Análisis de datos y modelado predictivo mediante aprendizaje automático para apoyar la operación de una central fotovoltaica: caso Puerto Peñasco y aplicaciones potenciales de IA generativa

Tesis para obtener el grado de Maestro en Ingeniería con opción terminal en Sistemas Eléctricos de Potencia

Por

MARIO SÁNCHEZ GUTIÉRREZ

Benemérita Universidad Autónoma de Puebla

FACULTAD DE INGENIERÍA

Investigación y Estudios de Posgrado

FACULTAD DE INGENIERÍA BUAP

Asesor

M.C. Ismael Albino Padilla

Coasesor

Dr. Juan Carlos Escamilla Sánchez

MARZO 2026

Abstract

La creciente integración de energías renovables en los sistemas eléctricos nacionales plantea desafíos operativos que van más allá de la simple ampliación de capacidad instalada. La naturaleza intrínsecamente variable de fuentes como la solar fotovoltaica introduce incertidumbre en la planificación del despacho, dificulta el equilibrio entre oferta y demanda en tiempo real, y exige infraestructuras de monitoreo, comunicación y análisis cualitativamente superiores a las requeridas por las centrales convencionales de generación base. En este contexto, los avances sostenidos en cómputo de alto desempeño, el desarrollo de ecosistemas de software de código abierto para análisis de datos, y la proliferación de sistemas de monitoreo de alta resolución han creado las condiciones técnicas para que la inteligencia artificial y el aprendizaje automático (ML) dejen de ser herramientas experimentales y se conviertan en componentes operativos viables dentro del sector eléctrico.

Esta investigación desarrolla un enfoque orientado al análisis de información operativa y al modelado predictivo, utilizando como caso de estudio la Central Fotovoltaica de Puerto Peñasco, en Sonora, México, la instalación de mayor escala en América Latina y una pieza estratégica en la integración de Baja California al Sistema Eléctrico Nacional (SEN). A partir de un conjunto de 2,016 registros horarios correspondientes al periodo diciembre de 2023 – agosto de 2025, que incluyen variables meteorológicas como temperatura ambiente, humedad relativa, velocidad del viento, visibilidad, presión atmosférica y cobertura de nubes, se construyen y comparan cuatro arquitecturas de modelos predictivos con el propósito de estimar la potencia de salida fotovoltaica. El objetivo no es únicamente técnico: se busca demostrar que este tipo de análisis es viable con los registros disponibles en una central real en operación, y que sus resultados tienen utilidad directa para la planeación y la toma de decisiones.

La metodología adoptada es de naturaleza mixta. Por un lado, se realiza una revisión sistemática del estado del arte en tecnologías de gestión de centrales fotovoltaicas, incluyendo sistemas SCADA, redes eléctricas inteligentes (*smart grids*), protocolos de comunicación industrial, algoritmos de optimización y diagnóstico, y modelos de pronóstico energético. Por otro lado, se lleva a cabo un trabajo analítico cuantitativo sobre los registros reales de Puerto Peñasco, que incluye exploración estadística de las series temporales, selección formal de características mediante eliminación recursiva con validación cruzada, y entrenamiento comparativo de modelos que van desde un regresor ficticio como línea base hasta arquitecturas de redes neuronales y bosques

aleatorios. El desempeño de cada modelo se evalúa mediante métricas estándar, error cuadrático medio (RMSE), error absoluto medio (MAE) y coeficiente de determinación (R^2); que permiten cuantificar objetivamente la precisión alcanzada y fundamentar la elección del enfoque más adecuado para el contexto operativo de la central.

De manera complementaria, y como extensión prospectiva del trabajo, se discuten las aplicaciones potenciales de la Inteligencia Artificial Generativa (GenAI) y los modelos de lenguaje de gran escala (LLMs) como capa de soporte para la interpretación de resultados, la generación automatizada de reportes operativos y la asistencia a la toma de decisiones en lenguaje natural. Esta dimensión no constituye el desarrollo de un sistema funcional dentro del alcance de la tesis, sino una exploración fundamentada en literatura reciente sobre las posibilidades que estas tecnologías abren para el personal técnico-operativo de instalaciones fotovoltaicas.

En conjunto, este trabajo contribuye al conocimiento disponible sobre la aplicación de técnicas de analítica e inteligencia artificial en la generación fotovoltaica a escala industrial, propone lineamientos metodológicos replicables en otras instalaciones, e identifica las condiciones de infraestructura de datos, comunicación y gobernanza de la información que hacen viable una adopción progresiva y sostenida de estas soluciones en el contexto del sistema eléctrico mexicano.

Agradecimientos

El presente trabajo no habría sido posible sin la ayuda y el consejo de mis padres, Blanca Estela Gutiérrez Domínguez y José Fidel Sánchez Fierro, quienes han sido mi mayor fuente de apoyo y motivación durante todo este proceso. No hay palabras suficientes para describir el amor, la paciencia y el sacrificio que han demostrado en cada paso de mi vida. Desde las primeras decisiones académicas hasta los momentos de mayor incertidumbre en este camino, siempre encontré en ellos una orientación serena y una convicción inquebrantable en mis capacidades.

A lo largo de este proceso, su respaldo incondicional fue fundamental para superar los momentos difíciles y seguir adelante con determinación. Su confianza en mí y las enseñanzas que han depositado a lo largo de los años me dieron la fortaleza necesaria para enfrentar los retos que implica una investigación de esta naturaleza, y por ello les estoy eternamente agradecido. Este logro es, en una medida que no sabría cuantificar, también el resultado de su entrega.

También quiero agradecer a mi querida hermana, Mariam Sánchez Gutiérrez, por su amor, su comprensión y por estar siempre a mi lado, compartiendo tanto las alegrías como las dificultades del camino. Su presencia constante ha sido un pilar emocional fundamental, y su cariño sincero me ha impulsado a seguir adelante en cada etapa, recordando siempre por qué vale la pena el esfuerzo.

Este trabajo no es únicamente el resultado de mi esfuerzo individual, sino también de la dedicación, el amor y la generosidad que siempre me han brindado quienes me rodean. Gracias por creer en mis sueños, por ser mi guía en los momentos de duda y por haberme dado la oportunidad de crecer, aprender y construir una versión más completa de mí mismo a través de esta experiencia.

Declaración del Autor

Yo declaro que el trabajo de esta tesis se ha realizado de acuerdo con los requisitos del Reglamento y Código de Prácticas para Programas de Títulos de Investigación de la Universidad y que no ha sido presentado para ningún otro premio académico. Excepto cuando se indique mediante referencia específica en el texto, el trabajo es obra del candidato. Los trabajos realizados en colaboración o con la ayuda de terceros se indican como tales. Las opiniones expresadas en la tesis son las del autor.

Firmado: Mario Sánchez Gutiérrez Fecha: Marzo 2026

ÍNDICE

Página

List of Tables **xi**

List of Figures **xiii**

Introducción **1**

Bibliography **7**

I Revisión de Tecnologías Emergentes para la Operación y Gestión de Centrales Fotovoltaicas: Tendencias y Retos Actuales **9**

1.1 *Smart Grids* en México y el Mundo 10

1.1.1 Alemania 11

1.1.2 Canadá 12

1.1.3 Chile 12

1.1.4 España 13

1.1.5 Estados Unidos 13

1.1.6 Japón 14

1.1.7 Uruguay 15

1.1.8 México 15

1.2 Tecnologías en el sector eléctrico actual 16

1.2.1 Recolección, procesamiento y análisis de datos 17

1.2.2 Utilización de algoritmos y herramientas de cómputo 19

1.3 Nuevas tecnologías en el sector eléctrico 20

1.3.1 Modelos predictivos de generación solar: estado del arte 21

1.3.2 Inteligencia Artificial Generativa como capa de soporte operativo 23

1.3.3 El caso de Puerto Peñasco como banco de prueba 24

Bibliography **27**

II Aprendizaje Automático Aplicado a la Operación Fotovoltaica: Metodología, Análisis de Datos y Modelos Predictivos sobre Registros Reales de Puerto Peñasco **31**

2.1 El Caso de Estudio: Central Fotovoltaica de Puerto Peñasco 32
2.1.1 Caracterización del emplazamiento y condiciones operativas 32
2.1.2 Descripción del conjunto de datos 34
2.2 Análisis Exploratorio de Datos 36
2.2.1 Distribución y comportamiento de la variable objetivo 37
2.2.2 Análisis de series temporales y patrones de generación 38
2.2.3 Análisis de correlaciones entre variables 38
2.3 Preprocesamiento y Preparación de los Datos 39
2.3.1 Partición del conjunto de datos en entrenamiento y prueba 40
2.3.2 Codificación de la variable categórica Season 41
2.3.3 Estandarización de las variables predictoras 41
2.4 Selección Formal de Características mediante RFECV 43
2.4.1 Fundamento metodológico 43
2.4.2 Resultado: número óptimo de características y su importancia relativa . . 44
2.4.3 Interpretación física del perfil de importancias 45
2.5 Descripción de los Modelos Implementados 46
2.5.1 Regresor Ficticio (DummyRegressor): Línea Base 46
2.5.2 Regresor por Descenso de Gradiente Estocástico (SGDRegressor) 47
2.5.3 Perceptrón Multicapa (MLPRegressor) 49
2.5.4 Bosque Aleatorio (RandomForestRegressor) 52
2.6 Protocolo de Evaluación y Métricas de Desempeño 55
2.6.1 Fundamento del protocolo de evaluación sobre conjunto de prueba independiente 55
2.6.2 Métricas de evaluación 55
2.6.3 Curvas de aprendizaje y diagnóstico 57
2.7 El Rol Potencial de los LLMs como Capa de Interpretación Operativa 58
2.8 Conclusiones del Capítulo II 59

Bibliography **61**

III Resultados, Análisis Comparativo y Perspectivas de Implementación: Del Modelo Predictivo a la Operación Fotovoltaica Inteligente en Puerto Peñasco **65**

3.1 Resultados del Desempeño Predictivo 67
3.1.1 Regresor ficticio (DummyRegressor): línea base 68
3.1.2 Regresor por descenso de gradiente estocástico (SGDRegressor) 70
3.1.3 Bosque aleatorio (RandomForestRegressor) 73

3.1.4 Perceptrón multicapa (MLPRegressor) 79
3.1.5 Tabla comparativa consolidada 83
3.2 Análisis Comparativo de los Modelos 87
3.2.1 La brecha SGD–RFR/MLP como evidencia de no linealidad 87
3.2.2 MLP versus RFR: la ganancia marginal de la complejidad 88
3.2.3 El regresor ficticio como referencia metodológica 89
3.2.4 Análisis de los patrones de residuos 89
3.3 Contextualización de los Resultados frente a la Literatura 90
3.3.1 Comparación con el estudio de referencia de Pasion et al. 90
3.3.2 Posicionamiento frente a estudios internacionales recientes 90
3.3.3 Factores que explican el posicionamiento de Puerto Peñasco 91
3.4 Implicaciones Operativas para la Central de Puerto Peñasco 91
3.4.1 Pronóstico para el despacho programado 91
3.4.2 Detección de anomalías y mantenimiento predictivo 92
3.4.3 Planificación de la generación y gestión de la interconexión 93
3.5 Perspectivas de Implementación y el Rol de la IA Generativa 93
3.5.1 Lineamientos para la implementación operativa 93
3.5.2 El rol potencial de los LLMs como capa de interpretación 94
3.6 Limitaciones del Estudio y Líneas de Trabajo Futuro 94
3.6.1 Limitaciones del estudio 94
3.6.2 Líneas de trabajo futuro 95
3.7 Conclusiones del Capítulo III 95

Bibliography **97**

List of Tables

Tabla **Página**

2.1 Estadísticos descriptivos del conjunto de datos de Puerto Peñasco. 36

3.1 Desempeño comparativo de los cuatro modelos sobre el conjunto de prueba ($n = 391$). Los valores en negritas indican el mejor desempeño por métrica. 85

List of Figures

Imagen **Página**

3.1 Predicciones del regresor ficticio contra valores reales de PolyPwr. La línea horizontal confirma la predicción constante de la media del conjunto de entrenamiento. 69
3.2 Diagrama de residuos del regresor ficticio. El patrón revela la estructura completa de la variabilidad de PolyPwr que el modelo no captura. 69
3.3 Curva de aprendizaje del regresor ficticio para R^2. 70
3.4 Curva de aprendizaje del regresor ficticio para RMSE. 70
3.5 Curva de aprendizaje del regresor ficticio para MAE. 71
3.6 Diagrama de distribución conjunta (JointPlot) del regresor ficticio. 71
3.7 Predicciones del SGD contra valores reales de PolyPwr. La dispersión pronunciada revela la incapacidad del modelo lineal para capturar las interacciones no lineales del problema. 72
3.8 Diagrama de residuos del SGD. El patrón fuertemente estructurado y no aleatorio confirma el sesgo sistemático del modelo lineal. 73
3.9 Diagrama de distribución conjunta (JointPlot) del SGD. 74
3.10 Curva de aprendizaje del SGD optimizado para R^2. 74
3.11 Curva de aprendizaje del SGD optimizado para RMSE. 75
3.12 Curva de aprendizaje del SGD optimizado para MAE. 75
3.13 Curva de validación del SGD. 76
3.14 Validación cruzada K-Fold (3 particiones) del SGD. 76
3.15 Predicciones del bosque aleatorio contra valores reales de PolyPwr. La concentración de los puntos sobre la línea de identidad refleja un R^2 de 0.991. 77
3.16 Validación cruzada K-Fold (3 particiones) del bosque aleatorio. Los tres pliegues alcanzan $R^2 = 0.99$. 78
3.17 Diagrama de residuos del bosque aleatorio. La distribución aproximadamente simétrica con incremento de dispersión para valores predichos altos indica heterocedasticidad moderada. 78
3.18 Curva de aprendizaje del bosque aleatorio para R^2. 79
3.19 Curva de aprendizaje del bosque aleatorio para RMSE. 79
3.20 Curva de aprendizaje del bosque aleatorio para MAE. 80

3.21 Diagrama de distribución conjunta (JointPlot) del bosque aleatorio. 80
3.22 Predicciones del MLP contra valores reales de PolyPwr. Con un R^2 de 0.994, el MLP exhibe la concentración más estrecha alrededor de la línea de identidad. 81
3.23 Validación cruzada K-Fold (3 particiones) del MLP. La consistencia con $R^2 = 0.99$ en cada pliegue confirma la estabilidad del modelo. 82
3.24 Diagrama de residuos del MLP. La distribución presenta heterocedasticidad similar a la del RFR, con un leve sesgo negativo en el rango de predicción media. 82
3.25 Curva de aprendizaje del MLP para R^2. 83
3.26 Curva de aprendizaje del MLP para RMSE. 83
3.27 Curva de aprendizaje del MLP para MAE. 84
3.28 Diagrama de distribución conjunta (JointPlot) del MLP. 84
3.29 Comparación del MAE por modelo. La reducción de más del 90% en el error absoluto entre los modelos no lineales y la línea base. 85
3.30 Comparación del RMSE por modelo. 85
3.31 Comparación del R^2 por modelo. El contraste entre el $R^2 \approx 0$ del Dummy, el 0.347 del SGD y los valores superiores a 0.99 del RFR y MLP evidencia la estructura no lineal del problema. 86
3.32 Distribución KDE de los valores predichos por cada modelo frente a los valores reales de PolyPwr. Las curvas del RFR y el MLP se superponen casi exactamente con la distribución bimodal real. 86

Introducción

En México, la generación eléctrica se sustentó durante décadas en fuentes convencionales como el carbón, el gas natural y los recursos hidroeléctricos. Este modelo energético, si bien permitió sostener el crecimiento industrial del país durante la segunda mitad del siglo XX, evidenció con el tiempo una fragilidad estructural significativa: su dependencia de insumos sujetos a la volatilidad de los mercados internacionales de hidrocarburos y a las fluctuaciones de los ciclos hidrológicos. Fue precisamente en ese contexto de vulnerabilidad cuando, a partir de la década de 1970, impulsada por la creación del Consejo Nacional de Ciencia y Tecnología (Conacyt) y la labor pionera de organismos como el Centro de Ecodesarrollo, A. C., comenzó a cobrar relevancia el aprovechamiento de energías limpias como alternativa estratégica de largo plazo. Este interés no surgió de forma espontánea: se intensificó como consecuencia directa de los conflictos geopolíticos en el Medio Oriente y, en particular, del embargo petrolero decretado en octubre de 1973 por la Organización de Países Árabes Exportadores de Petróleo (OAPEP), el alza sostenida en los precios del crudo que se prolongó durante toda esa década puso en evidencia ante el mundo entero la insostenibilidad de un modelo energético construido casi exclusivamente sobre el petróleo y sus derivados, e impulsó en numerosos países, México incluido, la búsqueda urgente de fuentes alternativas capaces de reducir esa dependencia estructural.

En ese contexto de búsqueda de alternativas, en 1976 el doctor Pablo Mulás del Pozo, quien había dirigido los Laboratorios del Reactor Nuclear del Centro Nuclear de México entre 1970 y 1973, y cuya trayectoria lo situaba en la intersección entre la investigación científica avanzada y la política energética nacional, presentó el reporte de investigación titulado “El aprovechamiento de la energía solar” [1]. Este documento representó un hito en la historia de la energía solar en México, no solo por la sistematización que ofrecía del potencial de distintas formas de aprovechamiento energético; incluyendo la térmica, la fotovoltaica, la fotoquímica, la eólica, los gradientes térmicos oceánicos, la biológica y las aplicaciones de climatización, sino también por la solidez cuantitativa de sus argumentos. En particular, el autor calculó que la radiación solar recibida sobre el territorio mexicano ascendía a 1.26×10^{22} Joules anuales, una cifra equivalente a cinco mil veces la energía primaria total consumida en el país durante 1975. Este resultado no era únicamente un dato de impacto: constituía la demostración técnica de que México poseía, en su recurso solar, una fuente de energía de escala prácticamente ilimitada para sus necesidades proyectadas, y que la pregunta pertinente no era si era posible aprovecharla, sino cuándo y mediante qué tecnologías

hacerlo de manera económicamente viable.

