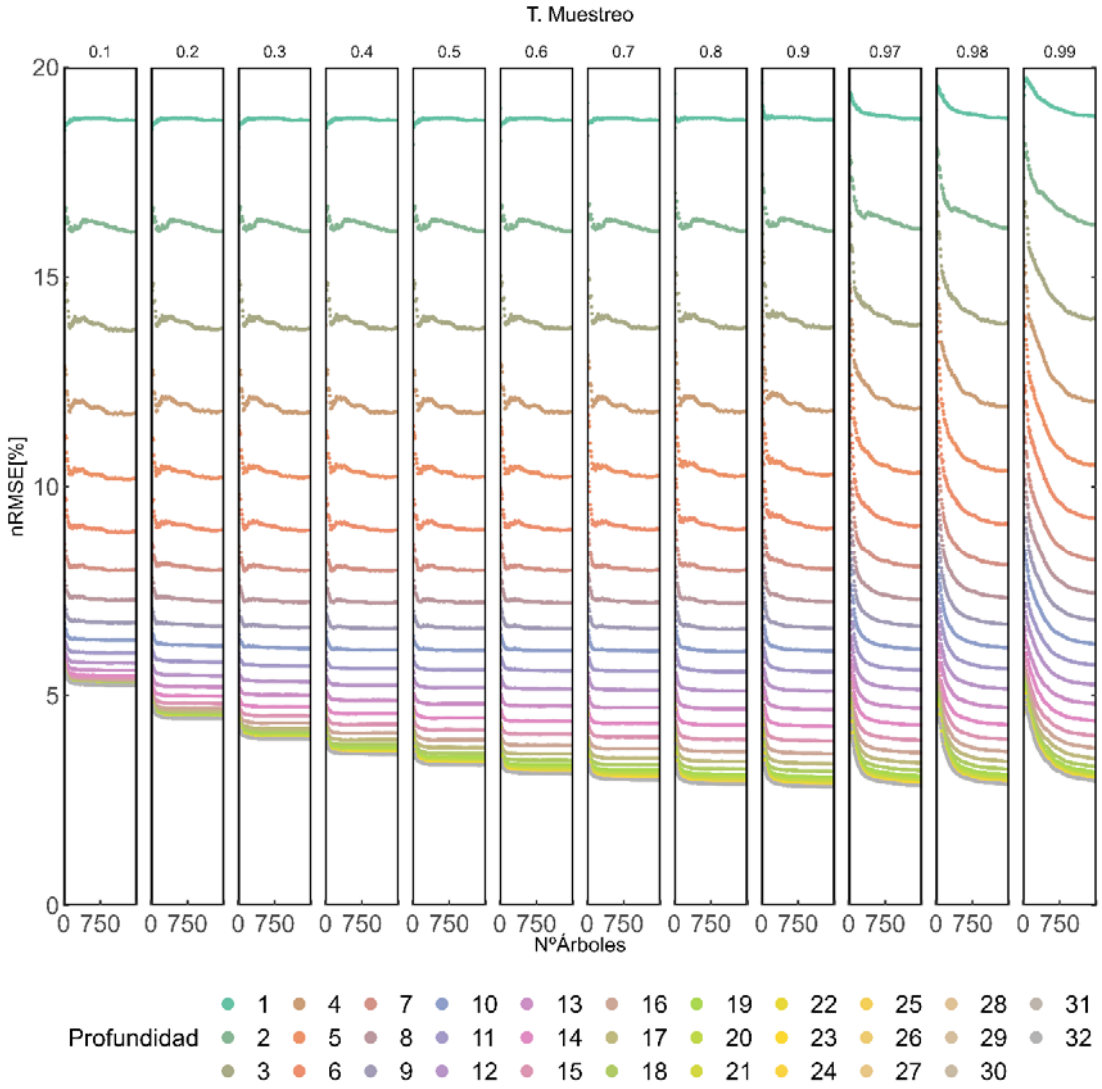


Figura 57. Representación del error de todos los modelos generados en la búsqueda. (RF)

Para poder estudiar los valores con mayor precisión, se realiza una gráfica ampliada y focalizada en un rango de valores, (ver Figura 58). En esta nueva gráfica se incluye también el resultado del error en el periodo de validación. El objetivo de analizar conjuntamente los errores de los modelos en el periodo de entrenamiento y en el periodo de validación, es observar si existe un sobre-ajuste en los modelos con algunas combinaciones de hiperparámetros.

La Figura 58 está formada por ocho gráficas distribuidas en dos filas, en las cuatro gráficas de la fila superior se presentan los resultados de los modelos aplicados en la muestra de entrenamiento, y en las cuatro gráficas situadas en la parte inferior, los resultados de los modelos de la muestra de validación. Al igual que en la figura anterior cada gráfica corresponde a un valor diferente del hiperparámetro T. Muestreo, con diferente color se indica la profundidad y en el eje de las abscisas se presenta el número de árboles.

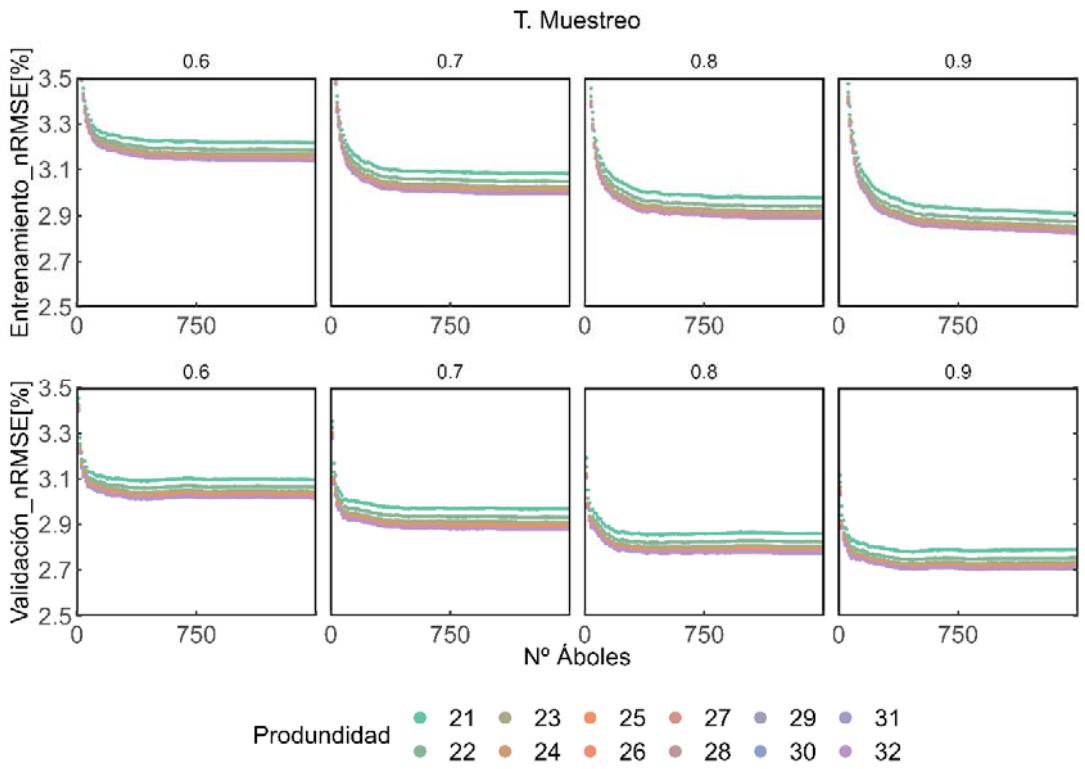


Figura 58. Detalle del error de una muestra de los modelos generados en la búsqueda. (RF, Profundidad mayor de 20, y T. Muestreo de 0.6 a 0.9). Aplicación estimación de irradiancia

Comparando los resultados de ambas filas se observa que para los mismos hiperparámetros los errores tienen la misma magnitud, independientemente de la muestra de trabajo, lo que demuestra que no existe un sobre-ajuste en este rango de hiperparámetros.

En la Figura 58, se aprecia que el valor óptimo de T. Muestreo, que minimiza el error, es 0.9. Respecto a los valores de profundidad y del número de árboles, se observa que la profundidad debe ser superior a 24 y el número de árboles superior a 500. Además, 1500 árboles no son suficientes para lograr la estabilidad del modelo. Para asegurar que las variaciones en el error, al variar el número de árboles y la profundidad, se mantengan por debajo de un valor umbral establecido, (ver ecuación (24)), se realiza una segunda búsqueda de hiperparámetros. Esta búsqueda se lleva a cabo de acuerdo con los valores presentados en la Tabla 22.

Tabla 22. Rango de valores de los hiperparámetros y número de modelos. Barrido 2 (RF). Aplicación estimación de irradiancia

T. Muestreo	0.9
Profundidad	24-32 delta 1
Nº Variables	2
Nº Árboles	1500-2000 delta 10
Nº Modelos	450

El error de los modelos con hiperparámetros definidos por el número de variables igual a 2, la tasa de muestreo 0.9, profundidad entre 24 y 32 y número máximo de árboles de 2000 se presentan en la Figura 59.

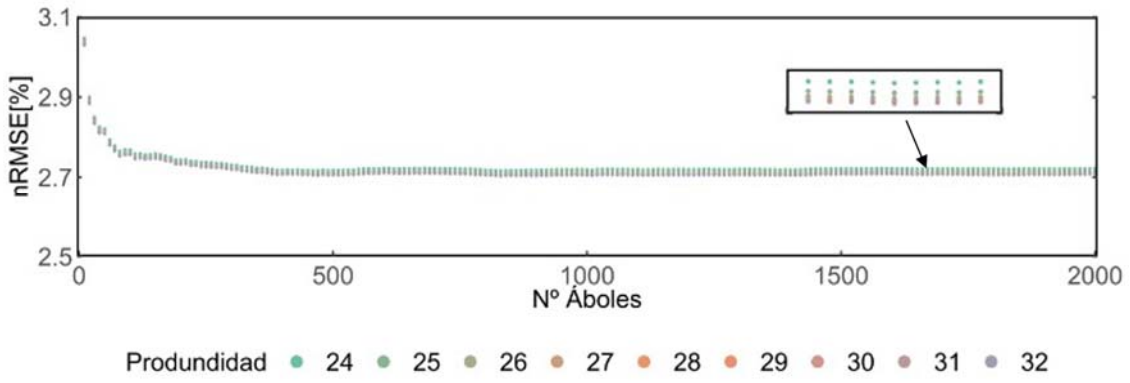


Figura 59. Detalle del error de una muestra de los modelos generados en la búsqueda. Barrido 2 (RF, Profundidad mayor de 24, Nº Variables 2 y T. Muestreo 0.9). Aplicación estimación de irradiancia

En la Tabla 23, se presentan los errores nRMSE de 279 posibles modelos, para analizar conjuntamente el valor de la profundidad y número de los árboles. En dicha Tabla 23, la región sombreada en verde corresponde a los casos estudiados cuyas combinaciones de hiperparámetros presentan menor error. Por otro lado, los casos que se encuentran en la región roja son aquellos que presentan un error más elevado. En esta tabla se observa que la profundidad adecuada es 29, y el número de árboles 1650.

Los valores de hiperparámetros quedan definidos de acuerdo con la Tabla 24, indicándose además el número de modelos que han intervenido en la selección.

Una vez definidos los hiperparámetros se entrena el modelo para definir los parámetros internos y se calcula el resultado para la muestra de prueba. El resultado se presenta en la Figura 60. Esta figura está formada por dos gráficas, la gráfica (a) representa el comportamiento del modelo sobre el mismo conjunto aleatorio de datos que el representado en la Figura 55, con el ánimo de que puedan ser comparados y la gráfica de dispersión (b), representa la relación de la irradiancia medida y la estimada con el modelo RF para todo el periodo de prueba, junto con su tendencia y la línea de identidad.

En esta Figura 60, se observa como el modelo RF se comporta ligeramente mejor que el modelo MLR. Pese a ello existen estimaciones, las correspondientes a los valores más altos, que no se ajustan bien a las medidas de la irradiancia. Este hecho queda corroborado en la gráfica de dispersión, en la cual se observa la tendencia diferente a altas irradiancias, es decir, al modelo le falta precisión a altas irradiancias. También se observa una dispersión de puntos asimétrica, correspondiente a valores de bajas y medias irradiancias, en los cuales el modelo ha dado una predicción superior a la medida.

Tabla 23. Error nRMSE. Selección de hiperparámetros de los modelos RF. Aplicación estimación de irradiancia

Nº Árboles \ Profundidad	24	25	26	27	28	29	30	31	32
500	2.875	2.867	2.864	2.863	2.862	2.862	2.861	2.861	2.861
550	2.869	2.862	2.858	2.856	2.856	2.855	2.855	2.855	2.855
600	2.868	2.861	2.857	2.856	2.855	2.855	2.854	2.854	2.854
650	2.865	2.858	2.854	2.852	2.852	2.851	2.851	2.851	2.851
700	2.864	2.857	2.852	2.851	2.850	2.850	2.850	2.850	2.849
750	2.862	2.855	2.851	2.849	2.849	2.848	2.848	2.848	2.848
800	2.859	2.852	2.848	2.847	2.846	2.846	2.845	2.845	2.845
850	2.858	2.851	2.847	2.846	2.845	2.845	2.844	2.844	2.844
900	2.855	2.848	2.844	2.842	2.842	2.841	2.841	2.841	2.841
950	2.855	2.848	2.844	2.842	2.842	2.841	2.841	2.841	2.841
1000	2.855	2.848	2.844	2.842	2.841	2.841	2.841	2.840	2.840
1050	2.854	2.846	2.842	2.841	2.840	2.839	2.839	2.839	2.839
1100	2.852	2.845	2.841	2.839	2.838	2.838	2.837	2.837	2.837
1150	2.849	2.842	2.838	2.836	2.835	2.834	2.834	2.834	2.834
1200	2.847	2.841	2.836	2.835	2.834	2.833	2.833	2.833	2.833
1250	2.847	2.840	2.836	2.834	2.833	2.833	2.832	2.832	2.832
1300	2.844	2.837	2.834	2.832	2.831	2.830	2.830	2.830	2.830
1350	2.843	2.836	2.833	2.831	2.830	2.829	2.829	2.829	2.829
1400	2.840	2.833	2.829	2.828	2.827	2.826	2.826	2.826	2.826
1450	2.841	2.834	2.830	2.828	2.827	2.827	2.826	2.826	2.826
1500	2.839	2.832	2.828	2.827	2.826	2.825	2.825	2.825	2.825
1550	2.839	2.832	2.828	2.826	2.825	2.825	2.824	2.824	2.824
1600	2.838	2.830	2.827	2.825	2.824	2.823	2.823	2.823	2.823
1650	2.836	2.829	2.825	2.823	2.823	2.822	2.822	2.822	2.822
1700	2.836	2.829	2.825	2.823	2.823	2.822	2.822	2.822	2.822
1750	2.837	2.830	2.826	2.824	2.823	2.823	2.823	2.823	2.823
1800	2.837	2.830	2.826	2.824	2.823	2.823	2.823	2.823	2.823
1850	2.836	2.829	2.825	2.824	2.823	2.823	2.823	2.823	2.823
1900	2.838	2.831	2.827	2.825	2.824	2.824	2.824	2.824	2.824
1950	2.839	2.832	2.828	2.826	2.825	2.825	2.824	2.824	2.824
2000	2.839	2.832	2.828	2.826	2.825	2.825	2.825	2.825	2.824

Tabla 24. Hiperparámetros óptimos y números de modelos para su obtención (RF). Aplicación estimación de irradiancia

T. Muestreo	0.9
Profundidad	29
Nº Variables	2
Nº Árboles	1650
Nº Modelos	58050

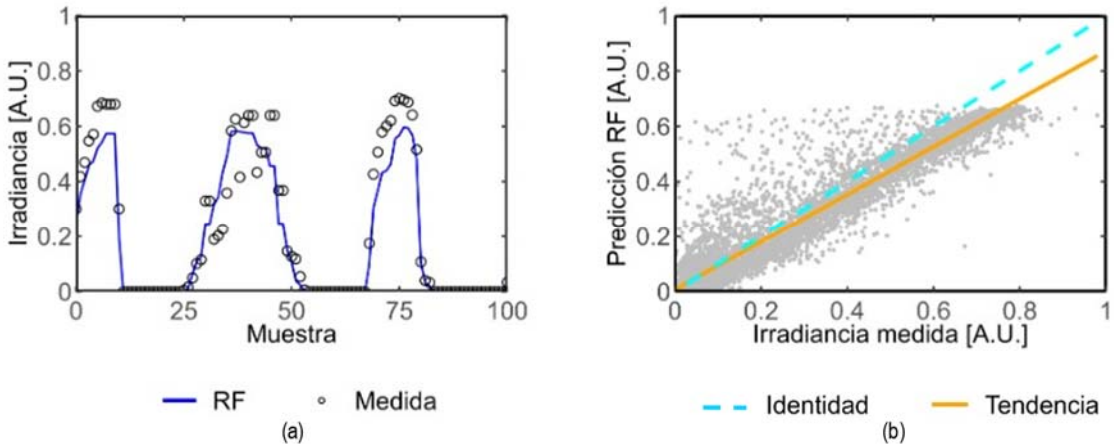


Figura 60. Muestra del resultado RF (a) y gráfica de dispersión del resultado RF (b) . Aplicación estimación de irradiancia

A continuación, se calcula el intervalo de confianza del resultado con el modelo QRF, para los intervalos de confianza del 90%, 95% y 99%, y los resultados se muestran en la Figura 61:

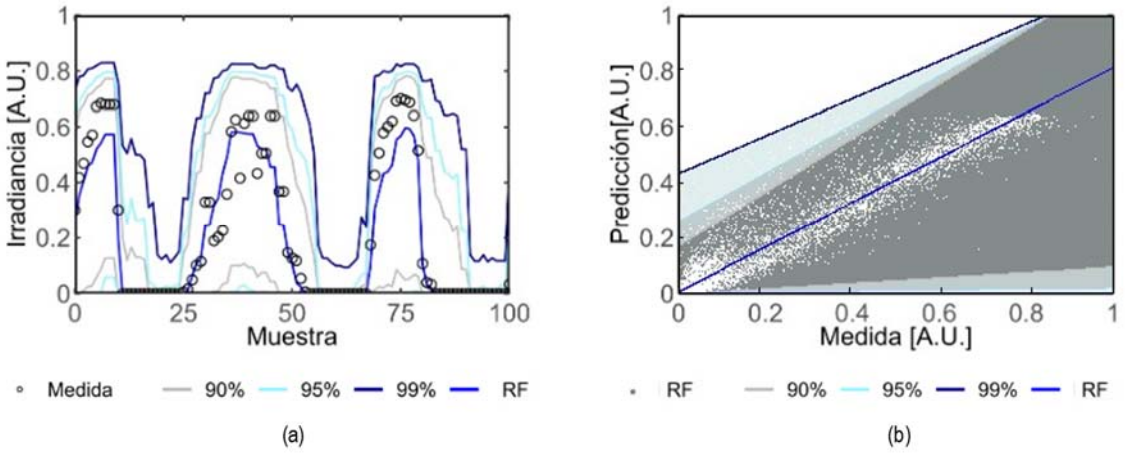


Figura 61. Muestra del intervalo de confianza del RF (a) y gráfica de dispersión con intervalo de confianza del resultado RF (b). Aplicación estimación de irradiancia

La Figura 61 muestra en la gráfica de la izquierda los intervalos en un conjunto de datos y en la gráfica de la derecha la muestra entera con la tendencia de sus intervalos. Los límites inferiores quedan en los valores muy bajos o prácticamente nulos y los límites superiores a medida que se incrementa el límite de confianza se amplían a bajas irradiancias y convergen a altas irradiancias. Se puede concluir que este modelo no tiene un buen comportamiento.