La respuesta a esa pregunta llegó, con décadas de distancia, de la mano del progreso tecnológico. Fue durante las primeras dos décadas del siglo XXI cuando los avances en la fabricación de semiconductores, los procesos de producción en masa y las economías de escala redujeron de manera dramática el costo de los módulos fotovoltaicos, transformando la energía solar de una tecnología de nicho, restringida a aplicaciones espaciales o a instalaciones remotas donde el tendido de red resultaba inviable, en una opción competitiva para la generación eléctrica a escala de servicio público. En México, ese punto de inflexión se materializó el 14 de octubre de 2012, cuando se inauguró la Central Solar Fotovoltaica "Santa Rosalía" en Baja California Sur, con una capacidad instalada de 1 megawatt (MW) y una producción neta anual estimada de 2,000 megawatt-hora (MWh) [2]. Aunque modesta en términos de escala, esta instalación fue un hito simbólico de primer orden: demostró que la generación fotovoltaica conectada a la red era técnicamente viable en condiciones reales de operación en el país y abrió el camino para un despliegue progresivamente más ambicioso. Para 2014 ya operaban nueve centrales de este tipo en territorio nacional y, de acuerdo con los registros de la Secretaría de Energía (SENER), el número de instalaciones continuó creciendo de forma sostenida durante la década siguiente, impulsado tanto por las condiciones del recurso solar como por marcos regulatorios que fueron abriendo gradualmente el sector a nuevos actores [3].

Ese crecimiento sostenido obedece, en buena medida, a las extraordinarias condiciones del recurso solar disponible en el territorio nacional. México registra una irradiación media anual que oscila entre 4.4 y 6.3 kWh/m^2, dependiendo de la región geográfica, valores que lo ubican consistentemente entre los países con mayor potencial fotovoltaico del mundo [4]. Esta riqueza del recurso no es, sin embargo, uniforme: la heterogeneidad climática y orográfica del país implica que las condiciones de aprovechamiento varían de forma significativa de una región a otra. Estados como Sonora, Chihuahua, Baja California y Sinaloa presentan una irradiación especialmente elevada con relativamente poca variabilidad estacional, lo que los convierte en emplazamientos ideales para centrales de gran escala. Zonas del sureste, en cambio, aunque con alta radiación total, enfrentan períodos de mayor nubosidad y precipitación que introducen variabilidad operativa adicional. Esta diversidad climática subraya la importancia de contar con herramientas analíticas capaces de caracterizar el recurso con precisión local y de modelar su comportamiento a distintas escalas temporales.

Bajo esta perspectiva, el 17 de febrero de 2023 se inauguró la primera etapa de la Central Fotovoltaica de Puerto Peñasco, ubicada en el municipio del mismo nombre en el estado de Sonora [5]. Desde el momento de su apertura, esta instalación fue reconocida como la más grande de América Latina y la quinta en capacidad a nivel mundial, atributos que por sí solos justificaría su relevancia. Sin embargo, el significado estratégico del proyecto va mucho más allá de los récords de capacidad: Puerto Peñasco materializa, por primera vez, la interconexión permanente de Baja California al Sistema Eléctrico Nacional (SEN) mediante una línea de transmisión de

290 kilómetros que constituye la segunda más importante de la Red Nacional de Transmisión. Esta línea pone fin a décadas de operación aislada de la península bajacaliforniana, que hasta entonces debía sostener su suministro eléctrico con una combinación de generación local y conexiones internacionales con el sistema estadounidense. En su etapa final de desarrollo, la central contempla una extensión aproximada de 2,000 hectáreas y una capacidad instalada cercana a 1 gigawatt (GW), suficiente para abastecer las necesidades eléctricas de aproximadamente 1.6 millones de mexicanos, lo que la convierte en un activo estratégico de primera magnitud para la transición energética del país.

La puesta en marcha de proyectos de la magnitud de Puerto Peñasco refleja el compromiso explícito de México con sus metas de transformación energética. La Secretaría de Energía (SENER), en el documento "Prospectiva de Energías Renovables 2016–2030", establece como objetivo nacional alcanzar el 45% de generación eléctrica mediante fuentes limpias para ese año [3]. Este objetivo no puede lograrse únicamente sumando capacidad instalada. El verdadero reto operativo se manifiesta cuando la participación de fuentes variables como la solar alcanza proporciones significativas dentro del mix de generación: en ese punto, la variabilidad inherente al recurso, condicionada por el ciclo día-noche, la nubosidad, las estaciones del año y los fenómenos meteorológicos extremos, comienza a ejercer presión sobre la capacidad del sistema para mantener el equilibrio entre generación y demanda en tiempo real. Gestionar esa variabilidad con eficiencia y sin comprometer la estabilidad del SEN requiere de herramientas analíticas que vayan mucho más allá de los esquemas de operación diseñados para fuentes de generación base. México, con una irradiación media anual superior a 6 kWh/m^2 en amplias franjas de su territorio, posee las condiciones físicas para liderar esta transición en la región [4], pero necesita construir en paralelo las capacidades analíticas, de infraestructura de datos y de gestión operativa que permitan traducir esa ventaja natural en operación confiable, eficiente y predecible.

En este escenario, la industria 4.0 y el *Internet of Things* (IoT) han transformado profundamente el sector eléctrico al ampliar radicalmente las capacidades de recolección e interpretación de señales operativas. La proliferación de sensores de bajo costo, las redes de comunicación de alta velocidad y las plataformas de almacenamiento y procesamiento de datos en la nube han hecho posible lo que hace dos décadas era técnicamente inviable en muchos contextos: monitorear en tiempo real y de forma continua cada componente de una central de generación, desde el nivel de string individual hasta el transformador de interconexión. Gracias a esta infraestructura, tareas como la predicción de generación y demanda, la detección temprana de degradación en módulos o inversores, el mantenimiento predictivo basado en datos y la optimización del punto de operación del campo fotovoltaico han pasado de ser aspiraciones de la ingeniería de sistemas a convertirse en prácticas consolidadas en instalaciones modernas de todo el mundo [6]. A este panorama se suma ahora la irrupción de la Inteligencia Artificial Generativa (GenAI), que abre una dimensión adicional de posibilidades: la de procesar y sintetizar grandes volúmenes de información técnica, generar automáticamente reportes operativos y asistir a los tomadores de

decisión en la interpretación de indicadores de desempeño complejos, todo ello en lenguaje natural y con tiempos de respuesta próximos al tiempo real [7].

En ese sentido, la Central Fotovoltaica de Puerto Peñasco reúne las características que la convierten en un caso de estudio especialmente valioso para explorar estas posibilidades. Su escala, su complejidad tecnológica, que incluye sistemas de almacenamiento en baterías, la riqueza del recurso solar en su emplazamiento y el papel crítico que desempeña en la integración de Baja California al SEN hacen de ella un laboratorio idóneo para validar estrategias de analítica avanzada con impacto operativo real. La capacidad de predecir con precisión su generación horaria, de detectar anomalías antes de que se conviertan en fallas costosas y de optimizar el ciclo de carga y descarga de sus sistemas de almacenamiento son problemas cuya solución mediante herramientas de ML tiene consecuencias directas sobre la confiabilidad del suministro a escala regional y sobre los costos de operación y mantenimiento de la instalación.

La investigación adopta un enfoque metodológico mixto que combina de forma deliberada y complementaria dos tradiciones de producción de conocimiento. Por un lado, la revisión sistemática de publicaciones científicas arbitradas, documentos técnicos institucionales y reportes de organismos internacionales, permite construir un marco conceptual sólido sobre el estado del arte en la aplicación de tecnologías de datos e inteligencia artificial al sector fotovoltaico, identificar las mejores prácticas documentadas a nivel internacional y fundamentar las decisiones de diseño del estudio. Por otro lado, el análisis cuantitativo de series de tiempo reales provenientes de la Central de Puerto Peñasco permite contrastar los referentes teóricos con la realidad de una instalación específica, en un contexto climático, operativo e institucional concreto. Esta combinación es particularmente apropiada para el tipo de pregunta que motiva esta tesis: no se trata únicamente de demostrar que las técnicas de ML funcionan en abstracto, algo que la literatura ha establecido ampliamente, sino de evaluar qué nivel de precisión y qué valor operativo son alcanzables con los datos disponibles en una central real en México, bajo las condiciones de cobertura temporal, calidad de los registros y diversidad de variables que caracterizan a ese entorno específico.

El objetivo central de este trabajo es proponer y sustentar, a partir de evidencia empírica y revisión del estado del arte, una estrategia de adopción de tecnologías basadas en datos e inteligencia artificial orientada a fortalecer la operación y la toma de decisiones en centrales fotovoltaicas en México, tomando como caso de estudio la Central de Puerto Peñasco. Para alcanzar ese propósito, la investigación se estructura en torno a tres componentes metodológicos articulados: en primer lugar, la definición de los requerimientos técnicos mínimos en materia de captura, integración y gestión de información operativa que son condición necesaria para implementar soluciones analíticas de este tipo; en segundo lugar, el desarrollo y validación de un modelo de aprendizaje automático entrenado con registros reales de la central, que sirva como demostración de viabilidad técnica en condiciones operativas reales; y en tercer lugar, el establecimiento de lineamientos de implementación que permitan identificar brechas existentes, evaluar condiciones de escalabilidad y definir los pasos requeridos para extender estas soluciones a otras instalaciones del parque

fotovoltaico nacional. Como extensión complementaria, se discuten las aplicaciones potenciales de Inteligencia Artificial Generativa y modelos de lenguaje como herramientas de soporte a la interpretación y la comunicación de resultados, planteadas como una línea de desarrollo futuro con fundamento en la literatura científica más reciente.

Los tres capítulos que conforman esta tesis abordan, de manera progresiva y articulada, los distintos componentes del objetivo central. Cada uno cumple una función específica dentro de la estructura argumentativa del trabajo y sirve como base para el siguiente.

En el primer capítulo se presenta una revisión del panorama tecnológico actual en la operación y gestión de centrales fotovoltaicas. Se examina la evolución de herramientas como los sistemas de adquisición de datos y control supervisorio (SCADA), las redes eléctricas inteligentes (*smart grids*), los modelos de aprendizaje automático aplicados a la generación solar, el Internet de las Cosas (IoT) y los protocolos de comunicación industrial. Esta revisión analiza la experiencia de países que han avanzado de forma significativa en la modernización de sus sistemas eléctricos; incluyendo Alemania, España, Canadá, Estados Unidos, Japón y Uruguay, entre otros, identificando en cada caso las condiciones técnicas, regulatorias y financieras que han favorecido la implementación exitosa de estas tecnologías. El capítulo concluye con un análisis del estado actual en México y establece el puente hacia el trabajo empírico del capítulo siguiente, describiendo el conjunto de datos de Puerto Peñasco y la metodología de modelado adoptada.

El segundo capítulo está dedicado al trabajo analítico con los registros reales de la central. A partir de series temporales de frecuencia horaria correspondientes a un periodo continuo de casi dos años, desde diciembre de 2023 hasta agosto de 2025, se realizan análisis exploratorios orientados a caracterizar los patrones de generación, la estacionalidad, la variabilidad intradiaria e interestacional y las correlaciones entre las variables meteorológicas disponibles y la potencia de salida. Sobre esa base se construyen, entrenan y comparan cuatro arquitecturas de modelos predictivos cuyo desempeño se cuantifica mediante métricas estándar. El capítulo incorpora también una reflexión sobre el rol que los LLMs podrían desempeñar como herramienta de síntesis e interpretación de estos resultados en un entorno operativo real.

En el tercer capítulo se presentan los hallazgos integrados de la investigación junto con las conclusiones generales del trabajo. Se discuten los alcances y las limitaciones del enfoque adoptado, se identifican las principales áreas de oportunidad para trabajos futuros y se delinean recomendaciones concretas para el escalamiento de las soluciones desarrolladas, prestando especial atención a los requisitos de infraestructura de monitoreo, arquitectura de datos, interoperabilidad y gobernanza de la información que son condición necesaria para una adopción progresiva y sostenible en el contexto del sistema eléctrico mexicano.

En conjunto, estos tres capítulos configuran un recorrido metodológicamente coherente que parte del análisis conceptual del estado del arte, avanza hacia la aplicación práctica con base en evidencia real y culmina en la generación de conocimiento aplicable tanto en el ámbito académico como en el operativo. La propuesta no se limita a documentar lo que es posible en el plano teórico,

sino que apunta a identificar qué es viable en el contexto específico de México, qué brechas deben cerrarse para escalar estas soluciones y qué criterios deben orientar las decisiones de implementación de quienes operan y gestionan centrales fotovoltaicas en el país.

BIBLIOGRAPHY

[1] P. Mulás del Pozo, *El aprovechamiento de la energía solar*, Centro de Ecodesarrollo, A.C., Ciudad de México, México, Rep. Téc., 1976.

[2] Secretaría de Energía (SENER), "Inauguración del primer campo fotovoltaico de la CFE 'Santa Rosalía'," SENER, Ciudad de México, México, Oct. 2012. [En línea]. Disponible: https://www.gob.mx/sener

[3] Secretaría de Energía (SENER), *Prospectiva de Energías Renovables 2016–2030*, Secretaría de Energía, Ciudad de México, México, 2016. [En línea]. Disponible: https://www.gob.mx/cms/uploads/attachment/file/177622/Prospectiva_de_Energ_as_Renovables_2016-2030.pdf

[4] Agencia Internacional de Energías Renovables (IRENA), *Renewable Power Generation Costs in 2022*, IRENA, Abu Dhabi, UAE, 2023.

[5] Secretaría de Energía (SENER), "Inauguración de Central Fotovoltaica y Firma del Decreto para Nacionalización del Litio," SENER, Ciudad de México, México, Feb. 2023. [En línea]. Disponible: https://www.gob.mx/sener/articulos/inauguracion-de-central-fotovoltaica-y-firma-del-decreto-para-nacionalizacion-del-litio-fortalecen-politica-energetica-de-la-4t

[6] International Energy Agency (IEA), *Digitalization and Energy*, IEA, París, Francia, 2017.

[7] T. Kim, H. Li, R. Bhatt, y B. Palmintier, *Generative AI for Power Grid Operations*, National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep. NREL/TP-5D00-91176, 2024.

CHAPTER I

Revisión de Tecnologías Emergentes para la Operación y Gestión de Centrales Fotovoltaicas: Tendencias y Retos Actuales

La expansión sostenida de la generación eléctrica a partir de fuentes renovables ha dejado de ser una proyección optimista de política energética para convertirse en una realidad verificable en los sistemas eléctricos de un número creciente de países. En México, esta tendencia se hizo visible a escala de servicio público desde inicios de la década pasada, siendo la Central Solar "Santa Rosalía" en Baja California Sur, inaugurada en octubre de 2012 con una capacidad de 1 MW, el primer referente concreto de esa transición [1]. Sin embargo, lo que comenzó como una demostración técnica modesta se ha transformado en poco más de una década en un proceso de despliegue acelerado, impulsado por la caída sostenida de los costos de los módulos fotovoltaicos, los incentivos regulatorios y la creciente urgencia de diversificar la matriz energética nacional.

No obstante, integrar generación fotovoltaica en un sistema eléctrico de gran escala es un proceso que implica desafíos técnicos de una complejidad que va mucho más allá de instalar paneles y conectarlos a la red. A diferencia de las centrales de generación base, como las térmicas o las hidroeléctricas con embalse, cuya producción puede ajustarse activamente en respuesta a las variaciones de la demanda, las centrales fotovoltaicas operan en función de las condiciones de irradiancia disponibles en cada momento, las cuales dependen de factores externos no controlables: el ángulo de incidencia solar, la presencia de nubes, la temperatura ambiente, el polvo acumulado en los módulos o incluso eventos meteorológicos extremos. Esta naturaleza variable introduce fluctuaciones en la inyección de potencia a la red que pueden manifestarse en escalas temporales que van desde segundos, ante el paso de una nube, hasta meses, por la variación estacional de la altura solar. La red eléctrica, diseñada en su mayor parte bajo el paradigma de la generación

controlable, debe ser capaz de absorber esas fluctuaciones sin comprometer la estabilidad de la frecuencia, la calidad del voltaje ni la continuidad del suministro para los usuarios finales.

Ante este desafío estructural, la International Energy Agency (IEA) sostiene que la modernización de las redes eléctricas mediante digitalización, sensórica avanzada y herramientas de software constituye el mecanismo central para sostener la confiabilidad del sistema a medida que aumenta la participación de fuentes renovables variables en el mix de generación[1] [2]. Esta no es una postura académica abstracta: es una conclusión que emerge del análisis de los sistemas eléctricos que han alcanzado las mayores cuotas de penetración renovable, como los de Alemania, España o Dinamarca, y que muestra, de forma consistente, que la capacidad de gestión inteligente de la red es tan importante como la capacidad instalada misma. Más aún, la IEA proyecta que la inversión global en infraestructura de redes inteligentes deberá más que duplicarse antes de 2030 para mantener una trayectoria de descarbonización compatible con las metas de emisiones netas cero, advirtiendo que el déficit de inversión es especialmente pronunciado en los mercados emergentes y las economías en desarrollo, entre los que se cuenta México [3].

Bajo este panorama, las Redes Eléctricas Inteligentes (REI), denominadas *smart grids* en la literatura anglosajona, emergen como la arquitectura operativa de referencia para gestionar la transición energética en curso. La IEA las define como redes que emplean tecnologías digitales, sensores y programas informáticos para equilibrar la oferta y la demanda de electricidad en tiempo real, minimizando costos y preservando la estabilidad del sistema [3]. Sin embargo, esta definición, aunque precisa en su concisión, no captura la riqueza funcional de lo que una red inteligente implica en la práctica. En términos operativos, una REI no es simplemente una red convencional con medidores digitales: es una infraestructura que integra múltiples capas de tecnología sensórica distribuida, comunicaciones bidireccionales de alta velocidad, plataformas de procesamiento en tiempo real, algoritmos de optimización y herramientas de soporte a la decisión para coordinar de forma dinámica actores heterogéneos que incluyen generadores de distinta naturaleza, operadores de red, sistemas de almacenamiento, vehículos eléctricos y consumidores que pueden ser también productores de energía. La madurez que exhiben hoy muchas de las tecnologías que componen esta arquitectura hace que su adopción sea, por primera vez en la historia del sector, técnica y económicamente viable en contextos tan diversos como el de México.

1.1 *Smart Grids* en México y el Mundo

La modernización de las redes eléctricas no ha seguido un camino único ni universal. Cada sistema eléctrico nacional presenta una combinación específica de características: perfil de generación, antigüedad de la infraestructura, estructura regulatoria, capacidad institucional y

[1] El *mix* de generación designa el conjunto de tecnologías y fuentes de energía que contribuyen a la producción eléctrica total de un sistema en un periodo determinado. Su composición define tanto el perfil de emisiones del sistema como las características de variabilidad y controlabilidad de la oferta eléctrica; a mayor participación de fuentes variables como la solar o la eólica, mayor es la exigencia sobre la capacidad de regulación y gestión de la red.

grado de penetración renovable; que ha condicionado el ritmo, el enfoque y las prioridades de su proceso de transformación. Sin embargo, las experiencias documentadas en distintas regiones del mundo ofrecen lecciones transferibles sobre qué condiciones favorecen el éxito, qué obstáculos son recurrentes y qué enfoques han demostrado mayor eficacia para cerrar la brecha entre el marco normativo y la implementación real. En las subsecciones siguientes se describen iniciativas representativas de ocho países o regiones, identificando en cada caso los elementos más relevantes para el contexto mexicano.

1.1.1 Alemania

Alemania representa quizás el caso más documentado y estudiado de transformación acelerada del sistema eléctrico hacia una matriz dominada por fuentes renovables. La *Energiewende*, término que puede traducirse como "transición energética", no es simplemente un programa de política pública; es un proceso de transformación sistémica que abarca la generación, el transporte, la distribución y el consumo de energía, y que ha requerido modificaciones profundas tanto en la infraestructura física como en los marcos regulatorios, los modelos de negocio y la cultura operativa del sector. En este contexto, la digitalización del sistema energético no es un objetivo en sí mismo, sino una condición estructural para hacer posible la operación eficiente de una red que, en ciertos momentos, obtiene la mayor parte de su electricidad de fuentes variables como la eólica y la solar.

El Ministerio Federal de Economía y Acción Climática (BMWK) impulsó en 2023 lineamientos específicos para acelerar el despliegue de medidores inteligentes y habilitar la gestión activa de la demanda en tiempo real [4]. El diagnóstico ministerial parte de una constatación práctica: sin una infraestructura digital que conecte con precisión la producción y el consumo en cada momento, la integración de grandes volúmenes de energía renovable variable genera cuellos de botella, desperdicio de generación y costos de red que erosionan la viabilidad económica del proceso. En este sentido, los medidores inteligentes no son simplemente dispositivos de medición avanzada: son los nodos terminales de un sistema de información que permite al operador de red conocer en tiempo real el estado de la demanda en cada punto del sistema y gestionar activamente la flexibilidad disponible.

En paralelo, el programa SINTEG (*Smart Energy Showcases – Digital Agenda for the Energy Transition*), puesto en marcha desde 2016, estableció cinco regiones modelo a gran escala para desarrollar, probar y perfeccionar soluciones de operación de red y de mercado energético capaces de garantizar un suministro seguro y eficiente con alta penetración de fuentes variables [5]. El diseño metodológico de SINTEG es particularmente relevante: en lugar de desplegar tecnologías directamente a escala nacional, con los riesgos y costos que ello implicaría, el programa financia la implementación real de soluciones en contextos representativos, recopila evidencia sistemática sobre su desempeño técnico y económico, e identifica los ajustes regulatorios necesarios antes de escalar. Este modelo de pilotos estructurados y evaluados rigurosamente es uno de los aportes

metodológicos más valiosos de la experiencia alemana para otros países que están comenzando a transitar este camino.

1.1.2 Canadá

El caso canadiense ilustra con claridad cómo una política pública bien diseñada puede articular la inversión pública y privada en torno a objetivos concretos y medibles en el tiempo, generando un ciclo de aprendizaje institucional que trasciende los proyectos individuales. A través del *Smart Grid Program* coordinado por Natural Resources Canada (NRCan), el gobierno federal estructuró un esquema de financiamiento para proyectos de demostración y despliegue de redes inteligentes con un enfoque explícito en tres objetivos complementarios: la integración de fuentes renovables al sistema eléctrico, la reducción de emisiones de gases de efecto invernadero, y el uso más eficiente de los activos eléctricos ya existentes, evitando así la construcción de infraestructura nueva que podría no resultar necesaria si la demanda se gestiona de forma más inteligente [6].

Durante el periodo 2018–2023, el programa comprometió cerca de 100 millones de dólares canadienses en proyectos dirigidos en su mayoría por empresas de servicios públicos, lo que aseguró que las soluciones desarrolladas respondieran a necesidades operativas reales y no exclusivamente a visiones de laboratorio [7]. Un aspecto diferenciador del diseño del programa es su estructura de rendición de cuentas de largo plazo: los beneficiarios están obligados a reportar el impacto en la red de sus proyectos durante los cinco años posteriores a la recepción de la inversión, y NRCan utiliza esa información acumulada para orientar el diseño de futuros programas y políticas. Esta arquitectura de seguimiento genera un activo de conocimiento sectorial que va mucho más allá del proyecto individual: construye una base de evidencia sobre qué tecnologías funcionan en condiciones reales, en qué contextos y con qué resultados, permitiendo que las decisiones de política pública futura se apoyen en datos y no solamente en proyecciones teóricas.

1.1.3 Chile

Chile ofrece un caso de especial relevancia por la combinación de dos dinámicas simultáneas: un crecimiento muy acelerado de la penetración de energías renovables variables, en particular solar en el norte del país, y un esfuerzo paralelo de modernización de los instrumentos regulatorios y de gestión del sistema para acompañar esa transición sin comprometer la seguridad de suministro. La Política Energética 2050, publicada por el Ministerio de Energía, reconoció desde su formulación el rol creciente del consumidor activo y planteó explícitamente que las herramientas de gestión de demanda y los esquemas de operación asociados a las redes inteligentes serían determinantes para el desempeño del sector en las décadas siguientes [8]. Esta anticipación estratégica fue relevante porque estableció un marco de largo plazo que orientó las inversiones y las reformas regulatorias subsecuentes.

Desde el plano técnico-normativo, la Comisión Nacional de Energía (CNE) realizó estudios especí-

ficos sobre los mecanismos de tarificación necesarios para viabilizar el despliegue de medición inteligente y la automatización de la red de distribución, concluyendo que las señales económicas correctas son tan importantes como la tecnología misma: sin incentivos tarifarios adecuados, ni los usuarios ni las empresas distribuidoras tienen razones suficientes para adoptar esquemas de gestión activa de la demanda [9]. En un plano más técnico, el Centro de Energía de la Universidad de Chile elaboró un diagnóstico integral que propone un conjunto articulado de medidas habilitantes para la digitalización del sector: pilotos de redes inteligentes en contextos urbanos y rurales representativos, masificación de medidores de intervalo de 15 minutos, estándares de interoperabilidad entre equipos de distintos fabricantes, y fortalecimiento de las capacidades de ciberseguridad ante una superficie de ataque creciente [10]. En el plano operativo, el Coordinador Eléctrico Nacional ha identificado requerimientos concretos de capacidad de respuesta y coordinación que se vuelven ineludibles a medida que la penetración renovable avanza [11].

1.1.4 España

España ofrece quizás el ejemplo más elocuente y contemporáneo de lo que puede lograrse cuando la planificación energética de largo plazo, la modernización progresiva de la infraestructura de red y el desarrollo de una industria nacional competitiva convergen en un proceso coherente y sostenido. El Plan Nacional Integrado de Energía y Clima (PNIEC), en su actualización 2023–2030, fija el objetivo de alcanzar el 81% de generación renovable en el mix eléctrico para ese año [12], una meta que desde el punto de vista operativo sería absolutamente inviable sin una red capaz de gestionar con precisión una producción que durante muchas horas del año provendría mayoritariamente de fuentes variables.

Los resultados ya son perceptibles y cuantificables. En mayo de 2024, la energía solar fotovoltaica se convirtió por primera vez en la tecnología líder del mix mensual español, aportando el 23.8% de la producción total y superando a todas las demás fuentes individuales de generación [13]. Durante ese mes, las energías renovables en conjunto cubrieron el 63% de la demanda eléctrica, y las tecnologías libres de emisiones, incluyendo la nuclear, alcanzaron el 81.4% del total generado. Estos valores no son el resultado de condiciones meteorológicas excepcionales ni de circunstancias coyunturales favorables: son el producto acumulado de más de una década de inversión en digitalización de la red de distribución, automatización de subestaciones y capacidades de pronóstico y despacho en tiempo real. En ese contexto, proyectos como i-DE de Iberdrola, con más de 270,000 km de red de distribución, 1,100 subestaciones, 98,000 estaciones de transformación y el 100% de sus medidores convertidos a inteligentes para 2025, ilustran la escala y la profundidad del proceso de transformación que está en curso en el sector [14].

1.1.5 Estados Unidos

La complejidad del sistema eléctrico estadounidense no tiene comparación entre los sistemas nacionales del mundo. Con más de 9,200 centrales de generación que suman más de 1 millón de

MW de capacidad instalada, y con una red de transmisión que se extiende a lo largo de cerca de 600,000 millas de líneas [15], el sistema eléctrico de Estados Unidos es en realidad un conjunto de redes interconectadas pero parcialmente independientes, las interconexiones del Este, del Oeste y de Texas, gobernadas por una combinación de autoridades federales, estatales, municipales y privadas que hace extraordinariamente compleja la coordinación de cualquier esfuerzo de transformación a escala nacional. Esta fragmentación jurisdiccional ha sido históricamente uno de los principales obstáculos para modernizar la red a un ritmo comparable al de sistemas más centralizados.

El hito legislativo que sentó las bases del marco federal actual fue el Título XIII de la *Energy Independence and Security Act* (EISA) de 2007, aprobada con amplio apoyo bipartidista. Este título otorgó respaldo legal explícito a las iniciativas de red inteligente del Departamento de Energía (DOE), estableció el rol del gobierno federal como coordinador de los esfuerzos nacionales de modernización, y creó el *Federal Smart Grid Task Force*; un órgano interagencial encargado de identificar las barreras regulatorias y técnicas para la adopción de redes inteligentes y proponer mecanismos para superarlas [16]. Los beneficios perseguidos son concretos y operacionalmente significativos: menor frecuencia y duración de las interrupciones del suministro, que cuestan a la economía estadounidense decenas de miles de millones de dólares anuales, mayor resiliencia ante fenómenos meteorológicos extremos de creciente frecuencia e intensidad, reducción de pérdidas técnicas en la red, y mayor integración de generación distribuida y almacenamiento.

1.1.6 Japón

La estabilidad del suministro eléctrico ha sido históricamente un valor cardinal del sistema energético japonés, derivado en parte de la experiencia traumática de los grandes apagones provocados por terremotos y tsunamis, y en parte de las exigencias de una economía industrial de alta precisión que no tolera interrupciones del servicio. Esta prioridad ha modelado profundamente el enfoque japonés hacia las redes inteligentes: no se trata únicamente de integrar más renovables, sino de construir un sistema resiliente ante perturbaciones de cualquier naturaleza, capaz de recuperarse autónomamente y de redistribuir recursos en tiempo real.

En 2022, el gobierno japonés adoptó el *Smart Energy Grid Architecture Model* (SGAM) de la Comisión Electrotécnica Internacional (IEC) como marco de referencia para la digitalización del sistema, ampliándolo para incorporar las interacciones entre los sistemas de electricidad, gas y calor en una visión energética integrada [17]. Para 2023 se publicó un plan maestro de interconexión a escala nacional orientado a facilitar la transmisión eficiente de energía entre zonas con excedentes y zonas con déficit, reduciendo la necesidad de *curtailment* de renovables. Y en 2024 entró en plena operación la gestión de área amplia de la oferta y la demanda, incluyendo mecanismos de alerta temprana ante situaciones de escasez y protocolos coordinados de respuesta ante crisis de suministro.

En términos de generación distribuida, la instalación de sistemas fotovoltaicos en techos residenciales saltó de menos de 5 GW a más de 20 GW en la última década, representando ya más del 9% de la producción eléctrica total del país [17]. Este crecimiento, acelerado por la decisión del gobierno metropolitano de Tokio de exigir la instalación de paneles solares en todas las viviendas nuevas a partir de 2025 [18], plantea retos de gestión de la red de distribución que sólo pueden abordarse con herramientas de monitoreo, predicción y control en tiempo real. El establecimiento de cuatro proyectos comunitarios demostrativos a gran escala en Kyoto, Yokohama, Kitakyushu y Toyota, con más de 5,000 hogares participantes en cada caso, refleja el enfoque japonés de validar en comunidades reales y en condiciones controladas antes de proceder a despliegues generalizados [17].

1.1.7 Uruguay

Uruguay constituye un referente internacional que merece atención particular, no por el tamaño de su economía o la escala de su sistema eléctrico, sino por la profundidad y la velocidad de la transformación energética que ha protagonizado. En menos de quince años, el país pasó de una matriz eléctrica dependiente en gran medida de la generación termoeléctrica de respaldo a una con participación renovable mayoritaria y creciente, incorporando generación eólica, solar y de biomasa de forma ordenada y sin comprometer la seguridad del suministro [19]. Esta transformación respondió a una estrategia energética deliberada, articulada por UTE, la empresa estatal de transmisión y distribución, que combinó la licitación de nueva capacidad renovable mediante subastas competitivas con la modernización simultánea de la infraestructura de red.

La medición inteligente fue el eje central de esta modernización en el segmento de distribución. Los dispositivos desplegados por UTE permiten el registro del consumo en intervalos de 15 minutos, la ejecución remota de reconexiones, cortes y cambios de tarifa, y la detección más rápida de interrupciones mediante la identificación automática de equipos que dejan de reportar [20]. Para 2023, la empresa había instalado 1 millón de estos dispositivos, equivalente a la práctica totalidad de su base de clientes residenciales, y una licitación en curso proyectaba completar la cobertura hacia finales de 2024 [21][22]. La escala de este despliegue, en un país con recursos económicos y técnicos similares a los de muchas regiones de México, hace de Uruguay un modelo de referencia especialmente pertinente para evaluar la viabilidad de programas comparables en el contexto latinoamericano.

1.1.8 México

En México, la modernización de las redes eléctricas cuenta con un soporte legal e institucional formalmente establecido. La Ley de Transición Energética, aprobada en 2015, creó el Programa de Redes Eléctricas Inteligentes con el mandato explícito de apoyar la modernización de la Red Nacional de Transmisión y las Redes Generales de Distribución, articulando líneas de acción en áreas como medición inteligente, interoperabilidad de sistemas, gestión activa de la

demanda y almacenamiento de energía [23]. A nivel programático, la Secretaría de Energía publicó el Programa de Redes Eléctricas Inteligentes para articular estas acciones y contribuir a la eficiencia operativa del SEN [24]. Complementariamente, el National Renewable Energy Laboratory (NREL) elaboró un reporte técnico sobre el estado de implementación de *smart grids* en México que identifica los avances existentes y las brechas que persisten en respuesta a los mandatos legales vigentes [25].

Sin embargo, la distancia entre el marco normativo y la implementación efectiva sigue siendo significativa y merece una lectura crítica. León-Trigo, documentan que los esfuerzos tecnológicos más avanzados en el país han estado concentrados en el segmento de microrredes residenciales mediante inversores inteligentes de generación distribuida, sin que existan todavía estrategias equivalentes articuladas y ejecutadas para instalaciones de generación a gran escala [26]. Esta disparidad pone de manifiesto una tensión que es habitual en procesos de transformación tecnológica: la existencia de un marco legal ambicioso no garantiza por sí sola la disponibilidad de los recursos financieros, la capacidad técnica instalada, los incentivos económicos correctos y la voluntad institucional de ejecución necesarios para cerrar la brecha entre lo que la ley establece y lo que el sistema implementa en la práctica. El reto mexicano no radica en la ausencia de instrumentos normativos, sino en la priorización presupuestal, el desarrollo de capacidades técnicas nacionales y la construcción de las infraestructuras habilitadoras, medición avanzada, comunicaciones de baja latencia, interoperabilidad entre sistemas y analítica de datos, que son condición necesaria para operar el SEN con la eficiencia que exige una penetración renovable creciente.