"Gradient Boosting"

El algoritmo GB [220] también requiere hiperparámetros, prácticamente iguales a los del algoritmo RF. Estos se muestran en la Tabla 25 junto con sus posibles valores. Para ajustarlos, se utiliza la técnica de búsqueda en red, realizando un único barrido con un mallado que cubre los valores predefinidos de los hiperparámetros. El mallado propuesto en este caso genera 74700 modelos diferentes.

Tabla 25. Rango de valores de los hiperparámetros y número de modelos (GB). Aplicación estimación de irradiancia

T. Muestreo	0.1-1 delta 0.1
Profundidad	1-30 delta 1
Nº Árboles	1-1235 delta 5
Nº Modelos	74700

La Figura 62. muestra la representación gráfica de los errores de estos modelos. Esta figura consta de diez gráficas, cada una correspondiente a un valor diferente de T. Muestreo. En cada gráfica, el eje de ordenadas representa el error de los modelos, el eje de abscisas representa el número de árboles y los diferentes colores indican las profundidades de los árboles.

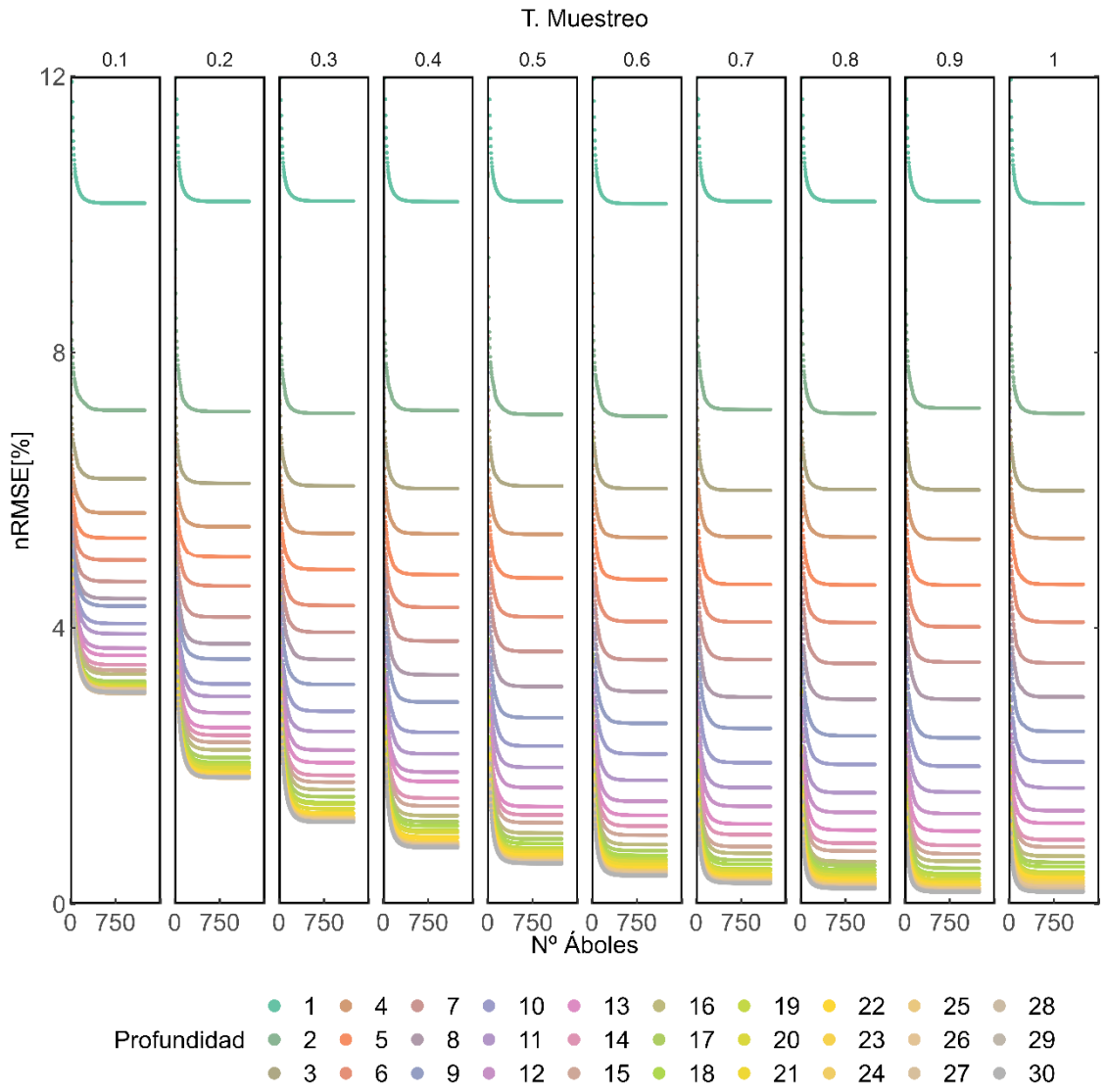


Figura 62. Representación del error de todos los modelos generados en la búsqueda. (GB). Aplicación estimación de irradiancia

En la Figura 62, se puede observar que a medida que aumenta el valor de T. Muestreo y la profundidad del árbol, el error disminuye. Esta disminución de errores es más pronunciada que la observada con RF y los bajos valores alcanzados por el error en el entrenamiento sugieren que podría haber un sobreajuste del modelo.

Para verificar si ha habido un sobre-entrenamiento en algunos de los modelos, se ha generado una nueva figura centrada en valores de T. Muestreo superiores a 0.5 y valores de Profundidad superiores a 5 (ver Figura 63). Esta figura consta de doce gráficas distribuidas en dos filas. En la fila superior se muestran las seis gráficas correspondientes a la muestra de entrenamiento y en la fila inferior las gráficas de la muestra de validación. Las gráficas se diferencian entre sí por el valor de T. Muestreo, y cada una tiene en el eje de abscisas el número de árboles, y en el eje de ordenadas el valor del error de los modelos, indicando con diferentes colores la profundidad del árbol.

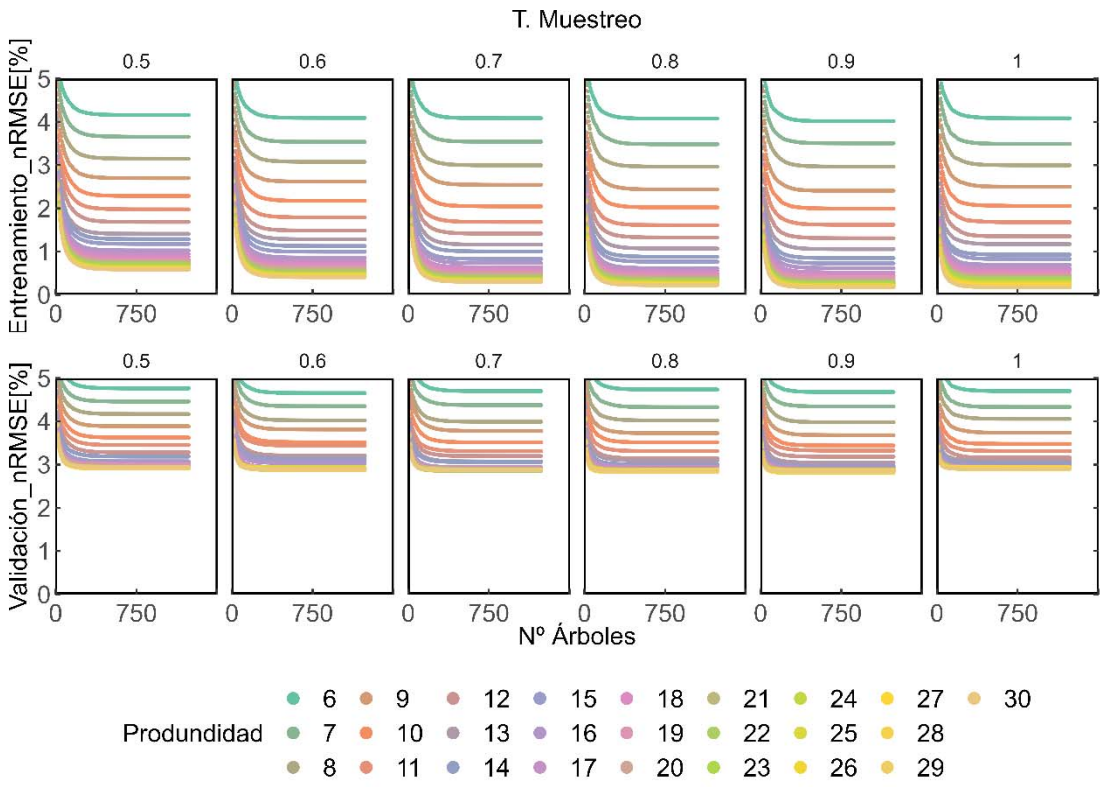


Figura 63. Detalle del error de los modelos generados en la búsqueda (GB, Profundidad mayor de 5, T. Muestreo mayor de 5). Aplicación estimación de irradiancia

Al analizar los errores de los modelos con idénticos hiperparámetros aplicados a diferentes muestras, se observa que, en la muestra de entrenamiento, algunas combinaciones de hiperparámetros reducen el error hasta valores muy pequeños. Sin embargo, en la muestra de validación, el error se mantiene en un valor asintótico ligeramente inferior al 3%. Esto se debe a que el aumento de los valores de T. Muestreo y Profundidad no mejora realmente el modelo, sino que lo sobre-ajusta.

La Figura 63 identifica los valores de hiperparámetros que minimizan el error sin sobre-entrenar el modelo. El valor óptimo de T. Muestreo se define como 0.5, por ser el mínimo valor a partir del cual no existe mejora sustancial en el error del modelo. La Profundidad óptima está en un rango de valores entre 6 y 12, para no tener sobre-entrenamiento y se recomienda un número de árboles superior a 500, para tener estabilidad en el error del modelo. La decisión final sobre los valores óptimos de Profundidad y Nº Árboles se toma empíricamente, estudiando todas las opciones posibles que minimicen el error sin causar un sobre-ajuste, como se muestra en la Tabla 26.

La Tabla 26 muestra los errores nRMSE de 119 modelos de combinaciones de hiperparámetros predefinidos. El criterio para seleccionar los hiperparámetros óptimos es buscar el valor menor de profundidad del árbol que no permita una diferencia sustancial entre los errores de entrenamiento y validación, controlando así el sobreajuste. Este criterio establece el valor óptimo de Profundidad en 9. Una vez definido este valor, el número óptimo de árboles se determina por el principio de estabilidad del modelo.

Tabla 26. Error nRMSE. Selección de hiperparámetros de los modelos GB. Aplicación estimación de irradiancia

Nº de Arboles \ Profundidad	6	7	8	9	10	11	12
400	4.772	4.467	4.178	3.895	3.633	3.460	3.292
450	4.768	4.464	4.174	3.891	3.629	3.456	3.288
500	4.766	4.461	4.172	3.889	3.627	3.454	3.286
550	4.765	4.460	4.171	3.887	3.626	3.452	3.284
600	4.764	4.459	4.170	3.887	3.625	3.452	3.284
650	4.763	4.459	4.169	3.886	3.624	3.451	3.283
700	4.763	4.458	4.169	3.886	3.624	3.451	3.283
750	4.762	4.458	4.169	3.886	3.624	3.451	3.283
800	4.762	4.458	4.168	3.886	3.624	3.450	3.283
850	4.762	4.458	4.168	3.885	3.624	3.450	3.282
900	4.762	4.458	4.168	3.885	3.624	3.450	3.282
950	4.762	4.458	4.168	3.885	3.624	3.450	3.282
1000	4.762	4.458	4.168	3.885	3.624	3.450	3.282
1050	4.762	4.458	4.168	3.885	3.624	3.450	3.282
1100	4.762	4.458	4.168	3.885	3.624	3.450	3.282
1150	4.762	4.458	4.168	3.885	3.624	3.450	3.282
1200	4.762	4.458	4.168	3.885	3.624	3.450	3.282

El resumen de los valores óptimos de los hiperparámetros para el modelo GB se presenta en la Tabla 27.

Tabla 27. Hiperparámetros óptimos y número de modelos para su obtención (GB). Aplicación estimación de irradiancia

T. Muestreo	0.5
Profundidad	9
Nº Árboles	850
Nº Modelos	74700

Entrenando el modelo GB con estos hiperparámetros, se calculan los parámetros internos del modelo y una vez definido este se calcula la predicción de la irradiancia para la muestra de prueba. El resultado obtenido se presenta en la Figura 64. Al igual que con los modelos anteriores la figura contiene dos gráficas, la parte izquierda muestra un ejemplo de su comportamiento en unas observaciones aleatorias y la parte derecha, muestra la gráfica de dispersión, para analizar su tendencia en conjunto.

El ajuste del modelo es mejor que los modelos estudiados con anterioridad, tanto en las observaciones puntuales tomadas como ejemplo, como en la figura de dispersión. La gráfica de dispersión presenta junto con los pares de puntos, la tendencia de estos y la función identidad. Esta figura permite observar como la tendencia es prácticamente la línea de identidad, mostrando así la precisión de este modelo, y la dispersión de los puntos indica que la incertidumbre del modelo es también pequeña.

A continuación, se calcula el intervalo de la incertidumbre del modelo para cada uno de los niveles de confianza, 90%, 95% y 99%, y se presenta el resultado de la muestra del periodo de prueba en la Figura 65.