1.2 Tecnologías en el sector eléctrico actual

La irrupción de tecnologías digitales en la operación del sector eléctrico ha transformado de manera fundamental la naturaleza de la gestión de las centrales de generación. Lo que en décadas anteriores dependía de inspecciones periódicas programadas, rutinas de mantenimiento basadas en el tiempo transcurrido y decisiones apoyadas en la experiencia acumulada del operador, hoy se apoya en flujos continuos de señales técnicas, ambientales y operativas que permiten conocer el estado de cada componente del sistema prácticamente en tiempo real y con una resolución temporal que puede llegar a segundos. Esta capacidad de observación granular tiene consecuencias directas sobre la eficiencia operativa, la confiabilidad del suministro y los costos de mantenimiento: cuanto más temprano se detecta una anomalía, ya sea una degradación incipiente en un módulo fotovoltaico, un desequilibrio de corriente en un string, o una variación en el coeficiente de temperatura de un inversor, menor es el costo de su corrección y menor el impacto sobre la generación.

En el caso específico de las instalaciones fotovoltaicas, esta capacidad de observación continua adquiere especial relevancia porque la producción es inherentemente variable y está condicionada

por factores externos no controlables. En este contexto, el valor de la información no está únicamente en describir lo que está ocurriendo en el sistema en un momento dado, sino en anticipar lo que ocurrirá en los próximos minutos, horas o días, y en detectar desviaciones respecto al comportamiento esperado que puedan indicar problemas incipientes. Esta dualidad, monitoreo en tiempo real más capacidad predictiva, define el espacio de aplicación de las tecnologías que se describen a continuación, y es consistente con la tendencia documentada por la IEA según la cual la digitalización constituye el principal mecanismo habilitador para sostener redes más flexibles y confiables bajo un escenario de alta penetración renovable [3].

1.2.1 Recolección, procesamiento y análisis de datos

En una central fotovoltaica moderna de escala industrial, la generación de información operativa es un proceso continuo, distribuido geográficamente y de alta densidad. El campo de paneles puede extenderse a lo largo de cientos de hectáreas, y en cada punto de ese campo existen variables relevantes para la gestión: la irradiancia incidente en plano inclinado y en plano horizontal, la temperatura de la célula fotovoltaica, la temperatura ambiente, la velocidad y dirección del viento, la corriente y el voltaje de salida de cada string y de cada inversor, los parámetros de red en el punto de conexión, frecuencia, tensión, potencia activa y reactiva, y el estado de carga de los sistemas de almacenamiento si la instalación los incorpora. Sensores de irradiancia, piranómetros, termómetros de módulo y ambiente, analizadores de calidad de energía y estaciones meteorológicas automáticas producen en conjunto cientos de señales que pueden actualizarse con frecuencias que van desde el segundo hasta el minuto, dependiendo del nivel de resolución temporal requerido por la aplicación. Gestionar esta densidad de señales de forma que sea útil para la operación y no simplemente un volumen inmanejable de datos brutos, requiere una arquitectura de adquisición y procesamiento estructurada en múltiples capas funcionales, cuyo núcleo son los sistemas SCADA (*Supervisory Control and Data Acquisition*) [27].

Un sistema SCADA de generación fotovoltaica opera típicamente en tres capas funcionales claramente diferenciadas. La primera es el nivel de campo, donde las unidades de telemetría remota (RTUs, *Remote Terminal Units*) y los controladores lógicos programables (PLCs, *Programmable Logic Controllers*) cumplen la función de capturar señales analógicas y digitales directamente desde los equipos: inversores trifásicos, transformadores de potencia, sistemas de almacenamiento en baterías, interruptores automáticos y transmitirlas hacia los niveles superiores mediante protocolos industriales estandarizados. Entre los más utilizados en el sector eléctrico se encuentran Modbus RTU/TCP, por su simplicidad y universalidad en dispositivos de campo, DNP3 diseñado específicamente para la automatización de la infraestructura eléctrica, con capacidades nativas de gestión de tiempo y transmisión de eventos, e IEC 61850, el estándar de referencia para subestaciones de potencia modernas, que garantiza la interoperabilidad semántica entre equipos de distintos fabricantes y soporta comunicaciones en tiempo real con latencias inferiores a los 4 milisegundos para funciones de protección críticas [28]. La segunda

capa es la red de comunicación, que puede materializarse mediante fibra óptica para los tramos de alta capacidad dentro de la central, radioenlaces de microondas para zonas donde el tendido de fibra no es viable, o conexiones celulares de baja latencia, incluyendo tecnologías 4G LTE y, en instalaciones de nueva generación, 5G para la transmisión hacia centros de operación remotos. La tercera capa es la de aplicación, donde el software SCADA almacena los registros en bases de datos de series temporales optimizadas, los visualiza en tiempo real mediante interfaces gráficas para los operadores de turno y los expone a través de APIs estandarizadas hacia otros sistemas de análisis, reportes y gestión de activos [27].

La integración del *Internet of Things* (IoT) ha complementado esta arquitectura tradicional al incorporar una nueva generación de dispositivos de bajo costo, bajo consumo energético y conectividad directa a plataformas de procesamiento en la nube. A diferencia de las RTUs convencionales, que son equipos robustos, costosos y relativamente rígidos en su configuración, los sensores IoT pueden instalarse en grandes cantidades, configurarse y actualizarse de forma remota mediante software, y comunicarse con plataformas centrales utilizando protocolos ligeros especialmente diseñados para entornos de conectividad variable. El protocolo MQTT (*Message Queuing Telemetry Transport*), basado en un modelo de publicación y suscripción, se ha consolidado como el estándar dominante en este tipo de despliegues por su eficiencia en redes de baja velocidad y ancho de banda limitado, su capacidad de mantener la entrega garantizada de mensajes incluso ante interrupciones transitorias de la conexión, y su relativa sencillez de implementación [28]. En paralelo, plataformas de *edge computing*, equipos de procesamiento instalados directamente en la instalación o en sus proximidades, permiten ejecutar algoritmos de análisis sobre los datos localmente, reduciendo la latencia en las decisiones críticas, disminuyendo el volumen de transmisión hacia los sistemas centrales y garantizando la continuidad de ciertas funciones de monitoreo incluso cuando la conexión a la red pública se interrumpe [29].

El valor operativo de toda esta infraestructura se materializa, en última instancia, en la capacidad de extraer conocimiento accionable a partir de los registros acumulados. Las series temporales generadas por una central fotovoltaica de escala industrial contienen información de una riqueza extraordinaria: patrones de generación diaria y estacional que reflejan el comportamiento del recurso solar en el emplazamiento, correlaciones entre variables ambientales y producción que revelan las relaciones físicas subyacentes, y señales de degradación o fallo incipiente que pueden ser imperceptibles para el operador humano durante semanas o meses pero que son detectables mediante técnicas estadísticas y algoritmos de aprendizaje automático. Modelos como las redes neuronales recurrentes con celdas de memoria a largo plazo (LSTM), los ensambles de árboles de decisión por gradiente (*Gradient Boosted Trees*) y las máquinas de soporte vectorial han demostrado en múltiples estudios su capacidad para identificar estas señales débiles, anticipar caídas de rendimiento con días o semanas de anticipación y generar alertas tempranas que permiten programar intervenciones de mantenimiento en ventanas operativas de bajo costo [29].

1.2.2 Utilización de algoritmos y herramientas de cómputo

La riqueza de los registros operativos descritos en la sección anterior únicamente se convierte en valor operativo cuando existe la capacidad computacional y algorítmica de procesarlos de forma eficiente, extraer los patrones relevantes y traducir ese conocimiento en acciones concretas. En el ámbito de las centrales fotovoltaicas, los algoritmos aplicados pueden agruparse en tres grandes categorías según su propósito funcional: la optimización del rendimiento energético, el control en tiempo real del sistema y el diagnóstico de fallas con soporte al mantenimiento predictivo [30].

Los algoritmos de optimización del rendimiento energético tienen como objetivo central maximizar la energía eléctrica extraída del campo fotovoltaico bajo las condiciones variables de irradiancia y temperatura que caracterizan a cualquier emplazamiento solar. El algoritmo más fundamental dentro de esta categoría es el de seguimiento del punto de máxima potencia (MPPT, *Maximum Power Point Tracking*), implementado de forma nativa en cada inversor fotovoltaico para ajustar continuamente el punto de operación del campo, es decir, la relación entre voltaje y corriente a la que trabaja el generador, de modo que se extraiga en cada instante la máxima potencia disponible. La curva I-V de un campo fotovoltaico varía continuamente con la irradiancia y la temperatura de la célula, y el punto de máxima potencia se desplaza a lo largo de ella de forma no lineal; seguirlo con precisión bajo condiciones de variación rápida de nubosidad requiere algoritmos de convergencia que sean al mismo tiempo suficientemente rápidos para responder a transitorios en segundos y suficientemente estables para no oscilar en condiciones de irradiancia uniforme. Las versiones avanzadas de MPPT, incluyendo las basadas en lógica difusa, las que emplean redes neuronales para predecir la posición del punto óptimo antes de que las condiciones cambien, o las que utilizan métodos de perturbación y observación con paso adaptativo que se ajusta dinámicamente en función de la velocidad de variación de la irradiancia, permiten reducir las pérdidas por seguimiento incompleto de forma significativa, especialmente en días de alta variabilidad nubosa [30].

La segunda categoría de algoritmos, orientada al diagnóstico de fallas y al soporte al mantenimiento predictivo, constituye probablemente el campo de mayor crecimiento y mayor impacto económico potencial en la operación de centrales fotovoltaicas durante la última década. El diagnóstico automático de fallas es un problema que combina características de clasificación supervisada, cuando se dispone de datos etiquetados sobre tipos de fallas anteriores y de detección de anomalías no supervisada, cuando se busca identificar comportamientos que se desvían del patrón esperado sin que necesariamente correspondan a una categoría de falla conocida. Algoritmos de clasificación como Random Forest y XGBoost han demostrado alta precisión en la detección temprana de un espectro amplio de anomalías: degradación acelerada de módulos por efecto de hotspots, cortocircuitos internos de string, desconexión parcial de inversores, pérdidas por sombreado no identificadas previamente, y anomalías en los sistemas de almacenamiento en baterías, entre otros [31]. La clave de su eficacia radica en la capacidad de comparar el rendimiento observado, potencia real generada, con el rendimiento esperado bajo las condiciones

meteorológicas registradas simultáneamente, e identificar desviaciones estadísticamente significativas que señalen la presencia de una falla antes de que esta se manifieste como una pérdida perceptible de generación o, en casos más graves, como un evento de seguridad. Las arquitecturas LSTM son especialmente adecuadas para la detección de degradación a largo plazo porque su diseño les permite capturar dependencias temporales a múltiples escalas simultáneamente, aprendiendo que la producción de un punto determinado en el tiempo, bajo condiciones de irradiancia y temperatura comparables a las de semanas o meses anteriores, presenta una tendencia de deterioro constante que señala la necesidad de una intervención preventiva [31].

En cuanto al ecosistema de herramientas de desarrollo, Python se ha consolidado como el lenguaje de programación dominante en este dominio, impulsado por la madurez y la riqueza de sus bibliotecas especializadas. La biblioteca scikit-learn proporciona implementaciones eficientes y consistentes de una amplia gama de algoritmos de ML clásico, incluyendo Random Forest, máquinas de vectores de soporte, regresión regularizada y métodos de ensamble, junto con utilidades para validación cruzada, selección de características y evaluación de modelos. TensorFlow y PyTorch son las plataformas de referencia para el diseño, entrenamiento y despliegue de arquitecturas de aprendizaje profundo, incluyendo redes recurrentes (LSTM, GRU) y redes convolucionales. Las bibliotecas pandas y NumPy cubren la capa de manipulación, limpieza y transformación de datos tabulares y series temporales, mientras que Matplotlib, Seaborn y Plotly permiten la visualización interactiva de resultados con un nivel de detalle que facilita tanto el análisis exploratorio como la comunicación de hallazgos a audiencias técnicas y no técnicas [31]. El desempeño de los modelos se evalúa mediante métricas estándar que permiten cuantificar objetivamente la calidad de las predicciones: el error cuadrático medio (RMSE) penaliza de forma cuadrática los errores grandes, lo que resulta especialmente informativo en contextos operativos donde las sobreestimaciones pronunciadas tienen consecuencias directas sobre el balance de la red; el error absoluto medio (MAE) ofrece una medida más robusta ante valores extremos y permite una interpretación directa en unidades físicas de potencia; y el coeficiente de determinación (R^2) proporciona una referencia adimensional que facilita la comparación entre modelos entrenados sobre conjuntos de datos con diferentes escalas de variabilidad.

1.3 Nuevas tecnologías en el sector eléctrico

La convergencia entre las capacidades de adquisición y gestión de datos descritas en las secciones anteriores y el avance sostenido de los algoritmos de aprendizaje automático ha abierto una nueva etapa en la operación de centrales fotovoltaicas: la del análisis predictivo orientado a la toma de decisiones proactiva. Esta transición, de sistemas que describen lo que está ocurriendo a sistemas que anticipan lo que ocurrirá, representa un salto cualitativo en la filosofía de operación de este tipo de instalaciones. A diferencia de los enfoques reactivos, que actúan una vez que una anomalía o una desviación de rendimiento ya es visible y cuantificable, los modelos predictivos permiten

anticipar el comportamiento del sistema con horizontes temporales que van desde minutos hasta días, habilitando una gestión proactiva de la generación, el despacho, el mantenimiento y la interacción con el mercado eléctrico [32]. La pregunta ya no es únicamente "¿qué está pasando ahora mismo en la central?" sino también "¿qué nivel de generación podemos comprometer para la próxima hora, para mañana o para la próxima semana, y con qué nivel de confianza?".

1.3.1 Modelos predictivos de generación solar: estado del arte

La predicción de la potencia generada por un sistema fotovoltaico es, desde el punto de vista del aprendizaje automático, un problema de regresión sobre series temporales multivariadas, en el que la variable objetivo, la potencia de salida en un instante futuro, depende de un conjunto de predictores meteorológicos y operativos que evolucionan de forma continua e interactúan entre sí de manera no lineal. La temperatura ambiente, por ejemplo, no solo afecta directamente al rendimiento de las células fotovoltaicas a través del coeficiente de temperatura, que reduce la tensión de circuito abierto aproximadamente 0.4% por cada grado Celsius de aumento, sino que también está correlacionada con la irradiancia incidente, con la velocidad del viento que determina el efecto de enfriamiento de los módulos, y con la demanda eléctrica del sistema, lo que introduce dependencias cruzadas entre variables que los modelos lineales simples son incapaces de capturar con fidelidad.

La literatura especializada clasifica los modelos de pronóstico fotovoltaico según su horizonte temporal, y esta clasificación tiene implicaciones directas sobre qué arquitecturas de modelo son más adecuadas y qué valor operativo pueden generar. Los pronósticos de muy corto plazo (*very short term*), con horizontes que van de segundos a minutos, son relevantes principalmente para el control de inversores, el balanceo de frecuencia en tiempo real y la gestión de la variabilidad intrahoraria en los puntos de conexión a la red. Los pronósticos de corto plazo (*short-term*), que cubren horizontes de horas a un día, son los de mayor utilidad para el despacho programado, la participación en mercados de energía de día anterior y la coordinación con sistemas de almacenamiento. Los pronósticos de mediano plazo (*day ahead* y *week ahead*) sustentan la planeación de mantenimientos preventivos, que requieren programar ventanas de menor generación esperada para minimizar la pérdida de producción, la gestión de reservas de capacidad y las negociaciones comerciales de largo plazo [33].

Entre los algoritmos con mayor presencia en la literatura reciente sobre pronóstico fotovoltaico, los modelos de conjunto basados en árboles de decisión, en particular Random Forest (RF) y sus variantes potenciadas como XGBoost y LightGBM, han demostrado de forma consistente un desempeño superior en conjuntos de datos tabulares con variables meteorológicas como predictores [34]. Sus ventajas son múltiples y complementarias: son robustos ante la presencia de valores atípicos y datos faltantes; tienen la capacidad de capturar interacciones no lineales complejas entre predictores sin requerir una especificación explícita de esas interacciones; producen estimaciones de importancia de variable que permiten interpretar qué factores contribuyen más al

pronóstico en un emplazamiento dado; y no requieren la normalización de las variables predictoras, simplificando el preprocesamiento. Por su parte, las redes neuronales artificiales (ANN) y especialmente sus variantes recurrentes como las redes LSTM (*Long Short-Term Memory*) ofrecen ventajas distintas: son especialmente adecuadas cuando la dependencia temporal de la serie es un predictor relevante por derecho propio, es decir, cuando el comportamiento pasado de la potencia fotovoltaica aporta información genuinamente adicional sobre el comportamiento futuro que no está completamente capturada en las variables meteorológicas contemporáneas [35].

Un aspecto crítico y frecuentemente subestimado en la construcción de modelos predictivos es la selección formal de las variables predictoras a incluir. No todas las señales disponibles en un sistema SCADA contribuyen de forma equivalente a la precisión del pronóstico, y la inclusión indiscriminada de variables puede introducir ruido, aumentar la complejidad computacional del modelo e incrementar el riesgo de sobreajuste, especialmente en conjuntos de datos de tamaño moderado como los que suelen estar disponibles en centrales en etapas tempranas de operación. Estudios recientes han demostrado que variables como la temperatura ambiente, la irradiancia horizontal global (GHI), la cobertura de nubes y la hora del día concentran la mayor parte del poder predictivo en la mayoría de los emplazamientos, mientras que otras como la presión atmosférica, la velocidad del viento o la visibilidad pueden tener contribuciones marginales que varían significativamente según la geografía local y el microclima del sitio [34][35]. La selección formal de características mediante técnicas como la Eliminación Recursiva de Características (RFE, *Recursive Feature Elimination*), que evalúa sistemáticamente qué subconjunto de predictores maximiza el desempeño del modelo, permite construir modelos más parsimoniosos, más generalizables ante datos futuros y más interpretables para los operadores, al tiempo que reduce los requisitos computacionales en sistemas con recursos limitados [36].