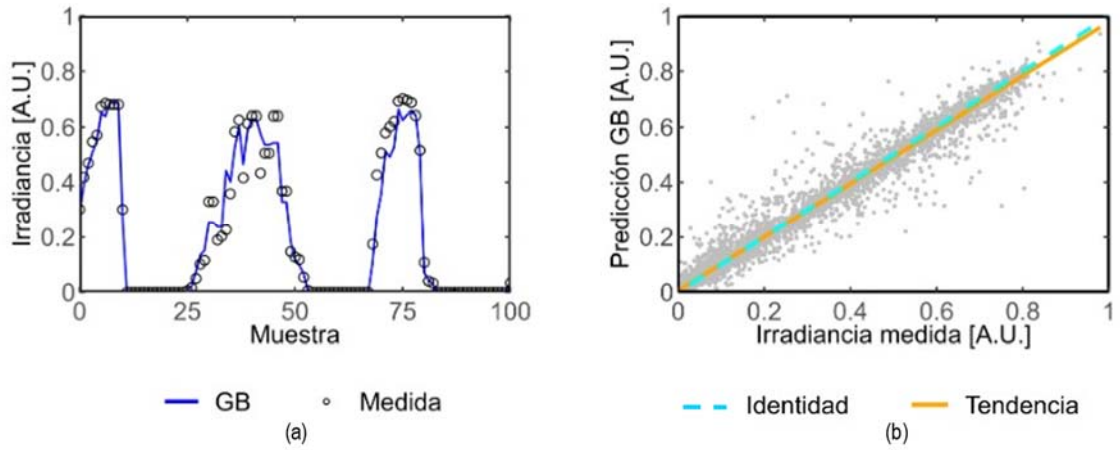


Figura 64. Muestra del resultado GB (a) y gráfica de dispersión del resultado GB (b). Aplicación estimación de irradiancia

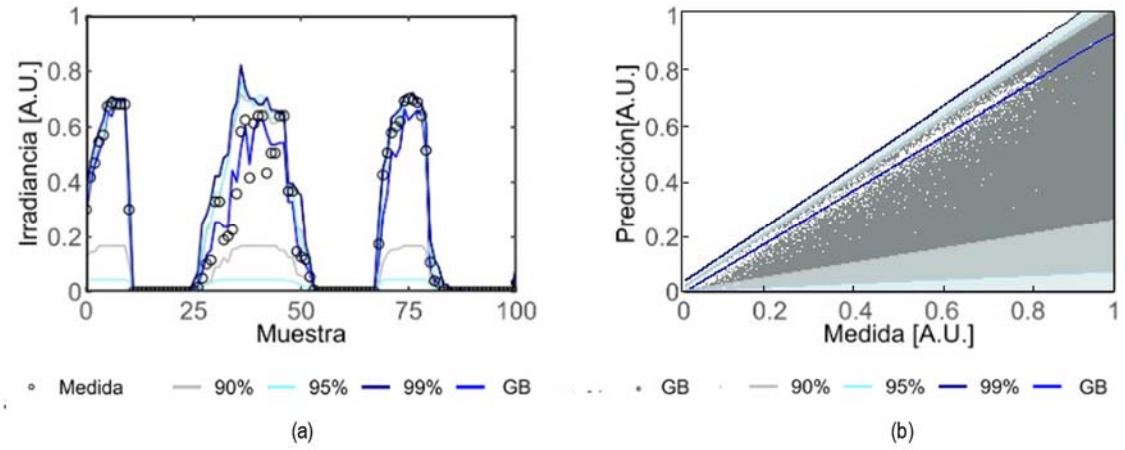


Figura 65. Muestra del intervalo de confianza del GB (a) y gráfica de dispersión con intervalo de confianza del resultado GB (b). Aplicación estimación de irradiancia

La amplitud de los intervalos es más estrecha que en los modelos anteriores, hecho que significa que el modelo tiene una baja incertidumbre como ya se deducía en la Figura 64. La precisión y baja incertidumbre del resultado obtenido con este modelo junto con las características de su intervalo de confianza hacen que el modelo GB sea una buena opción para estimar la irradiancia. Además, los hiperparámetros definidos para conseguir estos resultados corresponden a valores bajos que hacen que el modelo sea muy rápido de ejecutar y con bajas exigencias computacionales.

Redes neuronales

En la red neuronal MPNN [66], se optimizan sus hiperparámetros utilizando el método de hiper-bandas. Los hiperparámetros y sus valores se presentan en la Tabla 28

El método de hiper-bandas [186] requiere un criterio de parada predefinido por el número de pasadas del modelo, y por η, que al igual que en el caso de la estimación de la generación son 81 y 3 respectivamente (ver Sección 3.3). El número de configuraciones generadas con estos parámetros es 10550. Se han llevado a cabo tres barridos con diferentes semillas aleatorias, extrayendo de cada uno de ellos los 3400 mejores modelos (10200 modelos). Además, para hacer

la búsqueda más amplia, el método se aplica la metodología a modelos con el mismo número de capas ocultas. Así, se llevan a cabo tantas pruebas como modelos con diferentes capas se deseen generar, y en cada una de las pruebas se extraerán 10200 modelos. Se ha predefinido un máximo de 3 capas ocultas en los modelos, lo que da lugar a tres pruebas: una generando modelos con una capa oculta, otra con dos capas ocultas y una tercera con tres capas ocultas. De esta manera, se analizarán un total de 30600 modelos de redes neuronales.

Tabla 28. Rango de valores de los hiperparámetros y número de modelos. Método de Hiper-banda (ANN). Aplicación estimación de irradiancia

Prueba	1	2	3
Nº Capas	1	2	3
T. Aprendizaje	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001	0.01, 0.001, 0.0001
T. Descarte inicial	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1
T. Descarte Capa 1	0-0.9 delta 0.1	0-0.9 delta 0.1	0-0.9 delta 0.1
T. Descarte Capa 2	-	0-0.9 delta 0.1	0-0.9 delta 0.1
T. Descarte Capa 3	-	-	0-0.9 delta 0.1
Nº Neuronas Capa 1	2-512 delta 2	2-512 delta 2	2-512 delta 2
Nº Neuronas Capa 2	-	2-512 delta 2	2-512 delta 2
Nº Neuronas Capa 3	-	-	2-512 delta 2
Nº Modelos	31650 (extraídos 10200)	31650 (extraídos 10200)	31650 (extraídos 10200)

La representación del error de los modelos analizados en la búsqueda, en función de la variación de los hiperparámetros se muestra en la Figura 66 en tres gráficas diferentes. La primera de ellas corresponde a los modelos que tienen una capa oculta, la segunda a los modelos que tienen dos capas ocultas y la tercera gráfica a los modelos que tienen tres capas ocultas. En cada una de estas gráficas se representa con diferente color la tasa de descarte inicial, el error de los resultados de los modelos en el eje ordenadas, y en el eje abscisas el número de neuronas de la primera capa oculta.

Observando la Figura 66 se puede ver el amplio rango de errores que recorren los modelos analizados en la búsqueda y que además existe un valor del error mínimo que alcanzan los modelos estable e independiente del número de neuronas, por lo que el rango de los hiperparámetros y el número de modelos analizados son suficientes para determinar una combinación óptima. También se observa que los modelos de tres capas no mejoran los resultados de los modelos de dos capas, siendo estos mucho más complejos por lo que tampoco es necesario ampliar la búsqueda a modelos con más capas ocultas.

La estratificación de colores indica claramente que el valor del hiperparámetro T. Descarte inicial debe ser nulo, es decir todas las observaciones se incluirán en el modelo inicialmente. La nube de puntos indica que los mejores modelos de dos capas tienen un error ligeramente inferior a los mejores modelos de una capa, sin embargo, los modelos de tres capas tienen un error muy similar a los de dos y aplicando el principio de simplicidad de los modelos, elegimos los modelos de dos capas por ser la mejor opción para predecir la irradiancia.

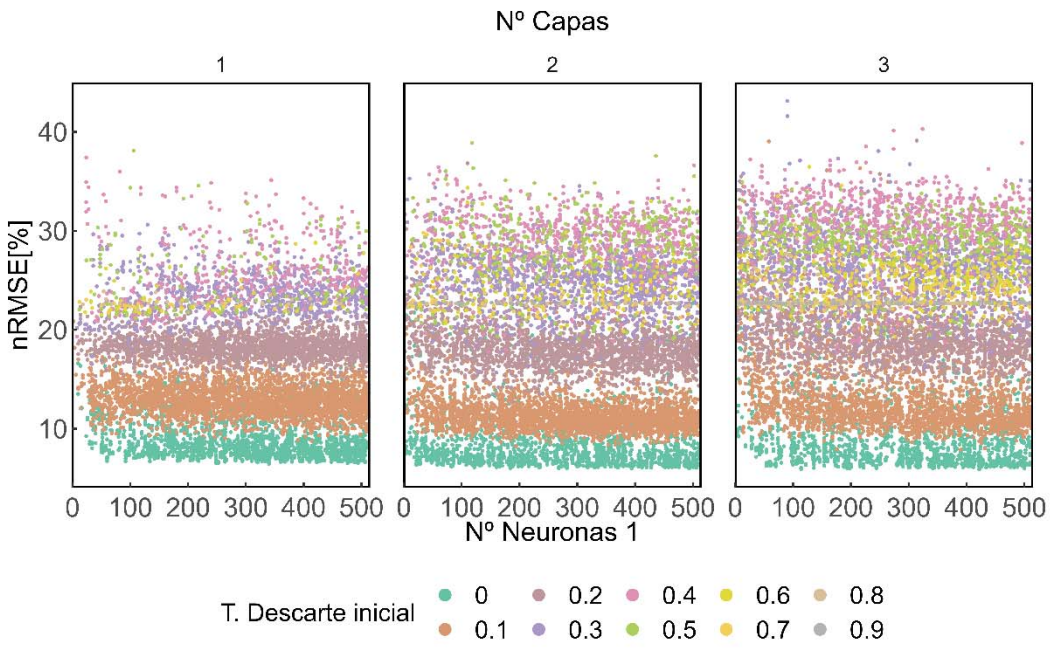


Figura 66. Representación del error de los modelos generados en la búsqueda. Método de Hiper-bandas. (ANN). Aplicación estimación de irradiancia

Centrando el análisis en los modelos de dos capas sin T. Descarte inicial, se dibuja una nueva figura, (ver Figura 66), en la que se representan los modelos clasificados en tres gráficas según el valor de la tasa de aprendizaje con la que han sido entrenados. En el eje de ordenadas se presenta el error normalizado de los modelos, en el eje de abscisa el número de neuronas de la primera capa oculta y con color se identifica el valor del hiperparámetro T. Descarte de la primera capa oculta.

En la Figura 67, se observa que la tasa de aprendizaje que minimiza el error corresponde al valor de 0.001, pero no se aprecia con claridad el valor que debe de tener la tasa de descarte de la primera capa oculta.

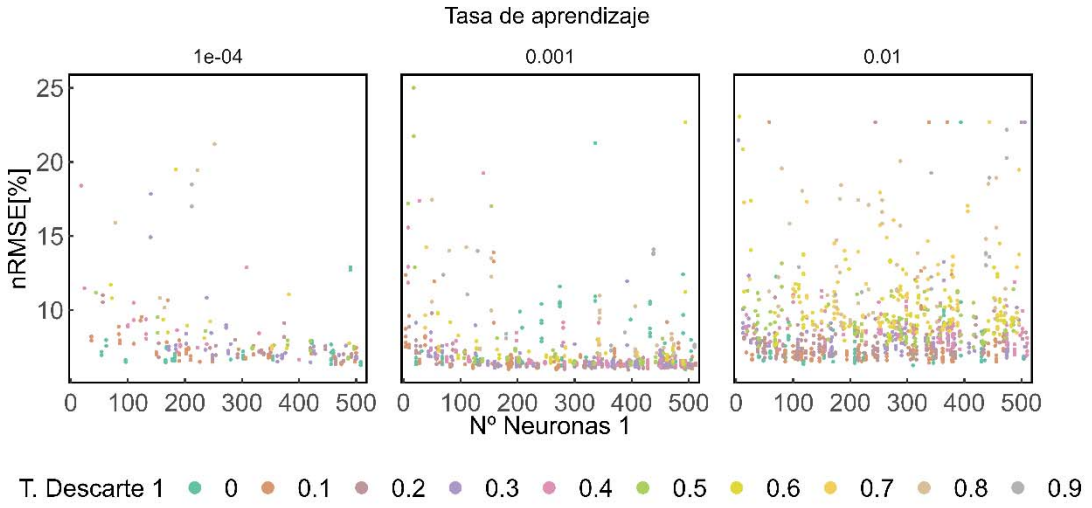


Figura 67. Detalle del error de los modelos generados en la búsqueda. Método de Hiper-bandas (ANN, modelos de dos capas y T. Descarte inicial nula). Aplicación estimación de irradiancia

Para analizar el valor de las tasas de descarte de manera específica se realiza la Figura 68, en la

que se representan solo los modelos de dos capas, con tasa de descarte nula y tasa de aprendizaje de 0.001. Esta figura se divide en 10 gráficas diferentes, y en cada una de ellas se representan los modelos con un valor en el hiperparámetro T. Descarte de la primera capa oculta concreto. En el eje de abscisas se indica el número de neuronas de la primera capa y en el eje de ordenadas el error normalizado de los modelos. Además, con diferente color se muestra el hiperparámetro T. Descarte de la segunda capa oculta.

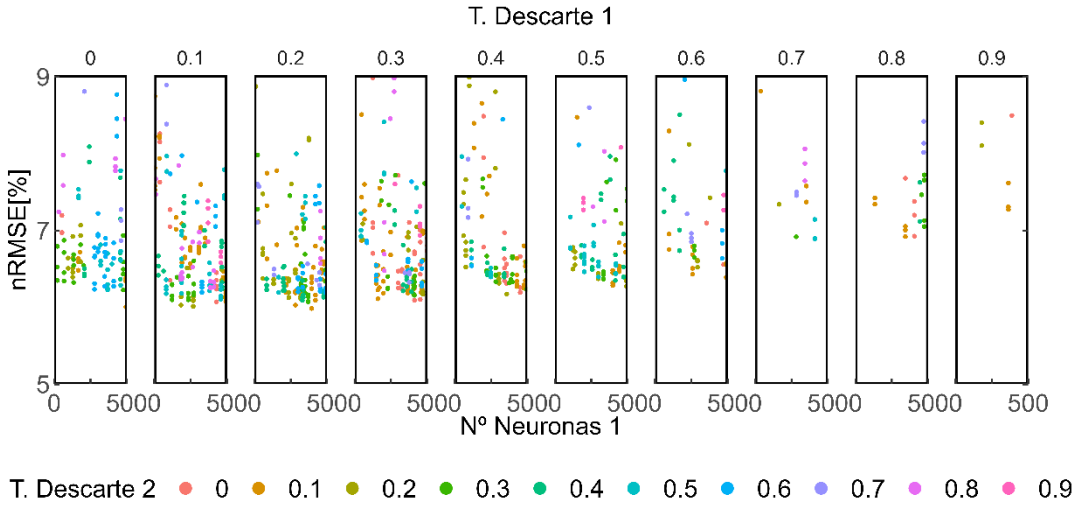


Figura 68. Detalle del error de los modelos generados en la búsqueda. Método de Hiper-bandas (ANN, modelos de dos capas, T. Aprendizaje de 0.001 y T. Descarte inicial nula). Aplicación estimación de irradiancia

En la Figura 68 tampoco se puede determinar un valor concreto de T. Descarte de la primera capa, pero sí que se observa que se alcanza un menor error con los valores inferiores o iguales a 0.3. En el caso de la tasa de descarte de la segunda capa, tampoco se identifica con claridad un valor concreto, pero se percibe que tiene que ser menor o igual al valor 0.4.

El número de modelos de 2 capas, tasa de descarte inicial nula y con bajas tasas de descarte en las capas ocultas, seleccionados con el método de hiper-bandas, es insuficiente para proseguir con el análisis de los demás hiperparámetros. Para incrementar la cantidad de modelos dentro de los rangos predefinidos de los hiperparámetros, se ha llevado a cabo un nuevo barrido. Este se realiza mediante una búsqueda aleatoria utilizando los valores de hiperparámetros descritos en la Tabla 29.

Tabla 29. Rango de valores de los hiperparámetros y número de modelos. Método de búsqueda aleatoria (ANN). Aplicación estimación de irradiancia

Prueba	1
Nº Capas	2
T. Aprendizaje	0.001
T. Descarte inicial	0
T. Descarte Capa 1	0-0.3 delta 0.1
T. Descarte Capa 2	0-0.4 delta 0.1
Nº Neuronas Capa 1	2-512 delta 2
Nº Neuronas Capa 2	2-512 delta 2
Nº Modelos	12000

Los modelos de dos capas de ambos métodos de búsqueda sin T. Descarte inicial, con T. Descarte de la primera capa oculta menor o igual de 0.3, con T. Descarte de la segunda capa inferior o igual a 0.4 y T. aprendizaje de 0.001, se muestran en la Figura 69. Esta figura es análoga a la

Figura 68. Detalle del error de los modelos generados en la búsqueda. Método de Hiper-bandas (ANN, modelos de dos capas, T. Aprendizaje de 0.001 y T. Descarte inicial nula). Aplicación estimación de irradiancia

, por lo que se representan en diferentes gráficas los modelos con diferente tasa de descarte en la primera capa. En este caso, el número de gráficas se recude a cuatro correspondiendo a cada uno de los valores considerados. En el eje de ordenadas se representa el error de los modelos y en el de abscisas los valores de neuronas de la primera capa oculta; con diferente color se indica los posibles valores de T. Descarte de la segunda capa oculta.