La evaluación final del desempeño de los modelos se realiza sobre un conjunto de prueba, reservado, que no participa en ninguna fase del entrenamiento ni de la selección de hiperparámetros. Esta separación es metodológicamente indispensable para obtener una estimación honesta de la capacidad de generalización del modelo: evaluar sobre datos que el modelo ya ha procesado durante el entrenamiento produce métricas optimistas que no reflejan el comportamiento real ante datos futuros. Las métricas estándar utilizadas en este dominio ofrecen perspectivas complementarias sobre la calidad del pronóstico: mientras el RMSE penaliza cuadráticamente los errores grandes y es por tanto especialmente informativo en contextos operativos donde las sobreestimaciones pronunciadas de la generación pueden comprometer el balance de la red, el MAE proporciona una medida más robusta ante valores extremos y permite una interpretación directa en unidades físicas de potencia. El R^2, por su parte, ofrece una referencia adimensional que facilita la comparación entre modelos entrenados sobre conjuntos de datos con diferentes escalas de variabilidad [33].

1.3.2 Inteligencia Artificial Generativa como capa de soporte operativo

En paralelo con la maduración de los modelos predictivos descritos hasta ahora, cuyo output es esencialmente numérico: un valor estimado de potencia, una clasificación de estado, una probabilidad de falla, la irrupción de la Inteligencia Artificial Generativa (GenAI) y los modelos de lenguaje de gran escala (LLMs, *Large Language Models*) está configurando una dimensión cualitativamente nueva de posibilidades en el sector eléctrico. La diferencia fundamental entre ambos tipos de herramientas no es de grado sino de naturaleza: los modelos discriminativos producen predicciones o clasificaciones sobre variables estructuradas, mientras que los LLMs son capaces de procesar texto, tablas y datos numéricos como entrada y producir como salida lenguaje natural estructurado: respuestas a preguntas, resúmenes, reportes, explicaciones o recomendaciones formuladas en términos comprensibles para usuarios sin formación especializada en ciencia de datos o estadística [37].

Esta capacidad de mediación entre los resultados del análisis cuantitativo y la comprensión humana de esos resultados tiene implicaciones prácticas de primera importancia para el contexto operativo de una central fotovoltaica. El personal técnico-operativo de instalaciones como Puerto Peñasco posee un profundo conocimiento del proceso físico y de los criterios operativos, pero no necesariamente una formación específica en interpretación de salidas de modelos de ML. Un LLM configurado como interfaz conversacional sobre los outputs de estos modelos puede responder a estas preguntas en tiempo real, adaptando el nivel de detalle y el vocabulario a las necesidades del usuario y generando explicaciones contextualizadas que conectan los resultados estadísticos con las implicaciones operativas concretas [37][38].

El National Renewable Energy Laboratory (NREL) ha documentado un espectro amplio de aplicaciones concretas de GenAI en operaciones de red eléctrica que ilustra este potencial: la síntesis automática de alertas operativas a partir de múltiples flujos de datos de monitoreo, la generación de resúmenes ejecutivos sobre el desempeño de la central para períodos específicos, turno, día, semana, mes, la respuesta a consultas sobre el historial operativo de equipos específicos, y la asistencia en la redacción de reportes técnicos para organismos regulatorios [37]. Cada una de estas aplicaciones transforma datos estructurados, tablas, series temporales, métricas de rendimiento, en comunicación legible y accionable, reduciendo la carga cognitiva de los operadores y liberando tiempo para tareas de mayor valor añadido.

Es importante, sin embargo, delimitar con precisión lo que los LLMs pueden y lo que no pueden hacer en el estado actual de su desarrollo. Estos modelos no reemplazan a los algoritmos especializados en precisión predictiva: un LLM no puede, por sí solo, predecir con fiabilidad cuántos megawatts generará un campo fotovoltaico dentro de tres horas, ni detectar un cortocircuito incipiente en un string a partir de señales de corriente. Su valor no está en el pronóstico cuantitativo de alta precisión, sino en tres funciones complementarias: la comunicación de resultados generados por otros modelos, la documentación asistida de procesos y hallazgos, y la democratización del

acceso a la información técnica para usuarios con distintos perfiles de formación. Emplearlos como capa de interpretación y comunicación sobre los outputs de modelos especializados, y no como sustitutos de estos, es el enfoque que la literatura científica más reciente identifica como el más maduro y el más adecuado para el tipo de aplicación que se contempla en este trabajo [37][38].

1.3.3 El caso de Puerto Peñasco como banco de prueba

Los conceptos y herramientas revisados a lo largo de este capítulo no constituyen únicamente un marco teórico de referencia: encuentran su aplicación directa en el análisis empírico que se desarrolla en el Capítulo II, utilizando registros operativos reales de la Central Fotovoltaica de Puerto Peñasco como base de evidencia. Esta vinculación entre el estado del arte y la práctica es deliberada y constitutiva del enfoque metodológico mixto adoptado en esta investigación: el Capítulo I establece el marco conceptual y justifica las decisiones de diseño, mientras que el Capítulo II proporciona la evidencia empírica que permite evaluar la viabilidad y el valor de esas herramientas en condiciones reales de operación.

El conjunto de datos disponible para el análisis comprende 2,016 registros horarios correspondientes al periodo diciembre de 2023 – agosto de 2025, cubriendo de forma representativa las cuatro estaciones del año: invierno, primavera, verano y otoño en un emplazamiento ubicado a 31.32° de latitud norte y 113.28° de longitud oeste, a 18 metros sobre el nivel del mar, en la costa norte del estado de Sonora. Esta ubicación geográfica es significativa desde el punto de vista del recurso solar: Sonora es uno de los estados con mayor irradiación media anual de México, y Puerto Peñasco en particular combina una alta disponibilidad del recurso con condiciones de temperatura que durante los meses de verano pueden superar los 40°C introduciendo así el efecto negativo de la temperatura sobre el rendimiento de las células como una variable operativa de peso significativo, y condiciones de viento que pueden ser intensas y variables en determinadas épocas del año. Esta diversidad de condiciones a lo largo del año es precisamente lo que hace del conjunto de datos un banco de prueba valioso: un modelo entrenado sobre estos registros ha sido expuesto a la variabilidad real del emplazamiento, lo que proporciona una estimación más honesta de su capacidad de generalización que la que ofrecerían datos restringidos a una sola estación o a condiciones meteorológicas uniformes.

La variable objetivo del análisis es la potencia de salida del generador fotovoltaico, designada en el conjunto de datos como PolyPwr, cuya magnitud oscila entre 0, durante las horas nocturnas y en condiciones de irradiancia mínima y 811.94 MW, el máximo registrado bajo condiciones de alta irradiancia y temperatura moderada, con una media de 225.9 MW y una desviación estándar de 288.96 MW. Esta elevada desviación relativa respecto a la media del orden del 128% es una característica definitoria de las series de potencia fotovoltaica: refleja la alternancia entre las horas nocturnas, en las que la producción es nula, y las horas centrales del día, en las que la producción puede alcanzar su valor máximo, así como la variabilidad adicional

introducida por las estaciones del año y por la nubosidad episódica. Un modelo que pretenda ser útil operativamente debe ser capaz de reproducir esta variabilidad con fidelidad: tanto los períodos de generación máxima, críticos para el despacho y las obligaciones comerciales de la central, como las transiciones rápidas de producción, que son los momentos de mayor exigencia para la gestión de la red.

Las variables predictoras disponibles en el conjunto de datos son ocho: temperatura ambiente (AmbientTemp), con valores que oscilan entre 6°C y 42.1°C y una media de 23.38°C; humedad relativa (Humidity), entre 4% y 96% con media de 46.61%; velocidad del viento (Wind.Speed), entre 0.32 y 80.32 m/s con media de 15.78 m/s; visibilidad horizontal (Visibility), entre 0.20 y 90 km con media de 43.18 km; presión atmosférica (Pressure), entre 1002.4 y 1026.6 hPa con media de 1012.13 hPa; cobertura de nubes (Cloud.Ceiling), entre 0 y 100 unidades con media de 18.76; hora del día (Hour); y mes del año (Month). La riqueza de este conjunto de predictores permite explorar cuáles de ellos tienen mayor capacidad explicativa sobre la generación fotovoltaica en este emplazamiento específico, una pregunta que tiene tanto interés científico: ¿qué factores dominan el comportamiento del recurso solar en el noroeste de México? como implicaciones prácticas para el diseño de sistemas de monitoreo más eficientes que prioricen la instrumentación de las variables con mayor impacto sobre el pronóstico.

La metodología de modelado adoptada en el Capítulo II sigue las recomendaciones más sólidas de la literatura para conjuntos de datos de este tipo y escala. Se parte de una partición entrenamiento/prueba sobre los registros ordenados cronológicamente, preservando la estructura temporal de la serie, que sería destruida por una partición aleatoria, seguida de la normalización de los predictores mediante estandarización para garantizar la comparabilidad entre variables de distintas unidades y rangos. La selección formal de características se realiza mediante RFE con validación cruzada (RFECV), que identifica el subconjunto óptimo de predictores en términos de su contribución al desempeño del modelo. Sobre ese subconjunto se entrenan de forma comparativa cuatro arquitecturas que cubren un espectro amplio de complejidad y capacidad expresiva: un regresor ficticio (*Dummy Regressor*) que produce siempre la media del conjunto de entrenamiento y sirve como línea base estadística mínima; una regresión por descenso de gradiente estocástico (SGD) con regularización, que representa la familia de modelos lineales y cuyo desempeño relativo permite cuantificar cuánta complejidad no lineal está presente en los datos; un Perceptrón Multicapa (MLP) con capas ocultas, que puede capturar relaciones no lineales complejas sin necesidad de una especificación explícita de su forma funcional; y un Bosque Aleatorio (*Random Forest Regressor*, RFR), que es un método de ensamble basado en múltiples árboles de decisión y cuya robustez e interpretabilidad lo han convertido en uno de los algoritmos de referencia en la literatura de pronóstico fotovoltaico [34][36].

La elección de este conjunto específico de arquitecturas responde a un principio metodológico fundamental: antes de adoptar el modelo más complejo disponible, es necesario entender cuánta complejidad están demandando los datos y cuánta precisión adicional aporta cada nivel de

complejidad. Un modelo sencillo que alcanza un desempeño cercano al de uno complejo es, en la mayoría de los contextos operativos, preferible: es más fácil de mantener, más transparente para los operadores, más robusto ante cambios en el régimen de operación y menos propenso al sobreajuste ante conjuntos de datos de tamaño limitado. Solo cuando queda demostrado que el modelo sencillo no captura adecuadamente la variabilidad de los datos está justificado escalar hacia arquitecturas de mayor complejidad. Esta progresión metodológica es la que estructura el análisis comparativo del Capítulo II y la que permite formular, al final de ese análisis, recomendaciones fundadas en evidencia sobre el nivel de complejidad algorítmica apropiado para el contexto operativo de Puerto Peñasco.

BIBLIOGRAPHY

[1] Secretaría de Energía (SENER), *Prospectiva de Energías Renovables 2016–2030*, SENER, Ciudad de México, México, 2016. [En línea]. Disponible: https://www.gob.mx/cms/uploads/attachment/file/177622/Prospectiva_de_Energ_as_Renovables_2016-2030.pdf

[2] International Energy Agency (IEA), *Digitalization and Energy*, IEA, París, Francia, 2017.

[3] International Energy Agency (IEA), "Smart Grids," IEA, París, Francia. [En línea]. Disponible: https://www.iea.org/energy-system/electricity/smart-grids. [Accedido: Feb. 23, 2026].

[4] Bundesministerium für Wirtschaft und Klimaschutz (BMWK), "Neustart für die digitale Energiewende," BMWK, Berlín, Alemania, May. 2023. [En línea]. Disponible: https://www.bmwk.de

[5] Bundesministerium für Wirtschaft und Klimaschutz (BMWK), "Smart Energy Showcases – Digital Agenda for the Energy Transition (SINTEG)," BMWK, Berlín, Alemania, 2016. [En línea]. Disponible: https://www.bmwk.de/Redaktion/EN/Artikel/Energy/sinteg.html

[6] Natural Resources Canada (NRCan), "Green Infrastructure: Smart Grid Program," NRCan, Ottawa, Canadá, 2021. [En línea]. Disponible: https://www.nrcan.gc.ca

[7] Natural Resources Canada (NRCan), "Smart Grid Program — Results and Outcomes," NRCan, Ottawa, Canadá, 2025. [En línea]. Disponible: https://www.nrcan.gc.ca

[8] Ministerio de Energía de Chile, *Política Energética de Chile: Energía 2050*, Ministerio de Energía, Santiago, Chile, 2015.

[9] Comisión Nacional de Energía (CNE), *Estudio de Tarificación de Redes Inteligentes*, CNE, Santiago, Chile, 2012.

[10] Centro de Energía, Universidad de Chile, *Medidas Habilitantes para la Digitalización del Sector Eléctrico Chileno*, Informe para el Ministerio de Energía, Santiago, Chile, 2023.

[11] Coordinador Eléctrico Nacional (CEN), *Reporte de Operación del Sistema Eléctrico Nacional 2024*, CEN, Santiago, Chile, 2024. [En línea]. Disponible: https://www.coordinador.cl

[12] Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO), *Plan Nacional Integrado de Energía y Clima 2023–2030*, MITECO, Madrid, España. [En línea]. Disponible: https://www.miteco.gob.es

[13] Red Eléctrica de España (REE), *Informe del Sistema Eléctrico Español — Mayo 2024*, REE, Madrid, España, 2024. [En línea]. Disponible: https://www.ree.es/es/datos/publicaciones/informe-de-datos

[14] Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO), *Plan Nacional Integrado de Energía y Clima 2021–2030 (PNIEC)*, MITECO, Madrid, España, 2020.

[15] U.S. Department of Energy (DOE), Office of Electricity, "About the Office of Electricity," U.S. DOE, Washington, D.C., USA. [En línea]. Disponible: https://www.energy.gov/oe/office-electricity. [Accedido: Feb. 23, 2026].

[16] U.S. Congress, *Energy Independence and Security Act of 2007*, Pub. L. No. 110-140, 121 Stat. 1492, 2007.

[17] D. N.-Y. Mah, Y.-Y. Wu, J. C.-M. Ip, y P. R. Hills, "The role of the state in sustainable energy transitions: A case study of large smart grid demonstration projects in Japan," *Energy Policy*, vol. 63, pp. 726–737, 2013, doi: 10.1016/j.enpol.2013.07.106.

[18] Tokyo Metropolitan Government, "Obligation to install renewable energy equipment in new buildings," Tokyo Metropolitan Government, Tokio, Japón. [En línea]. Disponible: https://www.metro.tokyo.lg.jp. [Accedido: Feb. 23, 2026].

[19] Clean Energy Ministerial (CEM), "Uruguay — Renewable Energy Country Profile," CEM, Washington, D.C., USA, 2024. [En línea]. Disponible: https://www.cleanenergyministerial.org

[20] Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE), "Medidores Inteligentes," UTE, Montevideo, Uruguay. [En línea]. Disponible: https://www.ute.com.uy. [Accedido: Feb. 23, 2026].

[21] REN21, *Renewables 2024 Global Status Report*, REN21 Secretariat, París, Francia, 2024.

[22] Presidencia de la República Oriental del Uruguay, "Medidores inteligentes: Uruguay avanza en la modernización de su sistema eléctrico," Presidencia de la República Oriental del Uruguay, Montevideo, Uruguay, 2024. [En línea]. Disponible: https://www.presidencia.gub.uy

[23] Cámara de Diputados del H. Congreso de la Unión, *Ley de Transición Energética*, Diario Oficial de la Federación, Ciudad de México, México, 2015.

[24] Secretaría de Energía (SENER), *Programa de Redes Eléctricas Inteligentes*, SENER, Ciudad de México, México, 2016.

[25] National Renewable Energy Laboratory (NREL), *Smart Grid Implementation in Mexico: Opportunities and Challenges*, NREL, Golden, CO, USA, Tech. Rep., 2019.

[26] L. I. León-Trigo, E. Reyes-Archundia, J. A. Gutiérrez-Gnecchi, A. Méndez-Patiño, y G. M. Chávez-Campos, "Smart Grids en México: Situación actual, retos y propuesta de implementación," *Ingeniería, Investigación y Tecnología*, vol. 20, no. 2, 2019, doi: 10.22201/fi.25940732e.2019.20n2.015.

[27] S. Ferlito, G. Coppola, G. De Vito, y M. Ferrara, "A Study on an IoT-Based SCADA System for Photovoltaic Utility Plants," *Electronics*, vol. 13, no. 11, p. 2065, May 2024, doi: 10.3390/electronics13112065.

[28] M. I. Mohammed, A. A. Al-Sumaidae, y A. H. Al-Timimi, "Design of a SCADA System for a Solar Photovoltaic Power Plant," *NTU Journal of Engineering and Technology*, vol. 2, no. 2, 2023.

[29] L. K. Narayanan, M. Padmanabhan, P. Thangavel, y S. Subramanian, "A Survey on AI- and ML-Based Demand Forecast Analysis of Power Using IoT-Based SCADA," en *Smart Energy and Electric Power Systems*, Elsevier, 2023, pp. 65–78.