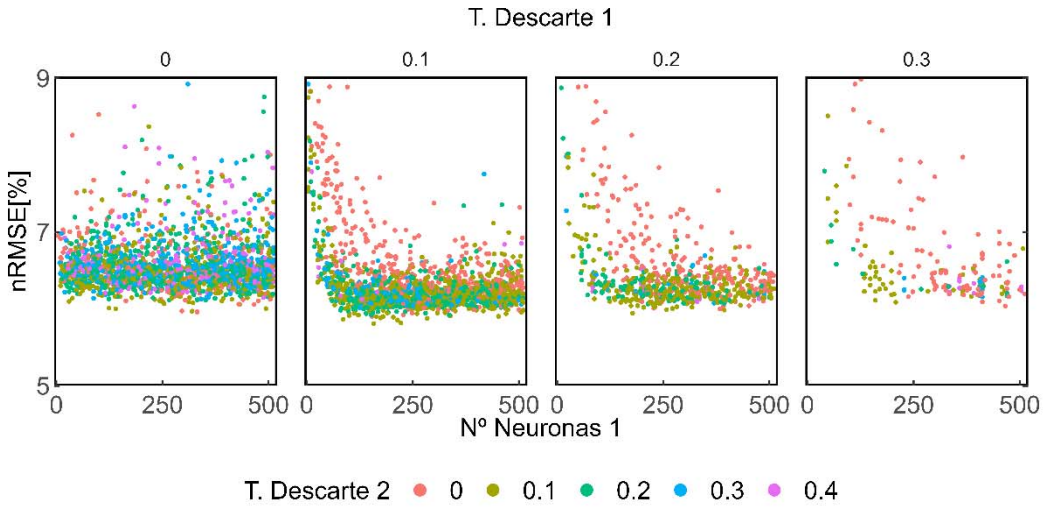
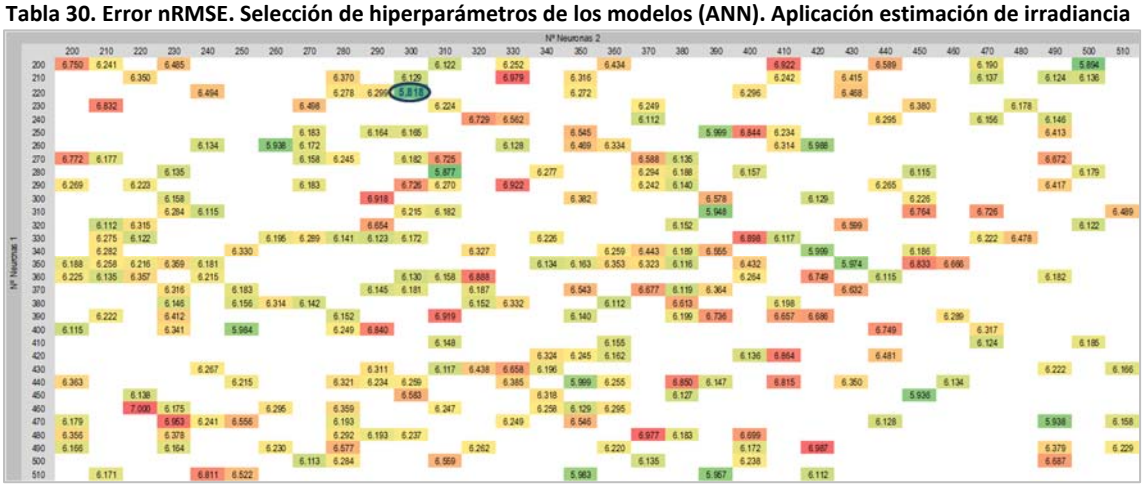


Figura 69. Detalle del error de los modelos generados en la búsqueda. Método de hiper-bandas y búsqueda aleatoria (ANN, modelos de dos capas con T. Descarte inicial nulo, T. Aprendizaje 0.001, T. Descarte capa 1 hasta 0.3, y T. Descarte capa 2 hasta 0.4). Aplicación estimación de irradiancia

La Figura 69 permite identificar el valor tanto de la tasa de descarte de la primera capa como de la segunda y en ambos casos el valor es coincidente y corresponde al 0.1. El número de neuronas se determina a partir de los 268 modelos seleccionados mediante los criterios anteriores y que se muestran en la Tabla 30.

Tabla 30. Error nRMSE. Selección de hiperparámetros de los modelos (ANN). Aplicación estimación de irradiancia



En dicha Tabla 30 se observa que el volumen de información a explorar en las redes neuronales es muy grande y también como los sistemas de búsqueda aleatoria e hiper-bandas, no han evaluado todas las posibilidades, aunque sí suficientes, para que no existan diferencias sustanciales entre el error de los modelos. Tomando el mejor de los modelos estudiado, se obtiene que el número de neuronas de la primera capa son 220, y el número de neuronas de la segunda capa son 300.

Según estos criterios y análisis los valores de los hiperparámetros para las redes neuronales se adjuntan en la Tabla 31.

Tabla 31. Hiperparámetros óptimos y número de modelos para su obtención (ANN). Aplicación estimación de irradiancia

Nº Capas	2
T. Aprendizaje	0.001
T. Descarte inicial	0
T. Descarte 1	0.1
T. Descarte 2	0.1
Nº Neuronas 1	220
Nº Neuronas 2	300
Nº Modelos	106950

Con el entrenamiento de los modelos se calculan los parámetros internos, pesos y sesgos de la red, y una vez definido el modelo se puede estimar la irradiancia en base a los datos de la muestra de prueba. Los resultados para cada una de las observaciones de la muestra se presentan en la Figura 70. La parte izquierda indica cómo se comporta el modelo en las observaciones puntuales escogidas al azar, mientras que la parte derecha a través de la gráfica de dispersión se puede observar el comportamiento general y la tendencia.

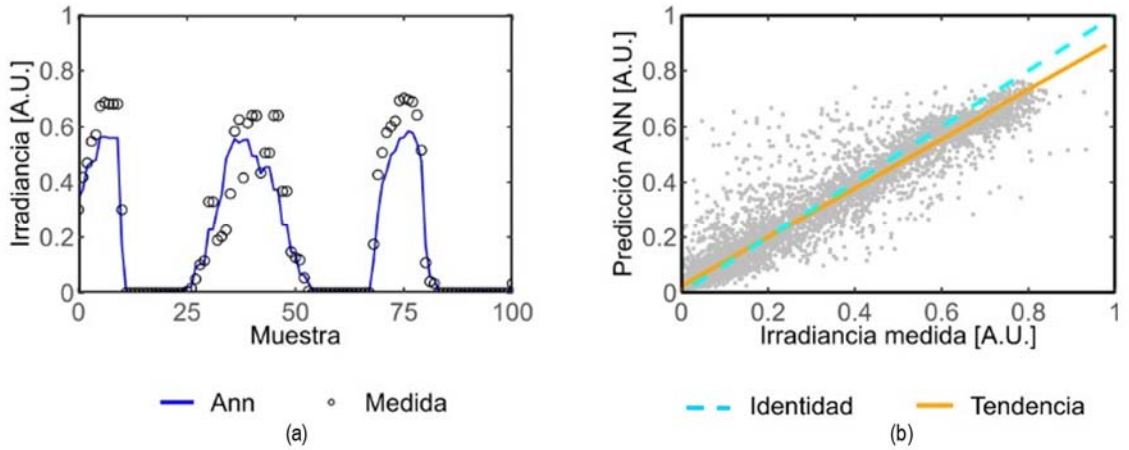


Figura 70. Muestra del resultado ANN (a) y gráfica de dispersión del resultado ANN (b). Aplicación estimación de irradiancia

El modelo es capaz de simular la mayoría de los valores medidos, tal y como se puede ver en la muestra ejemplo, (ver Figura 70 (a)), y en la gráfica de dispersión , (ver Figura 70 (b)), sin embargo, no es demasiado exacto con los valores de alta irradiancia ya que la línea de tendencia del modelo (anaranjada) difiere de la identidad (turquesa) a altas irradiancias. La nube de puntos visible por encima de la diagonal revela que el modelo sobre-estima considerablemente algunos registros de

los valores de bajas irradiancias. Si se compara con GB, se observa un peor comportamiento y la existencia de una mayor dispersión.

A continuación, se calcula el intervalo de confianza para los niveles del 90%, 95% y 99%, y se presenta el resultado de la muestra del periodo de prueba en la Figura 71.

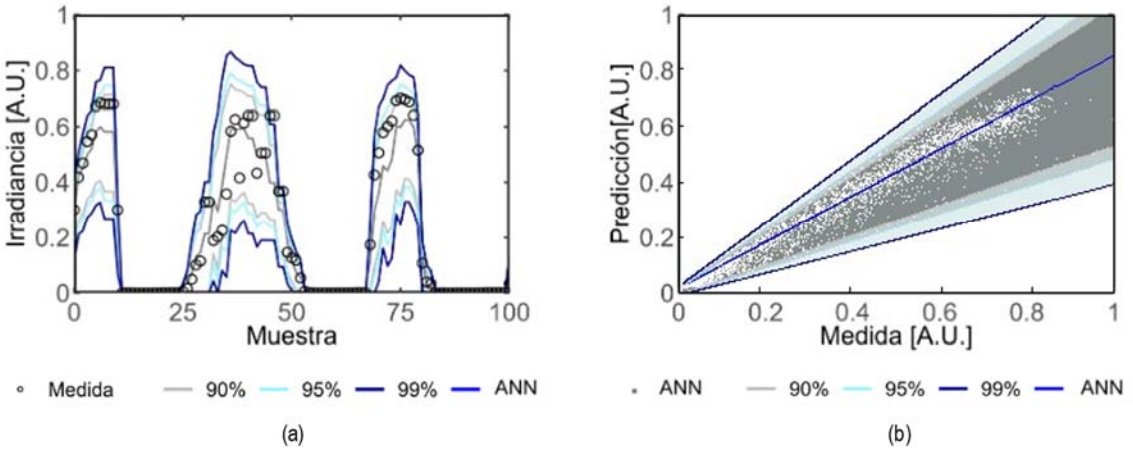


Figura 71. Muestra del intervalo de confianza de la ANN (a) y gráfica de dispersión con intervalo de confianza del resultado ANN (b). Aplicación estimación de irradiancia

La amplitud de los intervalos del modelo basado en redes neuronales es inferior a la que presentaban los anteriores modelos basados en otras metodologías. Sin embargo, su estimación puntual presenta un mayor error, posee muchos más hiperparámetros y son más complejas en los cálculos.

4.4. Discusión de los resultados

Una vez calculadas las estimaciones de los modelos y sus intervalos de confianza se comparan los resultados mediante el conjunto de métricas que ya han sido previamente definidas (ver Sección 2.7), y utilizadas en el análisis de los modelos aplicados a la generación (ver Sección 3.4), y que se enumeran en la Tabla 32.

Tabla 32. Resumen de las métricas de los diferentes modelos. Aplicación estimación de irradiancia

	MLR	RF	GB	ANN
nRMSE	7.4%	6.6%	3.0%	5.7%
nMAE	4.3%	3.4%	1.2%	3.4%
nBIAS	-0.9%	1.3%	0.1%	-0.6%
SS	-	10.9%	59.5%	23.0%
Mediana de la amplitud del intervalo 90% [A.U.]	19.1%	32.9%	13.4%	10.7%
Mediana de la amplitud del intervalo 95% [A.U.]	20.8%	41.2%	17.2%	12.6%
Mediana de la amplitud del intervalo 99% [A.U.]	23.0%	55.7%	21.0%	16.7%

En la Tabla 32 se observa como todas las métricas de los modelos de aprendizaje automático son inferiores al error que presenta el MLR. Además, en este caso de estudio la diferencia es superior a la obtenida al analizar los modelos que estimaban la producción de una planta fotovoltaica. Este

hecho muestra la importancia de los modelos de aprendizaje cuando las relaciones entre las variables no son lineales.

Entre los modelos de ML, el modelo GB, es el que mayor mejora presenta tal y como puede verse en el valor del SS. Aunque tiene un mayor intervalo de confianza que las redes neuronales, tiene una mayor rapidez y no necesita una demanda computacional alta, por lo que puede ser implementado fácilmente en las plantas para verificar la señal de la irradiancia in situ. En el nivel de confianza del 99%, los valores inferiores se quedan muy cercanos a cero, por lo que se recomienda trabajar con niveles de confianza inferiores.

Del análisis realizado, se puede concluir que, en la medida de la irradiancia el modelo GB es el más apropiado para estimar su valor a partir de datos climatológicos, la radiación teórica, el mes y la hora.

5. Conclusiones

La energía solar fotovoltaica es un componente crucial en el escenario energético mundial, gracias a su evolución tecnológica que le permite producir energía a costos competitivos. Sin embargo, el despliegue masivo de plantas fotovoltaicas conlleva desafíos para asegurar su optimización económica en varias fases del proceso, incluyendo la fabricación, instalación de módulos y la operación y mantenimiento durante la explotación de estas instalaciones.

Uno de los retos en la operación y mantenimiento de una planta es el análisis del rendimiento. Este análisis, que se realiza a través de la evaluación de su producción, tiene como objetivo detectar anomalías de manera temprana que afectan a su funcionamiento. Este enfoque proactivo permite orientar el mantenimiento hacia las áreas que más lo necesitan. Al hacerlo, no solo se maximiza el rendimiento de la planta, sino que también se optimizan los ingresos económicos derivados de su operación.

El análisis del funcionamiento de la planta se realiza utilizando los datos medidos comúnmente en la propia planta. Un factor que influye notablemente en el análisis del funcionamiento es el número de estaciones meteorológicas disponibles, que suele ser limitado debido a la competitividad económica en el sector. Debido a la escasez de estas estaciones, los módulos pueden estar situados lejos de los puntos de medición de la irradiancia solar. Esta distancia puede resultar en una discrepancia entre la irradiancia recibida por los módulos y la medida en la estación meteorológica, debido a la no uniformidad en la nubosidad que afecta a la relación entre la irradiancia y la producción, lo que complica el análisis de la producción y, por ende, del rendimiento de la planta.

Para dar respuesta a la necesidad planteada en esta tesis se han estudiado los modelos de aprendizaje automático existentes en la literatura para predecir la producción de plantas fotovoltaicas. Se ha optimizado su aplicación para el propósito planteado y se ha determinado cuál de ellos ofrece los mejores resultados para los casos de estudio propuestos. Además, se han analizado los modelos de caracterización de la irradiancia, lo que permite controlar la calidad de la variable más importante en la caracterización de la producción. Esto añade robustez al proceso.

En la revisión bibliográfica realizada, se han identificado las tres técnicas de aprendizaje automático que se desarrollan y aplican en la tesis. Las redes neuronales por ser las más utilizadas en la bibliografía. Los árboles de decisión, “Random Forest” y “Gradient Boosting”, por presentar buenos resultados en otras áreas de la ciencia. Adicionalmente, se ha trabajado con la regresión lineal múltiple, que es el modelo más sencillo y sirve como referencia para determinar la mejora aportada por el resto de los modelos.

Para garantizar la precisión en la predicción de producción e irradiancia en los métodos estudiados se necesita encontrar los hiperparámetros adecuados. El resultado de la búsqueda de los hiperparámetros en general revela que el error es más sensible a unos hiperparámetros concretos, que, generalmente, no existen valores únicos que minimicen el error y que estos además están relacionados entre sí. La metodología de la búsqueda está condicionada por el número de hiperparámetros y el rango de valores de estos, siendo posible utilizar para los modelos de árboles de decisión una búsqueda sistemática, mientras que para las redes neuronales hay que utilizar una combinación del método de hiper-banda y búsqueda aleatoria. La conclusión del estudio es que el análisis de los hiperparámetros es fundamental para garantizar un resultado adecuado en

el ajuste del algoritmo y la definición del modelo de producción de cada una de las plantas.

Para seleccionar el modelo no solo es importante valorar el error del modelo, sino que también hay que analizar el intervalo de confianza de la predicción, que define la incertidumbre del resultado. Para poder calcular los intervalos, se han utilizado los modelos “Quantile Regression”, “Quantile Random Forest” y “Quantile Gradient Boosting”, que, a través del cálculo de cuantiles, definen el intervalo de confianza. En el caso de las redes neuronales, se adaptan las capas de la red, para que los resultados del modelo tengan una distribución Normal que permite definir sus intervalos de confianza. No se han encontrado estudios similares en la literatura donde se proporcionen intervalos de confianza para las estimaciones.

Los resultados de los cálculos de los intervalos indican que los modelos basados en cuantiles proporcionan un intervalo asimétrico, con un límite superior más ajustado que el inferior, mientras que las redes neuronales ofrecen un intervalo simétrico. Los intervalos asimétricos son preferibles en este contexto, ya que reflejan tanto el comportamiento de la producción de las plantas como de la irradiancia. Ambas tienen un límite superior técnico que no pueden exceder, mientras que el límite inferior puede llegar a valores muy bajos dependiendo de la nubosidad. Para evaluar la bondad del intervalo, se utiliza la mediana de la amplitud del intervalo como valor de comparación siendo deseable intervalos con amplitudes lo más bajas posibles.

La aplicación de los tres modelos de aprendizaje automático a las plantas para predecir su producción muestra a través del SS que estos mejoran los resultados respecto al modelo referencia, la regresión lineal múltiple, en un rango del 14% al 30%. Esta mejora varía dependiendo de la planta y del modelo utilizado. Estos resultados justifican plenamente la realización del estudio.

La influencia de las características de la planta sobre la mejora de los resultados es inversamente proporcional a la complejidad de la misma. Es decir, cuanto más compleja sea la planta de modelizar, los modelos de aprendizaje automático aportan una mayor mejora. Así, en la planta PV1, de terreno complejo y que solo cuenta con una estación meteorológica, el modelo "Random Forest", tiene una mejora respecto a la referencia del orden del 30%, mientras que el mismo modelo en las plantas PV2 y PV3 presenta una mejora de alrededor del 25%. El modelo "Gradient Boosting" presenta una mejora para la planta PV1 del 26%, para la planta PV2 del 23% y para la planta PV3 del 22%, y las redes neuronales mejoran el resultado de la referencia en la planta PV1 un 16% y en las plantas PV2 y PV3 un 14%.

El modelo "Random Forest" es el que menor error de predicción presenta, su nRMSE se encuentra entre un 1.9% y un 5.4%, dependiendo de la planta, frente a la referencia cuyo rango de error se encuentra en 2.6% y 7.9% para las mismas plantas. Los errores nRMSE del modelo "Gradient Boosting" son muy próximos a los del modelo "Random Forest", entre el 2% al 5.9%, mientras que las redes neuronales presentan errores nRMSE ligeramente superiores a los modelos de árboles de decisión, entre el 2.2% y el 6.6%.

Los resultados del estudio de los intervalos de la predicción de la producción de la planta muestran que el intervalo del modelo "Random Forest" es mucho más amplio que el de los demás modelos de aprendizaje automático. Esto implica que, aunque el "Random Forest" tiene un error inferior al

de los demás modelos, presenta una mayor incertidumbre y dispersión en sus resultados. Este comportamiento del "Random Forest" puede ser problemático si se pretende utilizar el modelo para identificar eventos con producción anómala. Sin embargo, el modelo “Gradient Boosting” que tiene un error similar al modelo “Random Forest”, tiene un intervalo asimétrico y con la amplitud menor para cualquiera de los niveles de confianza. Por lo que el modelo “Gradient Boosting" se destaca como el método preferente para simular el comportamiento de la producción de una planta fotovoltaica a nivel inversor.

Los resultados que presenta el modelo “Gradient Boosting” son un error del 5.9% en la planta PV1, del 1.9% para la planta PV2 y del 3.2% para la planta PV3, siendo la mediana de la amplitud de su intervalo, en el nivel de confianza del 90%, de 0.6% para la planta PV1, 0.2% en la planta PV2 y 0.3% en la planta PV3. Además, este modelo tiene una mayor eficiencia en el procesamiento y cálculo.

Cuando se aplican las mismas metodologías a la predicción de la irradiancia, se observa que los valores de los ajustes de los hiperparámetros llevan a modelos más complejos en el caso de los árboles de decisión, compensando la falta de claridad entre las relaciones de las variables. Así, el algoritmo de “Random Forest” necesita un tamaño de árbol con una profundidad de 29 y 1650 árboles y el algoritmo de “Gradient Boosting” necesita 850 árboles con un tamaño de árbol marcado por una profundidad de 9, valores muy superiores a los que se manejaban en los modelos de producción. En el caso de las redes neuronales se trabaja con un modelo de 2 capas, con 220 neuronas en la primera capa y 300 en la segunda, valores que son similares al modelo más complejo de redes que es el que estima la producción en la planta PV1.

En cuanto a los errores de los modelos, se encuentran en el rango superior de los que se obtenían en la predicción de la producción, así, por ejemplo, en el modelo utilizado como referencia los errores en la predicción de la producción se encontraban en la horquilla 2.6% y 7.9% frente al error de la predicción de la irradiancia que es de un 7.4%. Es decir, predecir la irradiancia a partir de las variables de la estación meteorológica presenta el mismo error que la predicción de la energía a partir de las variables medidas en la planta cuando existen pendientes diferentes y hay grandes distancias. El diferencial de mejora que aportan los modelos más avanzados con respecto al modelo de referencia es de entre un 11% y un 59% dependiendo del modelo en cuestión, por lo que la elección del modelo tiene una mayor repercusión en este proceso.

El modelo que presenta la mayor mejora de resultado y por tanto el menor error en la predicción de la irradiancia es el modelo “Grandient Boosting” con un error nRMSE de 3%. El proceso encadenado de generación de los árboles que permite disminuir el error hace que sea muy adecuado para este complejo problema. El error del “Random Forest” en este caso es de 6.6%, peor que el que presentan las redes neuronales que es de 5.7%.

La incertidumbre de los modelos aplicados en el cálculo de la irradiancia también se ve aumentada en un orden de diez veces a la que presentan las mismas metodologías en el cálculo de la producción. El modelo que presenta la menor amplitud de intervalo de confianza son las redes neuronales, con el valor de la mediana de la amplitud de todas las observaciones del 10.7% para el nivel de confianza del 90%, 12.6% para el 95% y 16.7% para el 99%. El modelo “Gradient Boosting” tiene una amplitud de intervalos ligeramente mayor, 13.4% para el 90%, 17.2% para el

95% y 21% para el 99%, siendo este mucho más preciso en la estimación puntual y, además, más rápido y eficiente desde un punto de vista computacional. Por tanto "Gradient Boosting" también resulta el modelo más adecuado para estimar la irradiancia a partir de los datos meteorológicos.

De este estudio se puede concluir que los modelos de aprendizaje automático pueden caracterizar la producción de un inversor solo con las variables medidas en la planta, con unos errores bajos y una buena precisión del resultado. Disponer de la generación teórica permite identificar desvíos inducidos por ineficiencias de la planta. También estos modelos pueden predecir la irradiancia a partir de medidas climatológicas y de la ubicación exacta del punto de medición, pero con un mayor error e incertidumbre que la producción. Conocer la irradiancia ayuda a mejorar el cálculo de la producción, del "Performance Ratio" de la planta e incluso disponer de una medida de irradiancia en el caso de que la medida falle. El modelo más apropiado para predecir estas variables en el campo de la fotovoltaica es además el modelo "GB".

Si bien los modelos presentados son particulares para los escenarios expuestos, el procedimiento y la metodología utilizados en el estudio son de alcance general. Esto significa que el enfoque basado en modelos de aprendizaje automático, la metodología para el ajuste de los hiperparámetros y la forma de definir los intervalos de confianza son aplicables a cualquier planta fotovoltaica.

5.1. Limitaciones del estudio y nuevas líneas de trabajo

A pesar de los hallazgos significativos de esta investigación, es necesario reconocer algunas limitaciones, porque ellas son la base de nuevas investigaciones y su resolución ayudará a mejorar las plantas fotovoltaicas.

Una de las principales restricciones se relaciona con la limitación de la información. No ha sido fácil disponer de información de plantas de gran tamaño con varios años de operación. Este hecho ha repercutido, por ejemplo, en que se ha trabajado exclusivamente con plantas de estructura fija, ya que no se disponía de datos de plantas de gran tamaño con seguidor que tuvieran varios años de operación. Esta circunstancia, ha sido una ventaja porque ha permitido comparar los modelos en plantas de la misma tecnología y llegar a conclusiones interesantes, pero deja abierto el estudio al análisis de la implicación de incluir variables relacionadas con el seguidor.

También el estudio ha trabajado con plantas que están ubicadas en zonas óptimas para el desarrollo de la fotovoltaica, pero no ha trabajado con supuestos ubicados en climatologías más complejas como pueden ser zonas que presenten hielos durante parte del año o zonas desérticas con altas temperaturas. Futuras investigaciones podrían abordar esta limitación mediante el estudio de plantas que se están instalando ahora y que se ubican en lugares donde se den estos sucesos climatológicos.

Los resultados obtenidos mejoran el conocimiento de la eficiencia de las plantas fotovoltaicas, pueden ser aplicados a casos reales y ofrecen una base sólida sobre la cual se pueden construir nuevas líneas de trabajo. Así por ejemplo sería interesante continuar el trabajo estudiando parte de las limitaciones del estudio que ayudaran a entender cómo se comportan los modelos de aprendizaje automático en otros diferentes supuestos como, por ejemplo:

- Entornos en ubicaciones con climatologías extremas, con presencia de hielo, o en contraste en zonas desérticas de altas temperaturas.
- En plantas fotovoltaicas cuya estructura sea con seguidor solar.
- Analizar los resultados con diferentes tipos de inversores: inversor centralizado frente a tipo "string inverter".

En general, los modelos de inteligencia artificial y en concreto los de aprendizaje automático y profundo, están en constante evolución, y se mejoran día a día. En el tiempo en que se ha llevado a cabo este estudio se han observado sus mejoras, esto implica que no es un estudio terminado, sino que también debe mantenerse en evolución comprobando si los avances en los métodos tienen aplicación en el ámbito de la fotovoltaica y si suponen una mejora respecto de las conclusiones obtenidas en este estudio

Bibliografía

[1] A.P. Talayero, J.J. Melero, A. Llombart, N.Y. Yürüşen, Machine Learning models for the estimation of the production of large utility-scale photovoltaic plants, Solar Energy. 254 (2023) 88–101. https://doi.org/10.1016/j.solener.2023.03.007.

[2] A.P. Talayero, N.Y. Yürüşen, A.L.-E.J.J. Melero, Anomaly Detection At Inverter Level Via Machine Learning Algorithms Under the Absence of O&M Logbooks, in: 37th European Photovoltaic Solar Energy Conference (EUPVSEC), Libone, 2020: pp. 1381–1387. https://doi.org/10.4229/EUPVSEC20202020-5DO.4.2.

[3] A.P. Talayero, A. Llombart, J.J. Melero, Diagnosis of failures in solar plants based on performance monitoring, Renewable Energy and Power Quality Journal. 18 (2020) 128–133. https://doi.org/10.24084/repqj18.248.

[4] A.P. Talayero, J.J. Melero, A. Llombart, A. Casado, Operation and maintenance in solar plants: Eight study cases, Renewable Energy and Power Quality Journal. 1 (2018) 499–504. https://doi.org/10.24084/repqj16.363.

[5] N.Y. Yürüşen, B. Uzunoğlu, A.P. Talayero, A.L. Estopiñán, Apriori and K-Means algorithms of machine learning for spatio-temporal solar generation balancing, Renewable Energy. 175 (2021) 702–717. https://doi.org/10.1016/j.renene.2021.04.098.

[6] A.P. Talayero, N. Yıldırım Yürüşen, J.J. Melero, Mejora de las estrategias de mantenimiento en plantas de generación renovable a partir de los datos SCADA, VI Congreso Smart Grids. (2019). https://www.smartgridsinfo.es/comunicaciones/comunicacion-mejora-estrategias-mantenimiento-plantas-generacion-renovable-datos-scada (accessed November 1, 2023).

[7] IRENA, Renewable Capacity Statistics 2023, International Renewable Energy Agency, 2023. https://www.irena.org/Publications/2023/Mar/Renewable-capacity-statistics-2023 (accessed November 27, 2023).

[8] IEA, Energy Technology Perspectives 2023, International Energy Agency, 2023. https://www.iea.org/reports/energy-technology-perspectives-2023 (accessed November 27, 2023).

[9] J. Phillips, Determining the sustainability of large-scale photovoltaic solar power plants, Renewable and Sustainable Energy Reviews. 27 (2013) 435–444. https://doi.org/10.1016/j.rser.2013.07.003.

[10] Institute Fraunhofer, Photovoltaics Report, Fraunhofer ISE, 2023. https://www.ise.fraunhofer.de/content/dam/ise/de/documents/publications/studies/Photovoltaics-Report.pdf (accessed November 27, 2023).

[11] IEA, Electricity Market Report 2023, International Energy Agency, 2023. https://www.iea.org/reports/electricity-market-report-2023 (accessed November 24, 2023).

[12] SEIA, Solar Power Purchase Agreements, Solar Energy Industries Association. (2021). https://www.seia.org/research-resources/solar-power-purchase-agreements (accessed November 27, 2023).

[13] UNEF, Power Purchase Agreement – PPA, Union Española Fotovoltaica, 2018. https://www.unef.es/en/descargar-documento/2242dc938ff81a85ddaf12f2a3081aaa (accessed November 26, 2023).

[14] IEA-PVPS-Task1, Trends in Photovoltaic Applications, International Energy Agency, 2022. https://iea-pvps.org/wp-content/uploads/2020/02/5319-iea-pvps-report-2019-08-lr.pdf (accessed November 27, 2023).

[15] UNEF, Fomentando la biodiversidad y el crecimiento sostenible, 2023. https://www.unef.es/en/recursos-informes?idMultimediaCategoria=18 (accessed November 29, 2023).

[16] Ministerio para la Transición Ecológica y el Reto Demográfico, Plan Nacional Integrado de Energía y Clima 2021-2030, Gobierno de España, 2020. https://www.miteco.gob.es/es/prensa/pniec.aspx (accessed November 26, 2023).

[17] Parlamento y Consejo de la Unión Europea, Reglamento UE 2018/1999 del 11 de diciembre de 2018, Diario Oficial de la Unión Europea, Unión Europea, 2018. https://www.boe.es/doue/2018/328/L00001-00077.pdf (accessed November 23, 2023).

[18] IEA-PVPS-Task13, Review of Failures of Photovoltaic Modules, Internationa Eergy Agency, 2014. https://iea-pvps.org/wp-content/uploads/2020/01/IEA-PVPS_T13-01_2014_Review_of_Failures_of_Photovoltaic_Modules_Final.pdf (accessed November 27, 2023).

[19] IEA-PVPS-Task13, Analytical Monitoring of Grid-connected Photovoltaic Systems, Internationa Eergy Agency, 2014. https://iea-pvps.org/wp-content/uploads/2020/01/IEA-PVPS_T13-D2_3_Analytical_Monitoring_of_PV_Systems_Final.pdf (accessed November 27, 2023).

[20] Jaroslav Mencík, Reliability of Systems, in: Concise Reliability for Engineers, IntechOpen Book Series, 2016: pp. 33–42. https://doi.org/10.5772/62358.

[21] D.C. Jordan, T.J. Silverman, J.H. Wohlgemuth, S.R. Kurtz, K.T. VanSant, Photovoltaic failure and degradation modes, Progress in Photovoltaics: Research and Applications. 25 (2017) 318–326. https://doi.org/10.1002/pip.2866.

[22] A. Chokor, M. El Asmar, S. V Lokanath, A Review of Photovoltaic DC Systems Prognostics and Health Management : Challenges and Opportunities, The Annual Conference of the Prognostics and Health Management Society. (2016). https://doi.org/10.36001/phmconf.2016.v8i1.2505.

[23] S.R. Madeti, S.N. Singh, A comprehensive study on different types of faults and detection techniques for solar photovoltaic system, Solar Energy. 158 (2017) 161–185. https://doi.org/10.1016/j.solener.2017.08.069.

[24] M.K. Alam, F. Khan, S. Member, J. Johnson, J. Flicker, A Comprehensive Review of Catastrophic Faults in PV Arrays: Types, Dectection, and Mitigation Techniques, IEEE Journal of Photovoltaics. 5 (2015) 982–997. https://doi.org/10.1109/JPHOTOV.2015.2397599.

[25] IEA-PVPS-Task7, Reliability Study of Grid Connected PV Systems Field Experience and Recommended Design Practice Task 7, Internationa Eergy Agency, 2002. https://iea-pvps.org/wp-content/uploads/2020/01/rep7_08.pdf (accessed November 27, 2023).

[26] D. DeGraaff, R. Lacerda, Z. Campeau, S. Corp, Degradation mechanisms in Si module technologies observed in the field; their analysis and statistics, NREL 2011 Photovoltaic

Module Reliability Workshop. (2011) 1–25. https://www1.eere.energy.gov/solar/pdfs/pvmrw2011_01_plen_degraaff.pdf (accessed November 28, 2023).

[27] N. Bosco, T.J. Silverman, S. Kurtz, Climate specific thermomechanical fatigue of flat plate photovoltaic module solder joints, Microelectronics Reliability. 62 (2016) 124–129. https://doi.org/10.1016/j.microrel.2016.03.024.

[28] IEA-PVPS-Task13, Assessment of Photovoltaic Module Failures in the Field, Internationa Eergy Agency, 2017. https://iea-pvps.org/key-topics/report-assessment-of-photovoltaic-module-failures-in-the-field-2017/ (accessed November 27, 2023).

[29] E.E. Van Dyk, J.B. Chamel, A.R. Gxasheka, Investigation of delamination in an edge-defined film-fed growth photovoltaic module, Solar Energy Materials and Solar Cells. 88 (2005) 403–411. https://doi.org/10.1016/j.solmat.2004.12.004.

[30] D.C. Jordan, S.R. Kurtz, K. VanSant, J. Newmiller, Compendium of photovoltaic degradation rates, Progress in Photovoltaics: Research and Applications. 24 (2016) 978–989. https://doi.org/10.1002/pip.2744.

[31] D. Jordan, S. Kurtz, Photovoltaic module stability and reliability, in: The Performance of Photovoltaic (PV) Systems: Modelling, Measurement and Assessment, National Renewable Energy Laboratory (NREL), 2016: pp. 71–101. https://doi.org/10.1016/B978-1-78242-336-2.00003-3.

[32] D.C. Miller, E. Annigoni, A. Ballion, J.G. Bokria, L.S. Bruckman, D.M. Burns, X. Chen, J. Feng, R.H. French, S. Fowler, C.C. Honeker, M.D. Kempe, H. Khonkar, M. Kohl, L.E. Perret-Aebi, N.H. Phillips, K.P. Scott, F. Sculati-Meillaud, J.H. Wohlgemuth, Degradation in PV encapsulant strength of attachment: An interlaboratory study towards a climate-specific test, in: Conference Record of the IEEE Photovoltaic Specialists Conference, National Renewable Energy Laboratory, 2016: pp. 95–100. https://doi.org/10.1109/PVSC.2016.7749556.

[33] D.C. Miller, J. Bath, M. Köhl, T. Shioda, PV-Module Reliability: UV , Temperature , and Humidity, (2014) 0–13. https://www.nrel.gov/docs/fy14osti/62228.pdf (accessed November 28, 2023).

[34] B.P. Rand, J. Genoe, P. Heremans, J. Poortmans, Solar Cells Utilizing Small Molecular Weight Organic Semiconductors, Prog. Photovolt: Res. Appl. 15 (2007) 659–676. https://doi.org/10.1002/pip.

[35] E.S. Kopp, V.P. Lonij, A.E. Brooks, P.L. Hidalgo-Gonzalez, A.D. Cronin, I-V curves and visual inspection of 250 PV modules deployed over 2 years in tucson, Conference Record of the IEEE Photovoltaic Specialists Conference. (2012) 3166–3171. https://doi.org/10.1109/PVSC.2012.6318251.