[30] O. V. Gnana Swathika, S. Hemavathi, y P. Kannan, "Fault Diagnosis for Photovoltaic Systems: A Validated Industrial SCADA Framework," *Applied Sciences*, vol. 15, no. 23, p. 12656, Nov. 2025, doi: 10.3390/app152312656.

[31] G. Karthikeyan y A. Jagadeeshwaran, "Enhancing solar energy generation: comprehensive ML-based PV prediction and fault analysis," *Electric Power Components and Systems*, vol. 52, pp. 1497–1512, 2023, doi: 10.1080/15325008.2023.2289038.

[32] E. Ogliari, F. Grimaccia, S. Leva, y M. Mussetta, "Hybrid predictive models for accurate forecasting in PV systems," *Energies*, vol. 6, no. 4, pp. 1918–1929, 2013, doi: 10.3390/en6041918.

[33] T. Persson, M. Bacher, A. Shaker, y H. Madsen, "Multi-site solar power forecasting using gradient boosted regression trees," *Solar Energy*, vol. 150, pp. 423–436, 2017, doi: 10.1016/j.solener.2017.04.066.

[34] A. Mellit, S. A. Kalogirou, L. Hontoria, y S. Shaari, "Artificial intelligence techniques for sizing photovoltaic systems: A review," *Renewable and Sustainable Energy Reviews*, vol. 13, no. 2, pp. 406–419, 2009, doi: 10.1016/j.rser.2008.01.006.

[35] B. B. Abubakar, C. M. L. Ugwuanyi, y M. N. Olutola, “Deep learning models for solar irradiance forecasting: A comprehensive review,” *Renewable Energy Focus*, vol. 44, pp. 176–196, 2023, doi: 10.1016/j.ref.2023.01.003.

[36] I. Guyon y A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[37] T. Kim, H. Li, R. Bhatt, y B. Palmintier, *Generative AI for Power Grid Operations*, National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep. NREL/TP-5D00-91176, 2024.

[38] Z. Liu, D. Che, C. Gao, y B. Huang, “Revolutionizing solar energy resources: The central role of generative AI in elevating system sustainability and efficiency,” *Applied Energy*, vol. 381, p. 125079, 2025, doi: 10.1016/j.apenergy.2025.125079.

CHAPTER II

Aprendizaje Automático Aplicado a la Operación Fotovoltaica: Metodología, Análisis de Datos y Modelos Predictivos sobre Registros Reales de Puerto Peñasco

El Capítulo I estableció el marco conceptual y tecnológico que sustenta esta investigación: revisó la evolución de las redes eléctricas inteligentes en distintos contextos nacionales, describió la arquitectura de los sistemas de adquisición de datos en centrales fotovoltaicas, analizó el estado del arte en modelos predictivos de generación solar y delimitó el papel potencial de los modelos de lenguaje de gran escala como capa de soporte operativo. Ese recorrido no fue un fin en sí mismo, sino la fundamentación argumentativa de las decisiones metodológicas que se ejecutan en este capítulo: si el Capítulo I responde al *qué* y al *por qué* de las herramientas de aprendizaje automático en el contexto fotovoltaico, el presente capítulo responde al *cómo*, materializando esos conceptos en un pipeline analítico concreto aplicado sobre registros operativos reales de la Central Fotovoltaica de Puerto Peñasco.

La transición del marco teórico a la práctica no es trivial. Demostrar que un conjunto de técnicas funciona sobre datos controlados de laboratorio es un paso necesario pero insuficiente para justificar su adopción operativa en una instalación real; lo que verdaderamente importa desde el punto de vista de la ingeniería es determinar qué nivel de precisión es alcanzable cuando los datos provienen de un sistema en operación continua, con las imperfecciones, discontinuidades y peculiaridades que eso implica, y en qué medida ese nivel de precisión tiene valor para quien toma decisiones sobre el despacho, el mantenimiento y la gestión de activos de la central. Esta es la pregunta central que estructura el trabajo analítico de este capítulo.

Para responderla, el capítulo se organiza en torno a siete componentes metodológicos articulados. En primer lugar, se describe el caso de estudio con el nivel de detalle necesario para contextualizar

cada decisión posterior: la ubicación geográfica y el perfil climático del emplazamiento, las características del conjunto de datos disponible y el significado operativo de cada variable registrada. En segundo lugar, se presenta el análisis exploratorio de los datos, que permite caracterizar la distribución de la variable objetivo, identificar los patrones de generación a distintas escalas temporales y mapear las relaciones entre las variables predictoras y la potencia de salida. En tercer lugar, se documenta el proceso de preprocesamiento, incluyendo la estrategia de partición del conjunto de datos, la transformación de variables categóricas y la normalización de los predictores. En cuarto lugar, se describe la selección formal de características mediante eliminación recursiva con validación cruzada, que identifica el subconjunto de variables con mayor capacidad predictiva sobre los datos de Puerto Peñasco. En quinto lugar, se presentan las cuatro arquitecturas de modelos implementadas en orden creciente de complejidad, describiendo en cada caso sus fundamentos algorítmicos, sus ventajas para el tipo de problema abordado, los aspectos específicos de su configuración y los resultados de la búsqueda de hiperparámetros. En sexto lugar, se establece el protocolo de evaluación y se definen las métricas utilizadas para comparar el desempeño de los modelos de manera rigurosa y reproducible. El capítulo cierra con una reflexión fundamentada sobre el papel potencial de los LLMs como capa de interpretación y comunicación de los resultados del pipeline desarrollado, antes de presentar las conclusiones que sirven de puente hacia el análisis comparativo del Capítulo III.

2.1 El Caso de Estudio: Central Fotovoltaica de Puerto Peñasco

La selección del caso de estudio es una decisión metodológica que condiciona el alcance y la interpretabilidad de todos los análisis posteriores. En este trabajo, la Central Fotovoltaica de Puerto Peñasco constituye no sólo el origen de los datos sino también el contexto que da sentido operativo a cada resultado: las métricas de desempeño de los modelos, la relevancia de las variables predictoras y las recomendaciones de implementación sólo son interpretables en relación con las características físicas, geográficas y sistémicas de esa instalación específica. Por ello, antes de introducir los datos y el análisis, es necesario establecer con precisión el marco de referencia que define el espacio de validez del estudio.

2.1.1 Caracterización del emplazamiento y condiciones operativas

La Central Fotovoltaica de Puerto Peñasco se localiza en el municipio homónimo del estado de Sonora, en el noroeste de México, en las coordenadas 31.32° de latitud norte y 113.28° de longitud oeste, a 18 metros sobre el nivel del mar. Esta ubicación tiene implicaciones físicas directas y determinantes sobre el comportamiento del recurso solar: a una latitud de algo más de 31 grados norte, la altura máxima del sol sobre el horizonte varía de forma significativa a lo largo del año, alcanzando su punto más bajo en el solsticio de invierno y su cenit en el solsticio de verano. Esta variación estacional de la geometría solar se traduce en diferencias sustanciales en la irradiancia

diaria disponible entre los meses de diciembre y los de junio-julio, diferencias que los modelos predictivos deben ser capaces de capturar con fidelidad para ser útiles a lo largo de todo el año.

Sonora es, en términos de recurso solar, uno de los estados con las condiciones más favorables del territorio nacional. Investigaciones de la Universidad de Sonora han confirmado que el estado cuenta con los niveles de irradiación solar más elevados del país: en menos del 1% de su vasta superficie árida, el recurso solar disponible es suficiente, en términos teóricos, para satisfacer la demanda eléctrica de toda la nación [1]. Un análisis posterior basado en datos satelitales de NASA SSE y de la Red Meteorológica Mexicana (SMN) reafirmó estas condiciones para las distintas regiones de la entidad, identificando que las zonas occidental y noroeste, donde se ubica Puerto Peñasco, combinan una alta irradiancia promedio con terreno relativamente plano, lo que facilita el desarrollo de proyectos solares a gran escala [2]. Esta excepcional disponibilidad del recurso es uno de los factores que motivó la elección de Puerto Peñasco para la instalación de la central más grande de América Latina.

Sin embargo, el clima del noroeste sonorense introduce también condiciones que complejizan la operación fotovoltaica. Durante los meses de julio y agosto, las temperaturas ambiente pueden superar con regularidad los 40°C, lo que activa el efecto negativo de la temperatura sobre el rendimiento de las células fotovoltaicas de silicio. Este efecto está físicamente bien documentado: la eficiencia y la potencia de salida de los módulos fotovoltaicos de silicio cristalino disminuyen de forma aproximadamente lineal con el aumento de la temperatura de operación, con un coeficiente de temperatura de potencia máxima (γ) que típicamente oscila entre $-0.35\%/°C$ y $-0.50\%/°C$ según la tecnología específica [3]. En la práctica, esto significa que un módulo que opera a 60°C, valor alcanzable en verano en Sonora cuando la irradiancia es alta, puede producir entre un 10% y un 15% menos potencia que la especificada bajo las condiciones estándar de medición a 25°C. Esta degradación térmica es uno de los factores que los modelos predictivos deben ser capaces de capturar para producir estimaciones precisas durante los meses de mayor temperatura.

La presencia estacional del monzón mexicano introduce episodios adicionales de nubosidad y precipitación que generan variabilidad adicional en la irradiancia disponible durante los meses de julio, agosto y septiembre. Los vientos del Golfo de California pueden alcanzar velocidades considerables en determinadas épocas del año, con implicaciones sobre el enfriamiento convectivo de los módulos y sobre la acumulación de polvo y arena en sus superficies. Este conjunto de condiciones: alta insolación base, calor extremo en verano, episodios monzónicos y vientos variables; define un perfil climático de alta variabilidad que convierte al emplazamiento en un banco de prueba exigente y representativo de las condiciones operativas del norte árido de México.

Desde el punto de vista estratégico, la central cumple un papel que trasciende la simple generación de energía eléctrica. Como se describió en la Introducción de esta tesis, la línea de transmisión que la conecta al Sistema Eléctrico Nacional constituye el primer vínculo permanente de Baja California con el sistema interconectado de México, poniendo fin a décadas de operación peninsular aislada. Esta función de anclaje sistémico convierte la confiabilidad de la

generación de Puerto Peñasco en una variable con consecuencias que se extienden más allá de los límites de la instalación: una predicción precisa de su producción horaria es información relevante no solo para la operación interna de la central sino también para el despacho del operador del sistema en la región noroeste y para la planificación de la reserva de capacidad en el nodo de interconexión.

2.1.2 Descripción del conjunto de datos

Una vez establecido el marco físico y operativo del emplazamiento, es posible caracterizar con mayor precisión el conjunto de datos sobre el que se construye el trabajo analítico de este capítulo, comprendiendo no solo sus estadísticos descriptivos sino también el significado operativo de cada variable en el contexto específico de Puerto Peñasco.

El enfoque metodológico adoptado para la selección de variables sigue la línea de investigación establecida por Pasion et al. [4], cuyo trabajo seminal demostró que es posible modelar la producción de sistemas fotovoltaicos horizontales con alta precisión, alcanzando un R^2 de 0.94 mediante bosques aleatorios, utilizando únicamente variables meteorológicas y de localización geográfica, sin necesidad de datos de irradiancia directa. Esta aproximación resulta especialmente valiosa en el contexto de instalaciones operativas reales donde los datos de irradiancia pueden ser difíciles de obtener con la cobertura y calidad requeridas para el entrenamiento de modelos, como es el caso de Puerto Peñasco. La selección de predictores del presente trabajo extiende la propuesta de Pasion et al. incorporando variables temporales adicionales, año y marca temporal, cuya relevancia para el pronóstico fotovoltaico ha quedado respaldada por la literatura posterior [5]. Esta línea metodológica ha sido continuada de forma independiente por Aravena-Cifuentes et al. [6], quienes replicaron y extendieron el trabajo original añadiendo la variable de altitud y aplicando codificación *one-hot* a la estación del año, alcanzando una mejora significativa en el R^2, respecto a los resultados de referencia. Más relevante aún para el presente trabajo es su hallazgo central sobre la especificidad geográfica: los experimentos de modelado por ubicación muestran que el desempeño varía sustancialmente entre emplazamientos, con valores de R^2 que oscilan entre 0.4 en localizaciones con condiciones climáticas atípicas y valores cercanos a 0.8 en emplazamientos más representativos del conjunto de entrenamiento, y que cuando se omite un sitio del entrenamiento los resultados se degradan hasta producir R^2 negativos. Esta evidencia refuerza la decisión metodológica central de este trabajo: desarrollar un modelo calibrado específicamente para las condiciones de Puerto Peñasco en lugar de intentar generalizar un modelo multi-emplazamiento.

El conjunto de datos fue compilado a partir del período comprendido entre diciembre de 2023 y agosto de 2025, con una frecuencia de muestreo horaria [7]. El archivo de datos originales contiene 20,987 entradas distribuidas en 18 columnas. De estas entradas, 18,971 corresponden a horas nocturnas o periodos sin generación activa, en los que la irradiancia es nula o el sistema no se encontraba en operación, lo que las hace no informativas para el entrenamiento de los modelos de

predicción de potencia activa. Tras eliminar los registros con valor nulo en PolyPwr, el subconjunto efectivo de análisis quedó conformado por 2,016 registros con generación documentada, que representan el 9.6% del total de entradas del archivo. Esta proporción es consistente con la relación horas-luz/horas-totales esperada en la latitud de Puerto Peñasco a lo largo del ciclo anual completo. Las variables con relevancia operativa directa y meteorológica documentada fueron retenidas para el análisis; la variable de potencia de salida del generador fotovoltaico (PolyPwr) fue obtenida de registros operativos verificados de la central, mientras que las variables meteorológicas complementarias fueron recopiladas de bases de datos abiertas correspondientes al emplazamiento. Todo el pipeline de procesamiento, entrenamiento y evaluación fue implementado en una libreta Jupyter de autoría propia [8], cuyas primeras celdas de carga de datos ejecutan esta operación:

```
# 20,987 entradas totales → 2,016 registros con generación activa (9.6 %)
```

La secuencia `dropna` seguida de `query('PolyPwr > 0')` implementa dos criterios independientes: el primero elimina filas donde el sensor no registró valor alguno (telemetría interrumpida); el segundo elimina registros con valor cero que corresponden a horas nocturnas en las que el sistema está en reposo pero el sensor devuelve una lectura nula en lugar de un valor faltante.

La resolución temporal de una hora resulta apropiada para las aplicaciones de pronóstico orientadas al despacho programado y a la coordinación con el mercado de energía de día anterior, que constituyen el horizonte operativo central que justifica el desarrollo de este tipo de herramientas. La literatura reciente ha extendido este enfoque hacia arquitecturas de pronóstico multi-horizonte que predicen simultáneamente la potencia disponible para distintos intervalos temporales futuros [9]; sin embargo, el alcance de este trabajo se circunscribe al horizonte horario único, coherente con los requerimientos del mercado de día anterior.

Antes de cualquier análisis, el conjunto de datos requirió un proceso de limpieza orientado a eliminar los registros con valores faltantes en la variable objetivo. La presencia de valores nulos en PolyPwr es inherente a la naturaleza de los datos operativos: interrupciones en la telemetría, labores de mantenimiento y reinicios del sistema de adquisición de datos generan huecos que deben ser identificados y tratados antes del entrenamiento. Del total de 20,987 registros horarios en el archivo original, 2,016 contienen generación efectiva documentada; sobre este subconjunto activo se aplicó la partición cronológica 80/20. Tras remover los registros con valores faltantes en PolyPwr, el conjunto de entrenamiento efectivo quedó conformado por 1,625 registros limpios y el conjunto de prueba por 391 registros, cada uno de los cuales describe el estado del sistema en una hora específica mediante las variables predictoras seleccionadas y el valor real de PolyPwr.

Las variables predictoras disponibles cubren los principales factores determinantes del comportamiento del recurso solar en el emplazamiento, organizados en tres categorías funcionales. El primer grupo, de naturaleza espacio-temporal, incluye: la marca temporal completa (Time), que

codifica con precisión el instante de cada registro dentro de la serie histórica; el año (Year), que captura tendencias interanuales en la operación del sistema; la hora del día (Hour), que refleja el ciclo diurno de la posición solar; el mes del año (Month), que recoge la variación estacional de la declinación solar; la estación del año (Season), que agrupa meses en cuatro categorías; y la latitud del emplazamiento (Latitude), que determina la geometría solar máxima para la ubicación. La inclusión de la latitud como predictor, aunque constante en este estudio de caso único, sigue la práctica de Pasion et al. [4] para sistemas distribuidos geográficamente y resulta relevante para la generalización del modelo a otros emplazamientos. El segundo grupo corresponde a las variables meteorológicas: temperatura ambiente (AmbientTemp), humedad relativa (Humidity), velocidad del viento (Wind.Speed), visibilidad horizontal (Visibility), presión atmosférica (Pressure) y cobertura de nubes (Cloud.Ceiling). El tercer elemento es la variable objetivo: potencia de salida fotovoltaica (PolyPwr), expresada en megawatts.