[36] M. Gostein, J.R. Caron, B. Littmann, Measuring soiling losses at utility-scale PV power plants, 2014 IEEE 40th Photovoltaic Specialist Conference, PVSC 2014. (2014) 885–890. https://doi.org/10.1109/PVSC.2014.6925056.

[37] M. Saidan, A.G. Albaali, E. Alasis, J.K. Kaldellis, Experimental study on the effect of dust deposition on solar photovoltaic panels in desert environment, Renewable Energy. 92 (2016) 499–505. https://doi.org/10.1016/j.renene.2016.02.031.

[38] A. Sayyah, M.N. Horenstein, M.K. Mazumder, Energy yield loss caused by dust deposition on photovoltaic panels, Solar Energy. 107 (2014) 576–604. https://doi.org/10.1016/j.solener.2014.05.030.

[39] S. Noack-Schönmann, O. Spagin, K.P. Gründer, M. Breithaupt, A. Günter, B. Muschik, A.A. Gorbushina, Sub-aerial biofilms as blockers of solar radiation: Spectral properties as tools to characterise material-relevant microbial growth, International Biodeterioration and Biodegradation. 86 (2014) 286–293. https://doi.org/10.1016/j.ibiod.2013.09.020.

[40] M. Dhimish, V. Holmes, B. Mehrdadi, M. Dales, The impact of cracks on photovoltaic power performance, Journal of Science: Advanced Materials and Devices. 2 (2017) 199–209. https://doi.org/10.1016/j.jsamd.2017.05.005.

[41] J.I. Van Mölken, U.A. Yusufoglu, A. Safiei, H. Windgassen, R. Khandelwal, T.M. Pletzer, H. Kurza, Impact of micro-cracks on the degradation of solar cell performance based on two-diode model parameters, Energy Procedia. 27 (2012) 167–172. https://doi.org/10.1016/j.egypro.2012.07.046.

[42] S. Kajari-Schröder, I. Kunze, U. Eitner, M. Köntges, Spatial and orientational distribution of cracks in crystalline photovoltaic modules generated by mechanical load tests, Solar Energy Materials and Solar Cells. 95 (2011) 3054–3059. https://doi.org/10.1016/j.solmat.2011.06.032.

[43] W. Nsengiyumva, S.G. Chen, L. Hu, X. Chen, Recent advancements and challenges in Solar Tracking Systems (STS): A review, Renewable and Sustainable Energy Reviews. 81 (2018) 250–279. https://doi.org/10.1016/j.rser.2017.06.085.

[44] J.S. Jeong, N. Park, Field discoloration analysis and UV/temperature accelerated degradation test of EVA for PV, Conference Record of the IEEE Photovoltaic Specialists Conference. 1 (2013) 3010–3013. https://doi.org/10.1109/PVSC.2013.6745095.

[45] P. Manganiello, M. Balato, M. Vitelli, A Survey on Mismatching and Aging of PV Modules: The Closed Loop, IEEE Transactions on Industrial Electronics. 62 (2015) 7276–7286. https://doi.org/10.1109/TIE.2015.2418731.

[46] L. Menyhart, A. Anda, Z. Nagy, A new method for checking the leveling of pyranometers, Solar Energy. 120 (2015) 25–34. https://doi.org/10.1016/j.solener.2015.06.033.

[47] IEA-PVPS-Task13, Technical Assumptions Used in PV Financial Models Review of Current Practices and Recommendations, Internationa Eergy Agency, 2017. https://iea-pvps.org/wp-content/uploads/2020/01/Report_IEA-PVPS_T13-08_2017_Technical_Assumptions_Used_in_PV_Financial_Models.pdf (accessed November 27, 2023).

[48] JRC, PV Status Report 2017, European Commission, 2017. https://ec.europa.eu/jrc (accessed November 26, 2023).

[49] M.P. Almeida, O. Perpiñán, L. Narvarte, PV power forecast using a nonparametric PV model, Solar Energy. 115 (2015) 354–368. https://doi.org/10.1016/j.solener.2015.03.006.

[50] R. Platon, S. Pelland, Y. Poissant, Modelling the Power Production of a Photovoltaic

System: Comparison of Sugeno-Type Fuzzy Logic and PVSAT-2 Models, in: Europe Solar Conference (ISES), 2012. https://www.researchgate.net/publication/255908005_Modelling_the_Power_Production_of_a_Photovoltaic_System (accessed November 27, 2023).

[51] N. Gupta, R. Garg, P. Kumar, Sensitivity and reliability models of a PV system connected to grid, Renewable and Sustainable Energy Reviews. 69 (2017) 188–196. https://doi.org/10.1016/j.rser.2016.11.031.

[52] E. Alpaydin, Introduction to Machine Learning, fourth edi, The MIT Press, 2020.

[53] A. Chouder, S. Silvestre, Automatic supervision and fault detection of PV systems based on power losses analysis, Energy Conversion and Management. 51 (2010) 1929–1937. https://doi.org/10.1016/j.enconman.2010.02.025.

[54] K.H. Chao, S.H. Ho, M.H. Wang, Modeling and fault diagnosis of a photovoltaic system, Electric Power Systems Research. 78 (2008) 97–105. https://doi.org/10.1016/j.epsr.2006.12.012.

[55] S.A.S. Eldin, M.S. Abd-Elhady, H.A. Kandil, Feasibility of solar tracking systems for PV panels in hot and cold regions, Renewable Energy. 85 (2015) 228–233. https://doi.org/10.1016/j.renene.2015.06.051.

[56] F. Golestaneh, P. Pinson, H.B. Gooi, Very Short-Term Nonparametric Probabilistic Forecasting of Renewable Energy Generation; With Application to Solar Energy, Power Systems, IEEE Transactions On. PP (2016) 1–14. https://doi.org/10.1109/TPWRS.2015.2502423.

[57] R. Fazai, K. Abodayeh, M. Mansouri, M. Trabelsi, H. Nounou, M. Nounou, G.E. Georghiou, Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems, Solar Energy. 190 (2019) 405–413. https://doi.org/10.1016/j.solener.2019.08.032.

[58] A.M. Nobre, C.A. Severiano, S. Karthik, M. Kubis, L. Zhao, F.R. Martins, E.B. Pereira, R. Rüther, T. Reindl, PV power conversion and short-term forecasting in a tropical, densely-built environment in Singapore, Renewable Energy. 94 (2016) 496–509. https://doi.org/10.1016/j.renene.2016.03.075.

[59] M.B. Øgaard, A.F. Skomedal, H. Haug, E.S. Marstein, Robust and fast detection of small power losses in large-scale PV systems, IEEE Journal of Photovoltaics. 11 (2021) 819–826. https://doi.org/10.1109/JPHOTOV.2021.3060732.

[60] F. EaChollet, J.J. Alaire, Deep Learning with R, MEAP, Manning Publications, 2017.

[61] N. Gökmen, W. Hu, P. Hou, Z. Chen, D. Sera, S. Spataru, Investigation of wind speed cooling effect on PV panels in windy locations, Renewable Energy. 90 (2016) 283–290. https://doi.org/10.1016/j.renene.2016.01.017.

[62] A. Livera, M. Theristis, G. Makrides, G.E. Georghiou, On-line failure diagnosis of grid-connected photovoltaic systems based on fuzzy logic, in: 12th International Conference on Compatibility, Power Electronics and Power Engineering, CPE-POWERENG 2018, Proceedings - 2018 IEEE, 2018: pp. 1–6. https://doi.org/10.1109/CPE.2018.8372537.

[63] B. Li, C. Delpha, D. Diallo, A. Migan-Dubois, Application of Artificial Neural Networks to photovoltaic fault detection and diagnosis: A review, Renewable and Sustainable

Energy Reviews. 138 (2021) 110512. https://doi.org/10.1016/J.RSER.2020.110512.

[64] A. Mellit, G.M. Tina, S.A. Kalogirou, Fault detection and diagnosis methods for photovoltaic systems: A review, Renewable and Sustainable Energy Reviews. 91 (2018) 1–17. https://doi.org/10.1016/j.rser.2018.03.062.

[65] J.G. De Gooijer, R.J. Hyndman, 25 years of time series forecasting, International Journal of Forecasting. 22 (2006) 443–473. https://doi.org/10.1016/j.ijforecast.2006.01.001.

[66] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, first edi., Prmu, 2016. www.deeplearningbook.org (accessed November 26, 2023).

[67] K.P. Sinaga, M.S. Yang, Unsupervised K-means clustering algorithm, IEEE Access. 8 (2020) 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796.

[68] A.K. Yadav, S.S. Chandel, Identification of relevant input variables for prediction of 1-minute time-step photovoltaic module power using Artificial Neural Network and Multiple Linear Regression Models, Renewable and Sustainable Energy Reviews. 77 (2017) 955–969. https://doi.org/10.1016/j.rser.2016.12.029.

[69] S. Kurtz, J. Newmiller, A. Kimber, R. Flottemesch, E. Riley, T. Dierauf, J. McKee, P. Krishnani, Analysis of Photovoltaic System Energy Performance Evaluation Method, National Renewable Energy Laboratory, 2013. https://www.nrel.gov/docs/fy14osti/60628.pdf (accessed November 24, 2023).

[70] V. Beránek, T. Olšan, M. Libra, V. Poulek, J. Sedláček, M.Q. Dang, I.I. Tyukhov, New monitoring system for photovoltaic power plants' management, Energies. 11 (2018). https://doi.org/10.3390/en11102495.

[71] S. Daliento, A. Chouder, P. Guerriero, A.M. Pavan, A. Mellit, R. Moeini, P. Tricoli, Monitoring, diagnosis, and power forecasting for photovoltaic fields: A review, International Journal of Photoenergy. 2017 (2017). https://doi.org/10.1155/2017/1356851.

[72] J. Wang, Research on the Application of Machine Learning in Photovoltaic Power Station, 2021 3rd International Academic Exchange Conference on Science and Technology Innovation, IAECST 2021. (2021) 1927–1930. https://doi.org/10.1109/IAECST54258.2021.9695774.

[73] M.M. Rahman, J. Selvaraj, N.A. Rahim, M. Hasanuzzaman, Global modern monitoring systems for PV based power generation: A review, Renewable and Sustainable Energy Reviews. 82 (2018) 4142–4158. https://doi.org/10.1016/j.rser.2017.10.111.

[74] A. Triki-Lahiani, A. Bennani-Ben Abdelghani, I. Slama-Belkhodja, Fault detection and monitoring systems for photovoltaic installations: A review, Renewable and Sustainable Energy Reviews. 82 (2018) 2680–2692. https://doi.org/10.1016/j.rser.2017.09.101.

[75] UNE-EN 61724, Monitorización de sistemas fotovoltaicos Guías para la medida, el intercambio de datos y el análisis, (2000).

[76] S.R. Madeti, S.N. Singh, Monitoring system for photovoltaic plants: A review, Renewable and Sustainable Energy Reviews. 67 (2017) 1180–1207. https://doi.org/10.1016/j.rser.2016.09.088.

[77] N.M. Pearsall, Prediction and measurement of photovoltaic system energy yield, in: The

Performance of Photovoltaic (PV) Systems, 2017: pp. 183–208. https://doi.org/10.1016/B978-1-78242-336-2.00006-9.

[78] S. Ransome, P. Funtan, Why hourly averaged measurement data is insufficient to model PV performance accurately, in: 20th European Photovoltaic Solar Energy Conference, 2005: pp. 2752–2755. http://www.steveransome.com/PUBS/2005Barcelona_6DV_4_32.pdf (accessed November 27, 2023).

[79] D. Markovics, M.J. Mayer, Investigating the effect of training time for machine learning based photovoltaic power forecasting, 2022 8th International Youth Conference on Energy, IYCE 2022. (2022). https://doi.org/10.1109/IYCE54153.2022.9857544.

[80] M. Carpintero-Renteria, D. Santos-Martin, S. Koukoura, C. Nicolas-Martin, Photovoltaic electric power estimation with a machine learning algorithm based on neural networks and validated with deterministic approaches, Proceedings - 2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe, EEEIC / I and CPS Europe 2020. (2020). https://doi.org/10.1109/EEEIC/ICPSEurope49358.2020.9160751.

[81] Y. Wang, Y. Chen, H. Liu, X. Ma, X. Su, Q. Liu, Day-Ahead Photovoltaic Power Forcasting Using Convolutional-LSTM Networks, 2021 3rd Asia Energy and Electrical Engineering Symposium, AEEES 2021. (2021) 917–921. https://doi.org/10.1109/AEEES51875.2021.9403023.

[82] B. Navothna, S. Thotakura, Analysis on Large-Scale Solar PV Plant Energy Performance–Loss–Degradation in Coastal Climates of India, Frontiers in Energy Research. 10 (2022). https://doi.org/10.3389/fenrg.2022.857948.

[83] M.E.H. Jed, P.O. Logerais, C. Malye, O. Riou, F. Delaleux, M. El Bah, Analysis of the performance of the photovoltaic power plant of Sourdun (France), International Journal of Sustainable Engineering. 14 (2021) 1756–1768. https://doi.org/10.1080/19397038.2021.1971321.

[84] E. Fuster-Palop, C. Vargas-Salgado, J.C. Ferri-Revert, J. Payá, Performance analysis and modelling of a 50 MW grid-connected photovoltaic plant in Spain after 12 years of operation, Renewable and Sustainable Energy Reviews. 170 (2022). https://doi.org/10.1016/j.rser.2022.112968.

[85] M.E.H. Dahmoun, B. Bekkouche, K. Sudhakar, M. Guezgouz, A. Chenafi, A. Chaouch, Performance evaluation and analysis of grid-tied large scale PV plant in Algeria, Energy for Sustainable Development. 61 (2021) 181–195. https://doi.org/10.1016/j.esd.2021.02.004.

[86] N. Bansal, S.P. Jaiswal, G. Singh, Long term performance assessment and loss analysis of 9 MW grid tied PV plant in India, Materials Today: Proceedings. 60 (2022) 1056–1067. https://doi.org/10.1016/j.matpr.2022.01.263.

[87] F. Spertino, E. Chiodo, A. Ciocia, G. Malgaroli, A. Ratclif, Maintenance Activity, Reliability, Availability, and Related Energy Losses in Ten Operating Photovoltaic Systems up to 1.8 MW, IEEE Transactions on Industry Applications. 57 (2021) 83–93. https://doi.org/10.1109/TIA.2020.3031547.

[88] F. Spertino, A. Amato, G. Casali, A. Ciocia, G. Malgaroli, Reliability analysis and repair

activity for the components of 350 kW inverters in a large scale grid-connected photovoltaic system, Electronics (Switzerland). 10 (2021) 1–13. https://doi.org/10.3390/electronics10050564.

[89] Y. Yagi, H. Kishi, R. Hagihara, T. Tanaka, S. Kozuma, T. Ishida, M. Waki, M. Tanaka, S. Kiyama, Diagnostic technology and an expert system for photovoltaic systems using the learning method, Solar Energy Materials and Solar Cells. 75 (2003) 655–663. https://doi.org/10.1016/S0927-0248(02)00149-6.

[90] A. Woyte, J. Nijs, R. Belmans, Partial shadowing of photovoltaic arrays with different system configurations: Literature review and field test results, Solar Energy. 74 (2003) 217–233. https://doi.org/10.1016/S0038-092X(03)00155-5.

[91] M. Drif, A. Mellit, J. Aguilera, P.J. Pérez, A comprehensive method for estimating energy losses due to shading of GC-BIPV systems using monitoring data, Solar Energy. 86 (2012) 2397–2404. https://doi.org/10.1016/j.solener.2012.05.008.

[92] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de-Pison, F. Antonanzas-Torres, Review of photovoltaic power forecasting, Solar Energy. 136 (2016) 78–111. https://doi.org/10.1016/j.solener.2016.06.069.

[93] A. Alcañiz, D. Grzebyk, H. Ziar, O. Isabella, Trends and gaps in photovoltaic power forecasting with machine learning, Energy Reports. 9 (2023) 447–471. https://doi.org/10.1016/j.egyr.2022.11.208.

[94] J.F. Gaviria, G. Narváez, C. Guillen, L.F. Giraldo, M. Bressan, Machine learning in photovoltaic systems: A review, Renewable Energy. 196 (2022) 298–318. https://doi.org/10.1016/j.renene.2022.06.105.

[95] C. Ventura, G.M. Tina, Development of models for on-line diagnostic and energy assessment analysis of PV power plants: The study case of 1 MW Sicilian PV plant, Energy Procedia. 83 (2015) 248–257. https://doi.org/10.1016/j.egypro.2015.12.179.

[96] A. Manavi Alam, NAhid-Al-Masood, M.I. Asif Razee, M. Zunaed, Solar PV Power Forecasnting Using Traditional Methods and Machine Learning Techniques, in: Kansas Power and Energy Conference (KPEC), 2021: pp. 3–7. https://doi.org/10.1109/KPEC51835.2021.9446199.

[97] E. Garoudja, A. Chouder, K. Kara, S. Silvestre, An enhanced machine learning based approach for failures detection and diagnosis of PV systems, Energy Conversion and Management. 151 (2017) 496–513. https://doi.org/10.1016/j.enconman.2017.09.019.

[98] A. Mellit, A.M. Pavan, Performance prediction of 20 kWp grid-connected photovoltaic plant at Trieste (Italy) using artificial neural network, Energy Conversion and Management. 51 (2010) 2431–2441. https://doi.org/10.1016/j.enconman.2010.05.007.

[99] R.V.A. Monteiro, G.C. Guimarães, F.A.M. Moura, M.R.M.C. Albertini, M.K. Albertini, Estimating photovoltaic power generation: Performance analysis of artificial neural networks, Support Vector Machine and Kalman filter, Electric Power Systems Research. 143 (2017) 643–656. https://doi.org/10.1016/j.epsr.2016.10.050.

[100] M.O. Moreira, P.P. Balestrassi, A.P. Paiva, P.F. Ribeiro, B.D. Bonatto, Design of experiments using artificial neural network ensemble for photovoltaic generation forecasting, Renewable and Sustainable Energy Reviews. 135 (2021) 110450.

https://doi.org/10.1016/j.rser.2020.110450.

[101] A.M. Khalid, I. Mitra, W. Warmuth, V. Schacht, Performance ratio – Crucial parameter for grid connected PV plants, Renewable and Sustainable Energy Reviews. 65 (2016) 1139–1158. https://doi.org/10.1016/j.rser.2016.07.066.

[102] A. Livera, M. Theristis, G. Makrides, G.E. Georghiou, Recent advances in failure diagnosis techniques based on performance data analysis for grid-connected photovoltaic systems, Renewable Energy. 133 (2019) 126–143. https://doi.org/10.1016/j.renene.2018.09.101.

[103] S. Stettler, P. Toggweiler, E. Wiemken, W. Heidenreich, A.C. de Keizer, W.G.J.H.M. van Sark, S. Feige, M. Schneider, G. Heilscher, E. Lorenz, A. Drews5, D. Heinemann, H.G. Beyer, Failure detection routine for grid-connected PV systems as part of the PVSAT-2 project, in: 20th European Photovoltaic Solar Energy Conference & Exhibition, Barcelona (Spain), 2005: pp. 2490–2493. https://uol.de/f/5/inst/physik/ag/enmet/publications/solar/conference/2005/failure_detection_routine_for_grid_connected_pv_systems_as_part_of_PVSAT2_project.pdf.

[104] C.Y. Park, S.H. Hong, S.C. Lim, B.S. Song, S.W. Park, J.H. Huh, J.C. Kim, Inverter efficiency analysis model based on solar power estimation using solar radiation, Processes. 8 (2020) 1–19. https://doi.org/10.3390/pr8101225.

[105] H. Xing, B. Zhao, Z. Wang, Short-term Power Generation Prediction of Photovoltaic Panels Based on Meteorological Parameters and Support Vector Machine, Proceedings - 2020 Chinese Automation Congress, CAC 2020. (2020) 6018–6022. https://doi.org/10.1109/CAC51589.2020.9326631.

[106] J.S. and G.E.G. Andreas Livera, Alexander Phinikarides, George Makrides, Advanced failure detection algorithms and performance decision classification for grid-connected PV systems, in: PVSec 2017, EU PVSEC, 2007: pp. 23–42. https://www.eupvsec-proceedings.com/proceedings?paper=43125.

[107] T. AlSkaif, S. Dev, L. Visser, M. Hossari, W. van Sark, A systematic analysis of meteorological variables for PV output power estimation, Renewable Energy. 153 (2020) 12–22. https://doi.org/10.1016/j.renene.2020.01.150.

[108] Y. Zhao, L. Yang, B. Lehman, J.F. De Palma, J. Mosesian, R. Lyons, Decision tree-based fault detection and classification in solar photovoltaic arrays, in: Applied Power Electronics Conference and Exposition - APEC, Conference Proceedings - IEEE, 2012: pp. 93–99. https://doi.org/10.1109/APEC.2012.6165803.

[109] T. Tanev, R. Stanev, Modeling of photovoltaic power plant electricity generation using machine learning methods, 2021 17th Conference on Electrical Machines, Drives and Power Systems, ELMA 2021 - Proceedings. (2021) 1–4. https://doi.org/10.1109/ELMA52514.2021.9503066.

[110] M. De Benedetti, F. Leonardi, F. Messina, C. Santoro, A. Vasilakos, Anomaly detection and predictive maintenance for photovoltaic systems, Neurocomputing. 0 (2018) 1–10. https://doi.org/10.1016/j.neucom.2018.05.017.

[111] W. Chine, A. Mellit, V. Lughi, A. Malek, G. Sulligoi, A. Massi Pavan, A novel fault diagnosis technique for photovoltaic systems based on artificial neural networks, Renewable Energy. 90 (2016) 501–512. https://doi.org/10.1016/j.renene.2016.01.036.

[112] H. Mekki, A. Mellit, H. Salhi, Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules, Simulation Modelling Practice and Theory. 67 (2016) 1–13. https://doi.org/10.1016/j.simpat.2016.05.005.

[113] Syafaruddin, E. Karatepe, T. Hiyama, Controlling of artificial neural network for fault diagnosis of photovoltaic array, in: 16th International Conference on Intelligent System Applications to Power Systems, ISAP 2011, 2011: pp. 1–6. https://doi.org/10.1109/ISAP.2011.6082219.

[114] M.A. Sanz-Bobi, A. Muñoz San Roque, A. De Marcos, M. Bada, Intelligent system for a remote diagnosis of a photovoltaic solar power plant, in: 25th International Congress on Condition Monitoring and Diagnostic Engineering, Journal of Physics: Conference Series, 2012. https://doi.org/10.1088/1742-6596/364/1/012119.

[115] M. Trigo-Gonzalez, M. Cortés, J. Alonso-Montesinos, M. Martínez-Durbán, P. Ferrada, J. Rabanal, C. Portillo, G. López, F.J. Batlles, Development and comparison of PV production estimation models for mc-Si technologies in Chile and Spain, Journal of Cleaner Production. 281 (2021) 125360. https://doi.org/10.1016/j.jclepro.2020.125360.

[116] S. Theocharides, G. Makrides, E. George, A. Kyprianou, Machine Learning Algoriths for Photovoltaic System Power Output Prediction, 2018 IEEE International Energy Conference (ENERGYCON). (2018) 1–6. https://doi.org/10.1109/ENERGYCON.2018.8398737.

[117] E. Isaksson, M.K. Conde, Solar Power Forecasting with Machine Learning Techniques, KTH Vetenskap och konst, 2018. www.kth.se/sci.

[118] S. Theocharides, R. Alonso-Suarez, G. Giacosa, G. Makrides, M. Theristis, G.E. Georghiou, Intra-hour Forecasting for a 50 MW Photovoltaic System in Uruguay: Baseline Approach, Conference Record of the IEEE Photovoltaic Specialists Conference. (2019) 1632–1636. https://doi.org/10.1109/PVSC40753.2019.8980756.

[119] A.N. Sharkawy, M.M. Ali, H.H.H. Mousa, A.S. Ali, G.T. Abdel-Jaber, H.S. Hussein, M. Farrag, M.A. Ismeil, Solar PV Power Estimation and Upscaling Forecast Using Different Artificial Neural Networks Types: Assessment, Validation, and Comparison, IEEE Access. 11 (2023) 19279–19300. https://doi.org/10.1109/ACCESS.2023.3249108.

[120] X. Yang, M. Xu, S. Xu, X. Han, Day-ahead forecasting of photovoltaic output power with similar cloud space fusion based on incomplete historical data mining, Applied Energy. 206 (2017) 683–696. https://doi.org/10.1016/j.apenergy.2017.08.222.

[121] IEA-PVPS-Task14, Photovoltaic and Solar Forecasting: State of the Art, 2013. https://iea-pvps.org/key-topics/photovoltaics-and-solar-forecasting-state-of-art-report-t1401-2013/ (accessed November 27, 2023).

[122] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, Z. Hu, Photovoltaic and solar power forecasting for smart grid energy management, CSEE Journal of Power and Energy Systems. 1 (2015) 38–46. https://doi.org/10.17775/CSEEJPES.2015.00046.

[123] H. Haeberlin, C. Beutler, Normalized Representation of Energy and Power for Analysis of Performance and On-line Error Detection in PV-Systems, in: 13th EU PV Conference on Photovoltaic Solar Energy Conversion, Nice, France, 1995: pp. 1–4. https://www.semanticscholar.org/paper/Normalized-Representation-of-Energy-and-Power-for-Haeberlin/69e5d9080ffd1d09af8b5f250870a0a140387fa6.

[124] H. Zhu, P. Blakborow, Understanding Radiance (Brightness), Irradiance and Radiant Flux, 2011. https://www.energetiq.com/hubfs/Energetiq_March2019/PDF/Understanding Radiance (Brightness)%2C Irradiance and Radiant Flux - Mar 2011.pdf (accessed November 28, 2023).

[125] B. Kurtz, J. Kleissl, Measuring diffuse, direct, and global irradiance using a sky imager, Solar Energy. 141 (2017) 311–322. https://doi.org/10.1016/j.solener.2016.11.032.

[126] NREL, Best Practices Handbook for the collection and use of solar resource data for solar energy applications, 2012. https://www.nrel.gov/docs/fy21osti/77635.pdf.

[127] A. Lesterm, D.R. Myers, A method for improving global pyranometer measurements by modeling responsivity funtions, Solar Energy 80. (2006) 322–331. https://doi.org/10.1016/j.solener.2005.02.010.

[128] W.K. Tobiska, Measurement and modeling of solar EUV/UV radiation, Physics and Chemistry of the Earth, Part C: Solar, Terrestrial and Planetary Science. 25 (2000) 371–374. https://doi.org/10.1016/S1464-1917(00)00034-9.

[129] S. Mohanty, P.K. Patra, S.S. Sahoo, Prediction and application of solar radiation with soft computing over traditional and conventional approach - A comprehensive review, Renewable and Sustainable Energy Reviews. 56 (2016) 778–796. https://doi.org/10.1016/j.rser.2015.11.078.

[130] J. Polo, W.G. Fernandez-Neira, M.C. Alonso-Garc??a, On the use of reference modules as irradiance sensor for monitoring and modelling rooftop PV systems, Renewable Energy. 106 (2017) 186–191. https://doi.org/10.1016/j.renene.2017.01.026.

[131] M.A. Hassan, A. Khalil, S. Kaseb, M.A. Kassem, Independent models for estimation of daily global solar radiation: A review and a case study, Renewable and Sustainable Energy Reviews. 82 (2018) 1565–1575. https://doi.org/10.1016/j.rser.2017.07.002.

[132] D. Perez-Astudillo, D. Bachour, L. Martin-Pomares, Improved quality control protocols on solar radiation measurements, Solar Energy. 169 (2018) 425–433. https://doi.org/10.1016/j.solener.2018.05.028.

[133] A. Ohmura, E.G. Dutton, B. Forgan, C. Fröhlich, H. Gilgen, H. Hegner, A. Heimo, G. König-Langlo, B. McArthur, G. Müller, R. Philipona, R. Pinker, C.H. Whitlock, K. Dehne, M. Wild, Baseline Surface Radiation Network (BSRN/WCRP): New Precision Radiometry for Climate Research, Bulletin of the American Meteorological Society. 79 (1998) 2115–2136. https://doi.org/10.1175/1520-0477(1998)079<2115:BSRNBW>2.0.CO;2.

[134] C.N. Long, Y. Shi, An Automated Quality Assessment and Control Algorithm for Surface Radiation Measurements, The Open Atmospheric Science. 2 (2008) 23–37. https://doi.org/10.2174/1874282300802010023.

[135] I. Moradi, Quality control of global solar radiation using sunshine duration hours, Energy. 34 (2009) 1–6. https://doi.org/10.1016/j.energy.2008.09.006.

[136] R. Blaga, A. Sabadus, N. Stefu, C. Dughir, M. Paulescu, V. Badescu, A current perspective on the accuracy of incoming solar energy forecasting, Progress in Energy and Combustion Science. 70 (2019) 119–144. https://doi.org/10.1016/j.pecs.2018.10.003.

[137] R.H. Inman, H.T.C. Pedro, C.F.M. Coimbra, Solar forecasting methods for renewable

energy integration, Progress in Energy and Combustion Science. 39 (2013) 535–576. https://doi.org/10.1016/j.pecs.2013.06.002.

[138] S. Sobri, S. Koohi-Kamali, N.A. Rahim, Solar photovoltaic generation forecasting methods: A review, Energy Conversion and Management. 156 (2018) 459–497. https://doi.org/10.1016/j.enconman.2017.11.019.

[139] A. Teke, H.B. Yildirim, Ö. Çelik, Evaluation and performance comparison of different models for the estimation of solar radiation, Renewable and Sustainable Energy Reviews. 50 (2015) 1097–1107. https://doi.org/10.1016/j.rser.2015.05.049.

[140] C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, A. Fouilloy, Machine learning methods for solar radiation forecasting: A review, Renewable Energy. 105 (2017) 569–582. https://doi.org/10.1016/j.renene.2016.12.095.

[141] J. Zhang, B. Hodge, S. Lu, H.F. Hamann, B. Lehman, J. Simmons, E. Campos, V. Banunarayanan, J. Black, J. Tedesco, Baseline and target values for regional and point PV power forecasts : Toward improved solar forecasting, Solar Energy. 122 (2015) 804–819. https://doi.org/10.1016/j.solener.2015.09.047.

[142] X. Shao, S. Lu, H.F. Hamann, Solar radiation forecast with machine learning, Proceedings of AM-FPD 2016 - 23rd International Workshop on Active-Matrix Flatpanel Displays and Devices: TFT Technologies and FPD Materials. (2016) 19–22. https://doi.org/10.1109/AM-FPD.2016.7543604.

[143] J. Alonso-Montesinos, F.J. Batlles, Solar radiation forecasting in the short- and medium-term under all sky conditions, Energy. 83 (2015) 387–393. https://doi.org/10.1016/j.energy.2015.02.036.

[144] C. Rigollier, O. BAUER, L. WALD, On the clear sky model of the ESRA - European Solar Radiation Atlas - with respect to the Heliosar method, Solar Energy. 68 (2000) 33–48. https://doi.org/10.1016/S0038-092X(99)00055-9.

[145] A.B. Sproul, Derivation of the solar geometric relationships using vector analysis, Renewable Energy. 32 (2007) 1187–1205. https://doi.org/10.1016/j.renene.2006.05.001.

[146] O. Behar, A. Khellaf, K. Mohammedi, Comparison of solar radiation models and their validation under Algerian climate – The case of direct irradiance, Energy Conversion and Management. 98 (2015) 236–251. https://doi.org/10.1016/j.enconman.2015.03.067.

[147] R. Paulescu, Marius AU - Mares, Oana AU - Eugenia, Paulescu AU - Stefu, Nicoleta AU - Pacurar, Angel AU - Calinoiu, Delia AU - Gravila, Paul AU - Pop, Nicolina AU - Boata, Nowcasting solar irradiance using the sunshine numbe, Energy Conversion and Management ER. (2014). https://doi.org/10.1016/j.enconman.2013.12.048.

[148] D. Yang, Solar radiation on inclined surfaces: Corrections and benchmarks, Solar Energy. 136 (2016) 288–302. https://doi.org/10.1016/j.solener.2016.06.062.

[149] S.D. Miller, M.A. Rogers, J.M. Haynes, M. Sengupta, A.K. Heidinger, Short-term solar irradiance forecasting via satellite/model coupling, Solar Energy. 168 (2018) 102–117. https://doi.org/10.1016/j.solener.2017.11.049.

[150] A.A. Prasad, R.A. Taylor, M. Kay, Assessment of direct normal irradiance and cloud connections using satellite data over Australia, Applied Energy. 143 (2015) 301–311.

https://doi.org/10.1016/j.apenergy.2015.01.050.

[151] M. Tadj, K. Benmouiza, A. Cheknane, S. Silvestre, Improving the performance of PV systems by faults detection using GISTEL approach, Energy Conversion and Management. 80 (2014) 298–304. https://doi.org/10.1016/j.enconman.2014.01.030.

[152] M. Zamo, O. Mestre, P. Arbogast, O. Pannekoucke, A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production, Solar Energy. 105 (2014) 804–816. https://doi.org/10.1016/j.solener.2014.03.026.

[153] R. Perez, E. Lorenz, S. Pelland, M. Beauharnois, G. Van Knowe, K. Hemker, D. Heinemann, J. Remund, S.C. Müller, W. Traunmüller, G. Steinmauer, D. Pozo, J.A. Ruiz-Arias, V. Lara-Fanego, L. Ramirez-Santigosa, M. Gaston-Romero, L.M. Pomares, Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe, Solar Energy. 94 (2013) 305–326. https://doi.org/10.1016/j.solener.2013.05.005.

[154] D.P. Larson, L. Nonnenmacher, C.F.M. Coimbra, Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest, Renewable Energy. 91 (2016) 11–20. https://doi.org/10.1016/j.renene.2016.01.039.

[155] Z. Wang, I. Koprinska, M. Rana, Clustering Based Methods for Solar Power Forecasting, (2016) 1487–1494. https://doi.org/10.1109/IJCNN.2016.7727374.

[156] M. David, F. Ramahatana, P.J. Trombe, P. Lauret, Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models, Solar Energy. 133 (2016) 55–72. https://doi.org/10.1016/j.solener.2016.03.064.

[157] M. Bin Shams, S. Haji, A. Salman, H. Abdali, A. Alsaffar, Time series analysis of Bahrain's first hybrid renewable energy system, Energy. 103 (2016) 1–15. https://doi.org/10.1016/j.energy.2016.02.136.

[158] J.M. Corrêa, A.C. Neto, L.A. Teixeira Júnior, E.M.C. Franco, A.E. Faria, Time series forecasting with the WARIMAX-GARCH method, Neurocomputing. 216 (2016) 805–815. https://doi.org/10.1016/j.neucom.2016.08.046.