La siguiente tabla resume los estadísticos descriptivos del conjunto de datos completo:

Variable	**Categoría**	**Unidad**	**Mínimo**	**Máximo**	**Media**
PolyPwr	Objetivo	MW	0.00	811.94	225.90
Time	Espacio-temporal	–	–	–	–
Year	Espacio-temporal	–	2023	2025	–
Month	Espacio-temporal	–	1	12	–
Hour	Espacio-temporal	h	0	23	11.50
Latitude	Espacio-temporal	°	31.32	31.32	31.32
Season	Espacio-temporal	categórica	–	–	–
AmbientTemp	Meteorológica	°C	6.00	42.10	23.38
Humidity	Meteorológica	%	4.00	96.00	46.61
Wind.Speed	Meteorológica	m/s	0.32	80.32	15.78
Visibility	Meteorológica	km	0.20	90.00	43.18
Pressure	Meteorológica	hPa	1002.40	1026.60	1012.13
Cloud.Ceiling	Meteorológica	–	0.00	100.00	18.76

Table 2.1: Estadísticos descriptivos del conjunto de datos de Puerto Peñasco.

La desviación estándar de PolyPwr (288.96 MW) supera en valor absoluto a su propia media (225.90 MW), lo que revela que la variabilidad es el rasgo dominante de la serie y que el modelo que logre capturarla será genuinamente útil para la operación de la central.

2.2 Análisis Exploratorio de Datos

La descripción estadística del conjunto de datos presentada en la sección anterior es el punto de partida, pero no el destino, del trabajo analítico. El análisis exploratorio de datos (EDA, *Exploratory Data Analysis*) va más allá del resumen numérico para perseguir tres objetivos específicos y encadenados: caracterizar la distribución y el comportamiento de la variable objetivo para comprender la naturaleza del problema de predicción; identificar los patrones temporales que estructuran la serie de generación a distintas escalas: intradiaria, semanal y estacional; para determinar qué tipo de información temporal deben capturar los modelos; y mapear las relaciones

estadísticas entre las variables predictoras y la potencia de salida, lo que sirve tanto para orientar la selección formal de características como para interpretar físicamente el comportamiento del sistema. Este proceso de exploración es, en última instancia, el que convierte el conjunto de datos bruto en conocimiento accionable sobre el funcionamiento de la central.

2.2.1 Distribución y comportamiento de la variable objetivo

Los estadísticos descriptivos producidos por la libreta Jupyter [8] confirman cuantitativamente esta caracterización:

```
# PolyPwr: media = 225.90 MW, std = 288.96 MW, mediana = 89.72 MW, máx = 811.94 MW
# std > media → distribución asimétrica positiva; variabilidad dominante
```

La mediana de 89.72 MW, sustancialmente inferior a la media de 225.90 MW, refleja la asimetría positiva de la distribución: el grueso de los registros activos se concentra en valores bajos (horas de amanecer y atardecer), mientras que un subconjunto menor de registros en horas de plena irradiancia eleva la media hacia arriba. Esta asimetría implica que el regresor[1] ficticio que predice $\bar{y}_{train}$ sobreestimará sistemáticamente durante las horas marginales.

La distribución de PolyPwr en el conjunto de datos completo exhibe una morfología marcadamente bimodal que es característica de las series de potencia fotovoltaica en emplazamientos con ciclos diarios bien definidos [5]. El primer modo se concentra en torno a valores de cero megawatts y corresponde a las horas en las que el sol se encuentra por debajo del horizonte o la irradiancia es prácticamente nula; en la latitud de Puerto Peñasco, estas horas representan aproximadamente la mitad del total del dataset, con mayor proporción durante los meses de invierno y menor durante el verano. El segundo modo se extiende de forma relativamente difusa entre valores medios y altos, con un hombro pronunciado que refleja la diversidad de condiciones meteorológicas durante las horas de luz solar.

Esta distribución bimodal tiene consecuencias metodológicas directas. Un modelo que prediga siempre la media del conjunto de entrenamiento cometerá errores sistemáticamente grandes en ambos extremos de la distribución: sobreestimará durante todas las horas nocturnas y subestimará durante los períodos de máxima irradiancia. El coeficiente de determinación R^2 del regresor ficticio resultante se sitúa cercano o por debajo de cero, confirmando que la media aritmética es una predicción sustancialmente inferior a cualquier modelo que capture la estructura real de los datos. Cuantificar este punto de partida es metodológicamente indispensable: sin una línea base de referencia bien establecida, cualquier valor de R^2 de un modelo real carece de contexto interpretativo [10]. La desviación estándar de 288.96 MW, que supera en valor absoluto a la

[1] **Regresor**: en aprendizaje automático: modelo cuya salida es un valor numérico continuo, en contraposición al clasificador, cuya salida es una categoría discreta. La tarea de regresión consiste en aprender la función $f : \mathcal{X} \to \mathbb{R}$ que mapea un vector de características de entrada a un valor real; en este trabajo, ese valor es la potencia activa PolyPwr expresada en megawatts.

propia media de 225.90 MW, confirma que la variabilidad es el rasgo dominante de la serie y que el modelo que logre capturarla será genuinamente útil para la operación de la central.

2.2.2 Análisis de series temporales y patrones de generación

El análisis de la serie temporal de PolyPwr a distintas escalas revela tres capas de estructura que los modelos predictivos deben ser capaces de reproducir simultáneamente. La primera es el ciclo diurno: durante cualquier día del año, la potencia de salida sigue una curva que parte de cero en las horas nocturnas, aumenta de forma aproximadamente simétrica hasta alcanzar su máximo en torno al mediodía solar y desciende simétricamente hasta volver a cero al atardecer. La amplitud de esta curva varía considerablemente en función de las condiciones meteorológicas: en días completamente despejados puede superar los 700 MW, mientras que en días de cobertura nubosa densa puede mantenerse por debajo de los 100 MW incluso en las horas centrales.

La segunda capa de estructura es la estacional. La altura máxima del sol sobre el horizonte en la latitud de Puerto Peñasco varía de aproximadamente 35.2° en el solsticio de invierno a 82.3° en el solsticio de verano, una diferencia de casi 47 grados que tiene un impacto directo sobre la irradiancia incidente en el plano de los módulos. En términos prácticos, los días de diciembre y enero, incluso bajo cielos completamente despejados, generan valores máximos de PolyPwr sustancialmente inferiores a los registrados en junio y julio. Este efecto se ve además modulado por la temperatura: los meses de verano combinan una geometría solar más favorable con temperaturas más elevadas que reducen el rendimiento de las células, introduciendo una tensión entre los dos factores que los modelos no lineales capturan con mayor eficacia que los lineales [11].

La tercera capa corresponde a la variabilidad meteorológica interdiaria e intradiaria, que no responde a ningún patrón periódico estricto. El paso de sistemas nubosos asociados al monzón mexicano, los episodios de viento fuerte que generan sombreado parcial por polvo en suspensión, o las variaciones en la humedad relativa pueden reducir la generación en horas o días específicos de manera que no es predecible a partir de la geometría solar únicamente. Esta variabilidad residual es precisamente la que justifica incluir variables meteorológicas como predictores en los modelos: su función es capturar la fracción de la varianza de PolyPwr que no está determinada por la posición del sol sino por las condiciones atmosféricas del momento [4][12].

2.2.3 Análisis de correlaciones entre variables

Esta exploración se cuantifica en el cuaderno [8] mediante:

```
# r(AmbientTemp, PolyPwr) = +0.671  |  r(Hour, PolyPwr) = +0.595
# r(Humidity, PolyPwr)    = 0.329  |  r(Pressure, PolyPwr) = 0.122
```

La correlación negativa de Humidity ($r = -0.329$) con PolyPwr y la positiva de AmbientTemp ($r = 0.671$) reproducen exactamente el patrón documentado por Aravena-Cifuentes et al. [6] sobre el mismo dataset de referencia, lo que valida la coherencia estadística del conjunto de datos de Puerto Peñasco con la literatura.

La matriz de correlación calculada sobre las variables numéricas del conjunto de datos revela patrones de asociación estadística que son tanto físicamente interpretables como metodológicamente relevantes para el diseño de los modelos. Las correlaciones más altas con la variable objetivo corresponden, en orden decreciente de magnitud, a las variables temporales Hour y Month, y a la temperatura ambiente AmbientTemp. La correlación positiva con Hour refleja el ciclo diurno, mientras que la correlación con Month captura la variación estacional. La correlación positiva con AmbientTemp, aparentemente contraintuitiva dado que el calor reduce el rendimiento de las células, se explica porque en el conjunto de datos las temperaturas más altas están fuertemente correlacionadas con los períodos de mayor irradiancia, de modo que la temperatura actúa parcialmente como proxy de la radiación solar disponible, un resultado consistente con los hallazgos de Pasion et al. [4], quienes identificaron la temperatura ambiente como la variable individual más importante para la predicción de potencia fotovoltaica sin datos de irradiancia, y posteriormente confirmado por Aravena-Cifuentes et al. [6], cuyo análisis de correlaciones sobre el mismo dataset documentó una alta correlación positiva de PolyPwr con temperatura ambiente y cobertura de nubes, y una correlación negativa con humedad relativa y latitud.

Las correlaciones entre las propias variables predictoras revelan también dependencias con implicaciones metodológicas. La correlación negativa entre temperatura ambiente y humedad relativa refleja el comportamiento esperado en un clima árido: los días más calurosos tienden a ser los más secos, mientras que los episodios de humedad más alta, asociados a la influencia monzónica, corresponden a períodos de temperatura más moderada. La correlación negativa entre visibilidad y cobertura de nubes es igualmente esperable: una cobertura de nubes más alta reduce la visibilidad horizontal al introducir mayor dispersión atmosférica. Estas correlaciones cruzadas implican que no todas las variables predictoras aportan información genuinamente independiente: algunas pueden ser redundantes respecto a otras, y su inclusión simultánea en un modelo puede no añadir capacidad predictiva mientras incrementa la complejidad computacional. Este diagnóstico exploratorio refuerza la pertinencia de una etapa formal de selección de características, descrita en la sección 2.4, que aborde sistemáticamente esta cuestión.

2.3 Preprocesamiento y Preparación de los Datos

Los resultados del análisis exploratorio no solo describen el conjunto de datos sino que orientan directamente las decisiones de preprocesamiento. La presencia de una estructura temporal marcada indica que la partición de los datos debe respetar el orden cronológico; la existencia de una variable categórica de estación del año requiere una transformación adecuada antes del

entrenamiento; la heterogeneidad de escalas entre las variables predictoras indica la necesidad de estandarización; y la detección de registros incompletos en las series operativas impone un paso de limpieza previo al entrenamiento. En esta sección se documenta cada una de estas decisiones y su justificación metodológica. Todo el pipeline de preprocesamiento se implementó mediante la API de scikit-learn [13], que garantiza la consistencia y reproducibilidad de las transformaciones a lo largo de todas las etapas del flujo de trabajo.

2.3.1 Partición del conjunto de datos en entrenamiento y prueba

La partición del conjunto de datos en un subconjunto de entrenamiento y un subconjunto de prueba es una decisión metodológica con consecuencias directas sobre la validez de las métricas de desempeño reportadas. La práctica estándar en problemas de aprendizaje automático con datos independientes e idénticamente distribuidos es realizar esta partición de forma aleatoria; sin embargo, cuando los datos tienen estructura temporal, como es el caso de cualquier serie de registros operativos de una central de generación, la partición aleatoria destruye esa estructura y puede producir estimaciones de desempeño optimistas [14]. La razón es que si se mezclan aleatoriamente los registros antes de particionar, el conjunto de entrenamiento contendrá observaciones temporalmente adyacentes a las del conjunto de prueba, lo que permite al modelo aprender patrones de correlación temporal que no estarían disponibles en un escenario de pronóstico real.

Para preservar la integridad del protocolo de evaluación y garantizar que las métricas reportadas reflejen el desempeño que los modelos tendrían en condiciones de operación real, la partición se realizó respetando el orden cronológico de los registros: el 80% inicial del conjunto de datos se asignó al conjunto de entrenamiento, mientras que el 20% restante, correspondiente a los registros más recientes, constituyó el conjunto de prueba. Este enfoque es consistente con las recomendaciones de la literatura especializada en pronóstico de series temporales para sistemas energéticos [14][15].

La limpieza de los datos de entrenamiento, necesaria para eliminar registros con valores faltantes en la variable objetivo, resultó en un conjunto de entrenamiento efectivo de 1,625 registros limpios, con un conjunto de prueba de 391 registros. Estos valores son los que arroja directamente la celda de partición del cuaderno [8], que aplica la función `train_test_split` de scikit-learn con `shuffle=False` para preservar el orden temporal y reserva el 20% final del dataset:

```
# shuffle=False → partición cronológica: 1,625 entrenamiento | 391 prueba
```

Adicionalmente, se eliminaron del espacio de características las columnas no informativas para el modelado: Location, cuyo valor es constante para el emplazamiento único de Puerto Peñasco, Longitude, Altitude, igualmente constantes en este estudio de caso, y YRMODAHRMI, que contiene una representación textual de la marca temporal redundante con las variables temporales numéricas ya incluidas. La presencia de valores faltantes en los datos originales es habitual en

el contexto de la telemetría industrial: interrupciones momentáneas en la comunicación entre el equipo de campo y el sistema SCADA, labores de mantenimiento durante las cuales ciertos equipos de medición son desconectados, o fallas transitorias en los sensores pueden generar huecos en los registros que deben ser identificados y tratados antes del entrenamiento. La estrategia adoptada fue la eliminación de los registros incompletos del conjunto de entrenamiento, lo que garantiza que los modelos aprenden exclusivamente de observaciones con información completa. Esta decisión es conservadora pero metodológicamente sólida: el número de registros resultante es suficientemente amplio para entrenar los cuatro modelos sin riesgo de sobreajuste severo por escasez de datos [10].

2.3.2 Codificación de la variable categórica Season

La variable Season, que registra la estación del año en cuatro categorías, otoño (Fall), primavera (Spring), verano (Summer) e invierno (Winter), requirió una transformación antes de ser incorporada en los modelos. Los algoritmos de aprendizaje automático implementados no pueden operar directamente sobre variables categóricas nominales, ya que la representación numérica directa de las categorías (por ejemplo, 1=Fall, 2=Spring, 3=Summer, 4=Winter) introduciría una ordenación artificial que no existe en la realidad entre las estaciones del año. Para evitar este sesgo, se aplicó codificación *one-hot* (OneHotEncoder), que transforma la variable categórica original en cuatro variables binarias independientes, una por estación, donde cada variable toma el valor 1 si el registro corresponde a dicha estación y 0 en caso contrario. Este tratamiento es la práctica estándar para variables categóricas nominales en pipelines de aprendizaje supervisado [13], y fue igualmente adoptado por Aravena-Cifuentes et al. [6] en su extensión del mismo dataset de referencia, donde demostraron que la incorporación de la variable Season codificada en formato *one-hot* contribuye a mejorar el desempeño del modelo frente a la formulación original de Pasion et al. La implementación dentro del pipeline de scikit-learn garantiza que la transformación se aplique de forma consistente tanto al conjunto de entrenamiento como al de prueba.

En la implementación del cuaderno [8], esta transformación se encapsula dentro del `ColumnTransformer` que forma parte del pipeline principal:

```
# Season → {Season_Fall, Season_Spring, Season_Summer, Season_Winter}  {0, 1}
```

2.3.3 Estandarización de las variables predictoras

La estandarización de las variables predictoras mediante transformación a media cero y desviación estándar unitaria (StandardScaler) es un paso de preprocesamiento indispensable para los modelos que son sensibles a la escala de las variables de entrada. Formalmente, dado un predictor x_j con $j = 1, \ldots, p$, se estiman la media muestral $\hat{\mu}_j$ y la desviación estándar muestral $\hat{\sigma}_j$ sobre el conjunto de entrenamiento de n observaciones:

$$\underbrace{\hat{\mu}_j}_{\text{media}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_{ij}}_{\text{media muestral}}, \qquad \underbrace{\hat{\sigma}_j}_{\text{desv. estándar}} = \sqrt{\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\hat{\mu}_j\right)^2}_{\text{varianza muestral}}} \tag{2.1}$$

A continuación, cada valor del predictor se transforma mediante la puntuación z estándar:

$$x'_{ij} = \frac{\overbrace{x_{ij}-\hat{\mu}_j}^{\text{centrado}}}{\underbrace{\hat{\sigma}_j}_{\text{escalado}}} \tag{2.2}$$

de modo que la variable transformada x'_j tiene media muestral cero y desviación estándar unitaria sobre el conjunto de entrenamiento. Los parámetros $\hat{\mu}_j$ y $\hat{\sigma}_j$ se estiman exclusivamente sobre el conjunto de entrenamiento y se aplican sin reestimación al conjunto de prueba, es decir, si x_{ij}^{test} es una observación del conjunto de prueba, se transforma como $(x_{ij}^{test}-\hat{\mu}_j)/\hat{\sigma}_j$ con los mismos parámetros del entrenamiento, garantizando que el escalador no introduzca información futura durante la evaluación [10].

La necesidad de esta transformación en el contexto específico de este conjunto de datos es inmediata al examinar los rangos de las variables: la temperatura ambiente varía en 36 °C, la velocidad del viento en 80 m/s, la presión atmosférica en 24.2 hPa y la visibilidad en 89.8 km. Si estos predictores se ingresaran en su escala original a un modelo cuya función de pérdida o cuyos gradientes dependen de la magnitud absoluta de los valores de entrada, como es el caso del Regresor por Descenso de Gradiente Estocástico[2] (SGD) y el Perceptrón[3] Multicapa (MLP), las variables con mayor rango numérico dominarían artificialmente el proceso de optimización independientemente de su relevancia real para la predicción [16].

Este efecto es especialmente pronunciado en el caso del SGD, cuyos coeficientes son directamente proporcionales a la escala de las variables de entrada, y en el caso del MLP, donde las magnitudes de los pesos en cada neurona dependen de la escala de las activaciones. Para el Bosque Aleatorio, en cambio, la estandarización es indiferente al desempeño predictivo: los árboles de decisión realizan divisiones binarias invariantes ante transformaciones monotónicas, por lo que su inclusión en el pipeline unificado no introduce ningún sesgo.