[159] C. Voyant, C. Paoli, M. Muselli, M.L. Nivet, Multi-horizon solar radiation forecasting for Mediterranean locations using time series models, Renewable and Sustainable Energy Reviews. 28 (2013) 44–52. https://doi.org/10.1016/j.rser.2013.07.058.

[160] M. De Felice, A. Alessandri, P.M. Ruti, Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models, Electric Power Systems Research. 104 (2013) 71–79. https://doi.org/10.1016/j.epsr.2013.06.004.

[161] H. Jiang, Y. Dong, A nonlinear support vector machine model with hard penalty function based on glowworm swarm optimization for forecasting daily global solar radiation, Energy Conversion and Management. 126 (2016) 991–1002. https://doi.org/10.1016/j.enconman.2016.08.069.

[162] F. Wang, Z. Zhen, Z. Mi, H. Sun, S. Su, G. Yang, Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting, Energy and Buildings. 86 (2015) 427–438. https://doi.org/10.1016/j.enbuild.2014.10.002.

[163] B. Wolff, J. Kühnert, E. Lorenz, O. Kramer, D. Heinemann, Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data, Solar Energy. 135 (2016) 197–208. https://doi.org/10.1016/j.solener.2016.05.051.

[164] K.-P. Lin, P.-F. Pai, Solar power output forecasting using evolutionary seasonal decomposition least-square support vector regression, Journal of Cleaner Production. 134 (2016) 456–462. https://doi.org/10.1016/j.jclepro.2015.08.099.

[165] L.M. Aguiar, B. Pereira, P. Lauret, F. Díaz, M. David, Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting, Renewable Energy. 97 (2016) 599–610. https://doi.org/10.1016/j.renene.2016.06.018.

[166] R.C. Deo, M. Şahin, Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland, Renewable and Sustainable Energy Reviews. 72 (2017) 828–848. https://doi.org/10.1016/j.rser.2017.01.114.

[167] A. Bâra, C. George, C. Botezatu, A. Pîrjan, Comparative Analysis between Wind and Solar Forecasting Methods Using Artificial Neural Networks, in: 16th IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary, 2015: pp. 89–94. https://doi.org/10.1109/CINTI.2015.7382900.

[168] L. Ma, N. Yorino, Solar Radiation (Insolation) Forecasting Using Constructive Neural Networks, in: International Joint Conference on Neural Networks (/JCNN), 2016: pp. 4991–4998. https://doi.org/10.1109/IJCNN.2016.7727857.

[169] O.I. Abiodun, A. Jantan, A.E. Omolara, K.V. Dada, N.A.E. Mohamed, H. Arshad, State-of-the-art in artificial neural network applications: A survey, Heliyon. 4 (2018) e00938. https://doi.org/10.1016/j.heliyon.2018.e00938.

[170] L. Wang, O. Kisi, M. Zounemat-Kermani, G.A. Salazar, Z. Zhu, W. Gong, Solar radiation prediction using different techniques: Model evaluation and comparison, Renewable and Sustainable Energy Reviews. 61 (2016) 384–397. https://doi.org/10.1016/j.rser.2016.04.024.

[171] A. Bugała, M. Zaborowicz, P. Boniecki, D. Janczak, K. Koszela, W. Czekała, A. Lewicki, Short-term forecast of generation of electric energy in photovoltaic systems, Renewable and Sustainable Energy Reviews. 81 (2018) 306–312. https://doi.org/10.1016/j.rser.2017.07.032.

[172] A.K. Yadav, S.S. Chandel, Solar radiation prediction using Artificial Neural Network techniques: A review, Renewable and Sustainable Energy Reviews. 33 (2014) 772–781. https://doi.org/10.1016/j.rser.2013.08.055.

[173] M. Bou-Rabee, S.A. Sulaiman, M.S. Saleh, S. Marafi, Using artificial neural networks to estimate solar radiation in Kuwait, Renewable and Sustainable Energy Reviews. 72 (2017) 434–438. https://doi.org/10.1016/j.rser.2017.01.013.

[174] F.-V. Gutierrez-Corea, M.-A. Manso-Callejo, M.-P. Moreno-Regidor, M.-T. Manrique-Sancho, Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations, Solar Energy. 134 (2016) 119–131. https://doi.org/10.1016/j.solener.2016.04.020.

[175] F. Davò, S. Alessandrini, S. Sperati, L. Delle Monache, D. Airoldi, M.T. Vespucci, Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting, Solar Energy. 134 (2016) 327–338. https://doi.org/10.1016/j.solener.2016.04.049.

[176] M. Alanazi, A. Alanazi, A. Khodaei, Long-term solar generation forecasting, Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference. 2016-July (2016) 1–5. https://doi.org/10.1109/TDC.2016.7519883.

[177] M. Ghofrani, M. Ghayekhloo, R. Azimi, A novel soft computing framework for solar radiation forecasting, Applied Soft Computing. 48 (2016) 207–216. https://doi.org/10.1016/j.asoc.2016.07.022.

[178] T. Burianek, S. Misak, Solar irradiance forecasting model based on extreme learning machine, (2016) 0–4. https://doi.org/10.1109/EEEIC.2016.7555445.

[179] H.T.C. Pedro, C.F.M. Coimbra, Short-term irradiance forecastability for various solar micro-climates, Solar Energy. 122 (2015) 587–602. https://doi.org/10.1016/j.solener.2015.09.031.

[180] E. Akarslan, F.O. Hocaoglu, A novel adaptive approach for hourly solar radiation forecasting, Renewable Energy. 87 (2016) 628–633. https://doi.org/10.1016/j.renene.2015.10.063.

[181] R. Chauvin, J. Nou, S. Thil, S. Grieu, Intra-Day DNI Forecasting Under Clear Sky Conditions Using ANFIS, IFAC Proceedings Volumes. 47 (2014) 10361–10366. https://doi.org/10.3182/20140824-6-ZA-1003.02087.

[182] Y.E. Midilli, S. Parsutins, Optimization of Deep Learning Hyperparameters with Experimental Design in Exchange Rate Prediction, 2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University, ITMS 2020 - Proceedings. (2020) 2020–2023. https://doi.org/10.1109/ITMS51158.2020.9259300.

[183] L.A. Demidova, A.V. Filatov, Optimization of hyperparameters with constraints on time and memory for the classification model of the hard drives states, 2022 36th International Conference on Information Technologies, InfoTech 2022 - Proceedings. (2022). https://doi.org/10.1109/InfoTech55606.2022.9897074.

[184] C. Pan, J. Tan, Day-ahead hourly forecasting of solar generation based on cluster analysis and ensemble model, IEEE Access. 7 (2019) 112921–112930. https://doi.org/10.1109/ACCESS.2019.2935273.

[185] L. Zhang, P.N. Suganthan, A survey of randomized algorithms for training neural networks, Information Sciences. 364–365 (2016) 146–155. https://doi.org/10.1016/j.ins.2016.01.039.

[186] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Hyperband: A novel bandit-based approach to hyperparameter optimization, Journal of Machine Learning Research. 18 (2018) 1–52. https://doi.org/10.48550/arXiv.1603.06560.

[187] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research. 13 (2012) 281–305. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

[188] S. Theocharides, G. Makrides, A. Livera, M. Theristis, P. Kaimakis, G.E. Georghiou, Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing, Applied Energy. 268 (2020) 115023. https://doi.org/10.1016/j.apenergy.2020.115023.

[189] M. Pan, C. Li, R. Gao, Y. Huang, H. You, T. Gu, F. Qin, Photovoltaic power forecasting based on a support vector machine with improved ant colony optimization, Journal of Cleaner Production. 277 (2020) 123948. https://doi.org/10.1016/j.jclepro.2020.123948.

[190] J.V. Frank Hutter, Lars Kotthoff, Automated machine learning Methods, Systems, Challenges, The Springer Series on Challenges in Machine Learning, 2020. https://doi.org/10.1007/978-3-030-05318-5.

[191] M.P. Ranjit, G. Ganapathy, K. Sridhar, V. Arumugham, Efficient deep learning hyperparameter tuning using cloud infrastructure: Intelligent distributed hyperparameter tuning with Bayesian optimization in the cloud, IEEE International Conference on Cloud Computing, CLOUD. 2019-July (2019) 520–522. https://doi.org/10.1109/CLOUD.2019.00097.

[192] J. Wu, X.Y. Chen, H. Zhang, L.D. Xiong, H. Lei, S.H. Deng, Hyperparameter optimization for machine learning models based on Bayesian optimization, Journal of Electronic Science and Technology. 17 (2019) 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120.

[193] M. Chai, F. Xia, S. Hao, D. Peng, C. Cui, W. Liu, PV Power Prediction Based on LSTM with Adaptive Hyperparameter Adjustment, IEEE Access. 7 (2019) 115473–115486. https://doi.org/10.1109/ACCESS.2019.2936597.

[194] S. Zhou, L. Zhou, M. Mao, X. Xi, Transfer learning for photovoltaic power forecasting with long short-term memory neural network, Proceedings - 2020 IEEE International Conference on Big Data and Smart Computing, BigComp 2020. (2020) 125–132. https://doi.org/10.1109/BigComp48618.2020.00-87.

[195] R. Nguyen, Y. Yang, A. Tohmeh, H.G. Yeh, Predicting PV Power Generation using SVM Regression, 2021 IEEE Green Energy and Smart Systems Conference, IGESSC 2021. (2021). https://doi.org/10.1109/IGESSC53124.2021.9618677.

[196] R. Khelifi, M. Guermoui, B. Laouar, A Novel Hybrid Model for PV Power Forecasting Using Support Vector Machine and Grasshopper Optimization Algorithm: Case Study, SIENR 2021 - 6th International Symposium on New and Renewable Energies. (2021) 18–21. https://doi.org/10.1109/SIENR50924.2021.9631916.

[197] Shafer and Zhang, Introductory Statistics, LibreTexts, 2022.

[198] W.J. Padgett, G.J. Hahn, W.Q. Meeker, Statistical Intervals: A Guide for Practitioners and Researchers, Wiley, 2017. https://doi.org/10.2307/2290749.

[199] N.A. Weiss, Introductory Statistic, 9th ed., Addison-Wesley (PEARSON), 2012. http://www.pearson.com/us/higher-education/product/Weiss-Introductory-Statistics-10th-Edition/9780321989178.html.

[200] O. Kirchkamp, Resampling Methods, Technometrics. 44 (2002) 299–299. https://doi.org/10.1198/004017002320256611.

[201] R.J.H. and G. Athanasopoulos, 3.5 Prediction intervals | Forecasting: Principles and

Practice (2nd ed), in: Forecasting: Principles and Practice (2nd Ed), 2nd ed., Monash University, Australia, 2018. https://otexts.com/fpp2/prediction-intervals.html (accessed November 27, 2023).

[202] J.J. Buckley, Studies in Fuzziness and Soft Computing: Foreword, Springer-Verlag, 2011.

[203] B. Rasmussen, J.W. Hines, Prediction interval estimation techniques for empirical modeling strategies and their applications to signal validation tasks, Applied Computational Intelligence - Proceedings of the 6th International FLINS Conference. (2004) 549–556. https://doi.org/10.1142/9789812702661_0099.

[204] D.M. Lane, S. David, M. Hebl, R. Guerra, D. Osherson, H. Zimmer, Introduction to Statistics, Open Textbook Library, 2003. https://open.umn.edu/opentextbooks/textbooks/459.

[205] D. Alhakeem, P. Mandal, A.U. Haque, A. Yona, T. Senjyu, T.L. Tseng, A new strategy to quantify uncertainties of wavelet-GRNN-PSO based solar PV power forecasts using bootstrap confidence intervals, IEEE Power and Energy Society General Meeting. 2015-Septe (2015) 1–5. https://doi.org/10.1109/PESGM.2015.7286233.

[206] F. Rodríguez, A. Galarza, J.C. Vasquez, J.M. Guerrero, Using deep learning and meteorological parameters to forecast the photovoltaic generators intra-hour output power interval for smart grid control, Energy. 239 (2022) 122116. https://doi.org/10.1016/j.energy.2021.122116.

[207] T. Ishizaki, M. Koike, N. Ramdani, Y. Ueda, T. Masuta, T. Oozeki, T. Sadamoto, J.I. Imura, Interval quadratic programming for day-ahead dispatch of uncertain predicted demand, Automatica. 64 (2016) 163–173. https://doi.org/10.1016/j.automatica.2015.11.002.

[208] S. Keydana, Adding uncertainty estimates to Keras models with tfprobability, RStudio AI Blog. (2019). https://blogs.rstudio.com/ai/posts/2019-06-05-uncertainty-estimates-tfprobability/ (accessed November 27, 2023).

[209] TensorFlow, Probability, (n.d.). https://www.tensorflow.org/probability (accessed November 28, 2023).

[210] M. Ma, B. He, R. Shen, Y. Wang, N. Wang, An adaptive interval power forecasting method for photovoltaic plant and its optimization, Sustainable Energy Technologies and Assessments. 52 (2022) 102360. https://doi.org/10.1016/j.seta.2022.102360.

[211] S. Dang, L. Peng, J. Zhao, J. Li, Z. Kong, A Quantile Regression Random Forest-Based Short-Term Load Probabilistic Forecasting Method, Energies. 15 (2022) 1–20. https://doi.org/10.3390/en15020663.

[212] D. Saattrup Nielsen, Quantile regression forests, Quantile Regression Forest. (2020). https://saattrupdan.github.io/2020-04-05-quantile-regression-forests/ (accessed November 28, 2023).

[213] TensorFlow, Regression with Probabilistic Layers in TensorFlow Probability, (n.d.). https://blog.tensorflow.org/2019/03/regression-with-probabilistic-layers-in.html?_gl=1*1ufllhl*_ga*MTYwNTUyNzAzLjE2ODIzMjAyOTk.*_ga_W0YLR4190T*MTY5MzM5MDg2OS42LjEuMTY5MzM5MTU3OC4wLjAuMA.. (accessed November 28, 2023).

[214] X. Gang Su, Linear regression analysis, World Scientific Publishing Co., 2009.

https://doi.org/10.1142/6986.

[215] J.M. Rojo Abuí, Regresión lineal múltiple, Ieg, 2007. http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/Regresion_lineal_multiple_3.pdf (accessed November 21, 2023).

[216] Zhi-Hua Zhou, Ensemble Methods Foundations and Algorithms, CRC Press Taylor &Francis Group, 2012.

[217] J.O. Alvear, Arboles de decision y Random Forest, Bookdown.Org. (2018). https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html (accessed November 21, 2023).

[218] C. Zhang, Y. Ma, Ensemble Machine Learning, Springer, 2012. https://doi.org/10.1007/978-1-4419-9326-7.

[219] L. Breiman, Random Forests, Machine Learning. 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.

[220] J.H. Friedman, Stochastic gradient boosting, Computational Statistics and Data Analysis. 38 (2002) 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

[221] J.A. Rodrigo, Árboles de decisión, random forest, gradient boosting y C5.0, Ciencia de Los Datos. (2007). https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_C50.html (accessed November 26, 2023).

[222] T. Islam, P.K. Srivastava, M. Gupta, X. Zhu, S. Mukherjee, Computational Intelligence Techniques in Earth and Environmental Sciences, Springer, 2014. https://doi.org/10.1007/978-94-017-8642-3.

[223] M. Claesen, B. De Moor, Hyperparameter Search in Machine Learning, in: The XI Metaheuristics International Conference, 2015: pp. 10–14. http://arxiv.org/abs/1502.02127.

[224] R. Koenker, Quantile Regression in R: A Vignette, Quantile Regression. (2010) 295–316. https://doi.org/10.1017/cbo9780511754098.011.

[225] N. Meinshausen, Quantile Regression Forests, Journal of Machine Learning Research. 67 (2006) 983–999. https://stat.ethz.ch/~nicolai/quantregforests.pdf (accessed November 27, 2023).

[226] L. Schiesser, Quantile Regression Forests - An R-Vignette, 2015. https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf (accessed November 27, 2023).

[227] F. Martínez-Álvarez, A. Troncoso, G. Asencio-Cortés, J. Riquelme, A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting, 2015. https://doi.org/10.3390/en81112361.

[228] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna Austria. (2021). https://www.r-project.org/ (accessed November 20, 2022).

[229] J.F.M. Pessanha, A.C.G. Melo, R.P. Caldas, D.M. Falcao, A Methodology for Joint Data Cleaning of Solar Photovoltaic Generation and Solar Irradiation, in: 2020 International

Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2020 - Proceedings, 2020. https://doi.org/10.1109/PMAPS47429.2020.9183488.

[230] C. Rigollier, O. Bauer, L. Wald, On the clear sky model of the ESRA - European Solar Radiation Atlas - With respect to the Heliosat method, Solar Energy. 68 (2000) 33–48. https://doi.org/10.1016/S0038-092X(99)00055-9.

[231] SoDa, Linke Turbidity factor, Ozone, Water Vapor and Angstroembeta - www.soda-pro.com, Solar Radiation Data. (n.d.). https://www.soda-pro.com/help/general-knowledge/linke-turbidity-factor (accessed November 28, 2023).

[232] F. Chabane, N. Moummi, A. Brima, A New Approach to Estimate the Distribution of Solar Radiation Using Linke Turbidity Factor and Tilt Angle, Iranian Journal of Science and Technology - Transactions of Mechanical Engineering. 45 (2021) 523–534. https://doi.org/10.1007/s40997-020-00382-5.