[2]**Estocástico**: en optimización: calificativo que indica que el algoritmo incorpora aleatoriedad deliberada en su proceso de actualización. En el Descenso de Gradiente Estocástico, los parámetros del modelo se actualizan a partir del gradiente evaluado sobre una sola observación o un subconjunto reducido seleccionado aleatoriamente, en lugar del gradiente exacto sobre la totalidad de los datos de entrenamiento; esto introduce ruido controlado que puede ayudar al algoritmo a escapar de mínimos locales subóptimos y acelerar la convergencia en las fases iniciales del entrenamiento.

[3]**Perceptrón**: en aprendizaje automático: unidad computacional elemental inspirada en la neurona biológica, propuesta por Rosenblatt (1958). Calcula una combinación lineal ponderada de sus entradas y aplica una función de activación no lineal al resultado. El Perceptrón Multicapa apila múltiples capas de estas unidades, permitiendo al modelo aprender representaciones progresivamente más abstractas y capturar relaciones no lineales entre las variables de entrada y la variable objetivo.

En el cuaderno [8], el escalador se encapsula junto con cada estimador dentro de un objeto `Pipeline` de scikit-learn [13]:

```
# Pipeline: [StandardScaler → estimador]
# Los parámetros del escalador se estiman solo sobre el conjunto de entrenamiento
```

Esta arquitectura garantiza automáticamente el comportamiento correcto: durante `fit()`, el `StandardScaler` llama a `fit_transform()` sobre los datos de entrenamiento y almacena $\hat{\mu}_j$ y $\hat{\sigma}_j$; durante `predict()` sobre cualquier conjunto externo, llama a `transform()` con los mismos parámetros, nunca a `fit_transform()`, de modo que la separación entre los estadísticos del entrenamiento y los datos del conjunto de prueba es estructuralmente imposible de violar.

2.4 Selección Formal de Características mediante RFECV

El preprocesamiento establece los datos en el formato adecuado para el entrenamiento, pero no resuelve todavía la pregunta de cuáles de las variables disponibles deben efectivamente alimentar los modelos. Esta pregunta, que la sección 2.2.3 dejó planteada como hipótesis exploratoria, es abordada en esta sección de forma rigurosa mediante un procedimiento formal de selección de características.

2.4.1 Fundamento metodológico

El conjunto de variables predictoras disponibles, incluyendo las cuatro columnas binarias generadas por la codificación *one-hot* de Season, las variables espacio-temporales, y las variables meteorológicas, plantea la pregunta de si todas contribuyen de manera significativa a la capacidad predictiva de los modelos o si algunas son redundantes, ruidosas o de utilidad marginal. Esta pregunta no es trivial: la inclusión de variables irrelevantes o redundantes puede degradar el desempeño de los modelos al introducir ruido en el espacio de características e incrementar el riesgo de sobreajuste, especialmente en conjuntos de datos de tamaño moderado donde ese riesgo es más pronunciado [17].

La técnica de Eliminación Recursiva de Características con Validación Cruzada (RFECV, *Recursive Feature Elimination with Cross-Validation*) aborda este problema mediante un proceso iterativo que evalúa sistemáticamente el desempeño de los modelos entrenados con diferentes subconjuntos de variables, identificando el subconjunto de tamaño mínimo que maximiza la capacidad predictiva sin pérdida estadísticamente significativa respecto al uso de todas las variables disponibles [17]. El algoritmo opera de la siguiente manera: en cada iteración, entrena un modelo base sobre el conjunto de entrenamiento con el subconjunto actual de variables, evalúa su desempeño mediante validación cruzada, identifica la variable de menor importancia y la elimina del subconjunto, y repite el proceso hasta haber evaluado todos los tamaños posibles. El tamaño

óptimo es aquel que maximiza la métrica de validación cruzada. La validación cruzada interna del RFECV se realizó mediante *KFold*[4] con tres particiones ($k = 3$) y mezcla aleatoria de los datos.

La elección del modelo base para RFECV es relevante porque determina cómo se estima la importancia de cada variable. La implementación correspondiente en el cuaderno [8] instancia el estimador base y el objeto RFECV con la siguiente configuración:

```
# RFECV: estimador = ExtraTreesRegressor, cv = KFold(3), scoring = 'r2'
```

El parámetro `step=1` garantiza que el algoritmo evalúe todos los tamaños posibles de subconjunto, desde el conjunto completo de variables hasta un único predictor, seleccionando el tamaño que maximiza el R^2 de validación cruzada. En este trabajo se utilizó un ExtraTreesRegressor (Bosque de Árboles Extremadamente Aleatorizados) como estimador base por dos razones complementarias. En primer lugar, este tipo de modelo produce estimaciones de importancia de variable robustas basadas en la reducción de impureza promedio ponderada a lo largo de todos los árboles del ensamble [18]. En segundo lugar, el ExtraTreesRegressor es computacionalmente eficiente para aplicaciones de este tipo, ya que selecciona umbrales de división de forma aleatoria entre el mínimo y el máximo de cada variable, reduciendo el costo computacional de cada árbol sin comprometer la calidad de la estimación de importancias respecto al Random Forest estándar [19].

2.4.2 Resultado: número óptimo de características y su importancia relativa

El proceso de RFECV identificó un subconjunto óptimo de nueve características como el de mayor capacidad predictiva sobre los datos de Puerto Peñasco. La curva de desempeño en función del número de características muestra una mejora pronunciada al añadir las primeras más relevantes, y una estabilización seguida de ligera degradación al incorporar las menos informativas, confirmando que la selección de un subconjunto estrictamente menor que el conjunto completo beneficia la generalización de los modelos. La validación cruzada de RFECV alcanzó un R^2 de 0.9991, lo que indica que el subconjunto seleccionado es capaz de capturar prácticamente toda la variabilidad explicable de PolyPwr con el modelo base de ExtraTrees. Estos valores corresponden directamente a los que produce la celda de selección de características del cuaderno [8], cuya configuración central fue:

```
# Número óptimo de características: 9  |  R² de validación cruzada: 0.9991
```

El perfil de importancias relativas de las variables seleccionadas es el siguiente: Year (importancia relativa: 0.5044), Time (0.4576), Latitude (0.0124), Hour (0.0099), Month (0.0036), Date (0.0041),

[4] *K-Fold Cross-Validation*: técnica que divide el conjunto de datos en k subconjuntos de igual tamaño (*folds*). En cada una de las k iteraciones, el modelo se entrena sobre los $k-1$ folds restantes y se evalúa sobre el reservado; la métrica final es el promedio de las k evaluaciones. En este trabajo, $k = 3$.

AmbientTemp (0.0034), Wind.Speed (0.0035) y Humidity (0.0011). Las dos primeras variables, juntas, concentran aproximadamente el 96.2% de la importancia total del modelo. Los valores de FI(X_j) reportados corresponden al atributo `rfecv.estimator_.feature_importances_` del ExtraTreesRegressor entrenado sobre el subconjunto final seleccionado, calculados según la fórmula de reducción de impureza ponderada de la sección 2.5.4, ecs. (**??**)–(2.21).

2.4.3 Interpretación física del perfil de importancias

El resultado de RFECV no es únicamente estadístico: es también una fuente de conocimiento físico sobre el sistema que merece una interpretación cuidadosa. El peso absolutamente dominante de las variables temporales, Year, Time y Hour, sobre las variables meteorológicas puede parecer sorprendente a primera vista, pero es perfectamente coherente con el comportamiento de la generación fotovoltaica en un emplazamiento de alta insolación árida como Puerto Peñasco, y está respaldado por la literatura reciente sobre importancia de variables en modelos de bosques aleatorios aplicados a sistemas solares [12].

La razón fundamental es que la geometría solar, completamente determinista y capturada con precisión por las variables temporales, explica la mayor parte de la varianza de PolyPwr. La alternancia entre el día y la noche y la variación estacional de la declinación solar son los dos factores que más influyen sobre cuánta energía llega a los módulos en cada momento. Time como marca temporal granular captura el ciclo diurno con mayor resolución que Hour en solitario, al preservar la posición exacta en la serie temporal global, mientras que Year actúa como descriptor de la tendencia interanual que refleja el crecimiento de la capacidad instalada durante el período de datos, un patrón de largo plazo que ninguna variable meteorológica podría capturar. Las variables meteorológicas modulan esa generación potencial de forma relativamente menor en promedio, aunque pueden ser críticas en situaciones específicas como episodios de nubosidad densa o tormentas de polvo. Este resultado diverge parcialmente de los hallazgos de Pasion et al. [4], quienes identificaron la temperatura ambiente como la variable más importante en un contexto multi-emplazamiento; la diferencia se explica porque en ese estudio la variación geográfica entre sitios aportaba información adicional que las variables meteorológicas capturaban, mientras que en el caso de Puerto Peñasco, con un único emplazamiento, la variabilidad entre registros está dominada por la dimensión temporal.

Esta interpretación tiene además implicaciones prácticas para el diseño de futuros sistemas de monitoreo: sugiere que, para aplicaciones de pronóstico horario en este emplazamiento específico, la disponibilidad de una referencia temporal precisa es más crítica para la calidad del pronóstico que la instrumentación de sensores meteorológicos adicionales. Esto no significa que las variables meteorológicas sean irrelevantes, especialmente para el pronóstico de eventos de nubosidad extrema, sino que su contribución marginal sobre la base de la geometría solar es más modesta de lo que intuitivamente podría esperarse en un emplazamiento de tan alta insolación.

2.5 Descripción de los Modelos Implementados

Con el subconjunto de características seleccionado y los datos preparados, el siguiente paso es definir las arquitecturas de los modelos que se entrenarán y compararán. La selección de estas arquitecturas responde al principio metodológico de progresión de complejidad [10][20]: antes de adoptar el modelo más sofisticado disponible, es necesario entender cuánta complejidad están demandando los datos y cuánta precisión adicional aporta cada nivel adicional de sofisticación. Para articular esa progresión de manera sistemática, se implementaron cuatro arquitecturas en orden creciente de capacidad expresiva, desde el modelo más simple concebible hasta un método de ensamble de alta complejidad, diseño que permite no solo identificar el modelo de mejor desempeño absoluto, sino también diagnosticar la estructura del problema: si el modelo lineal funciona bien, la relación entre predictores y potencia es aproximadamente lineal; si solo los modelos no lineales complejos funcionan bien, la complejidad del problema justifica esa inversión adicional [5][11]. Todos los modelos se implementaron dentro de pipelines de scikit-learn [13] que integran, en secuencia, la estandarización de variables y el estimador correspondiente, garantizando que el preprocesamiento se aplique de forma consistente durante el entrenamiento y la inferencia.

2.5.1 Regresor Ficticio (DummyRegressor): Línea Base

El Regresor Ficticio, implementado con la estrategia de predicción por media (*mean strategy*), produce para toda observación del conjunto de prueba exactamente la misma predicción: la media aritmética de los valores de PolyPwr en el conjunto de entrenamiento. Formalmente, dado el conjunto de entrenamiento $\{y_1, y_2, \ldots, y_n\}$, la predicción del modelo para cualquier vector de entrada $\mathbf{x}$ es:

$$\underbrace{\hat{y}(\mathbf{x})}_{\substack{\text{predicción} \\ \text{para cualquier } \mathbf{x}}} = \underbrace{\bar{y}_{train}}_{\text{constante}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\text{media del entrenamiento}} \tag{2.3}$$

Este modelo no utiliza ninguna de las variables predictoras disponibles y, en consecuencia, no captura ninguna estructura en los datos. Su inclusión en el análisis comparativo tiene una justificación metodológica precisa: establece el umbral mínimo de referencia por debajo del cual ningún modelo estadístico tiene sentido de ser utilizado [10].

El regresor ficticio se instancia en el cuaderno [8] dentro del mismo esquema de `Pipeline` utilizado para los modelos más complejos:

```
# DummyRegressor(strategy='mean') → ŷ = 225.90 MW (constante, independiente de X)
```

La aserción final confirma formalmente que el modelo no utiliza ninguna variable predictora: la predicción es idéntica para cualquier observación del conjunto de prueba, independientemente de las condiciones meteorológicas o temporales del registro.

Si un modelo aprende algo genuinamente útil de los datos, su desempeño debe superar de forma significativa y consistente el del regresor ficticio. En términos del coeficiente de determinación R^2, el regresor ficticio produce un valor de exactamente cero por construcción, ya que $SS_{res} = SS_{tot}$ cuando $\hat{y}_i = \bar{y}$ para todo i. Un R^2 negativo no indica simplemente un modelo deficiente: indica formalmente que el modelo realiza predicciones peores que simplemente predecir siempre el valor medio, lo que es el peor resultado concebible para un regresor. Este umbral de comparación es indispensable para interpretar correctamente el desempeño de las arquitecturas más complejas presentadas a continuación.

2.5.2 Regresor por Descenso de Gradiente Estocástico (SGDRegressor)

El Regresor por Descenso de Gradiente Estocástico (SGD) pertenece a la familia de los modelos lineales regularizados y ajusta iterativamente un conjunto de coeficientes que ponderan linealmente las variables predictoras para minimizar una función de pérdida, actualizando los coeficientes tras procesar cada observación individual o un pequeño subconjunto, de ahí el calificativo "estocástico", en lugar de procesarlas todas de forma conjunta [16]. La predicción del modelo para un vector de entrada $\mathbf{x} \in \mathbb{R}^p$ tiene la forma:

$$\hat{y} = \underbrace{\mathbf{w}^\top \mathbf{x}}_{\text{combinación lineal}} + \underbrace{b}_{\text{sesgo}} = \underbrace{\sum_{j=1}^{p} w_j x_j}_{\text{suma ponderada de predictores}} + b \tag{2.4}$$

donde $\mathbf{w} \in \mathbb{R}^p$ es el vector de coeficientes y $b \in \mathbb{R}$ es el término independiente. Este modelo asume que la relación entre los predictores y la variable objetivo es lineal, lo que representa la hipótesis más restrictiva de entre todas las arquitecturas evaluadas.

La función de pérdida seleccionada mediante búsqueda de hiperparámetros fue la pérdida *squared epsilon-insensitive*, originalmente introducida en el contexto de las máquinas de vectores de soporte para regresión [21]. Esta función introduce una zona muerta de anchura 2ε centrada en la predicción, dentro de la cual los errores inferiores a ε no contribuyen al gradiente y, por tanto, no penalizan al modelo; los errores que superan ese umbral se penalizan cuadráticamente:

$$L_{\varepsilon^2}(y, \hat{y}) = \max\left(0, \underbrace{|y - \hat{y}|}_{\text{error absoluto}} - \underbrace{\varepsilon}_{\text{zona muerta}}\right)^2 \tag{2.5}$$

Esta elección dota al modelo de cierta robustez ante errores pequeños inevitables en los datos de entrenamiento, a la vez que penaliza con mayor fuerza las desviaciones grandes.

La regularización se implementa mediante penalización de tipo ElasticNet, propuesta por Zou y Hastie [22] como combinación convexa de las normas ℓ_1 y ℓ_2 sobre los coeficientes. El funcional de optimización global resulta:

$$J(\mathbf{w}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_{\varepsilon^2}(y_i, \hat{y}_i)}_{\text{pérdida empírica}} + \alpha \left[\underbrace{\rho \, \|\mathbf{w}\|_1}_{\text{penalización } \ell_1} + \underbrace{\frac{1-\rho}{2} \, \|\mathbf{w}\|_2^2}_{\text{penalización } \ell_2} \right] \tag{2.6}$$

donde $\alpha > 0$ controla la intensidad global de la regularización y $\rho \in [0,1]$ pondera el balance entre los dos términos de penalización. El término $\ell_1 = \|\mathbf{w}\|_1 = \sum_j |w_j|$ promueve la dispersión de la solución, llevando a cero los coeficientes de variables irrelevantes; el término $\ell_2 = \|\mathbf{w}\|_2^2 = \sum_j w_j^2$ estabiliza la solución ante la presencia de variables correlacionadas entre sí, evitando que coeficientes de magnitud arbitrariamente grande se anulen mutuamente. La ElasticNet hereda las ventajas de ambas regularizaciones, siendo especialmente apropiada cuando el número de predictores es moderado y existe cierta colinealidad entre variables, condición que se cumple en el conjunto de datos de Puerto Peñasco, donde variables como temperatura y hora del día presentan correlación positiva [22].

En cada iteración t, los coeficientes se actualizan en la dirección del gradiente negativo de la función de pérdida evaluada sobre una sola observación muestreada aleatoriamente $(\mathbf{x}_i, y_i)$:

$$\mathbf{w}^{(t+1)} = \underbrace{\mathbf{w}^{(t)}}_{\text{pesos actuales}} - \underbrace{\eta_t}_{\substack{\text{tasa de} \\ \text{aprendizaje}}} \underbrace{\nabla_{\mathbf{w}} J\left(\mathbf{w}^{(t)}; \mathbf{x}_i, y_i\right)}_{\text{gradiente estocástico}} \tag{2.7}$$

La aleatoriedad en la selección de la observación introduce ruido en el gradiente que, paradójicamente, puede ayudar al modelo a escapar de mínimos locales subóptimos y acelerar la convergencia en las fases iniciales del entrenamiento. Con la estrategia de tasa de aprendizaje adaptativa (*adaptive*) seleccionada como óptima, η_t se reduce por un factor constante cada vez que el error de entrenamiento no mejora durante un número predefinido de épocas consecutivas, con valor inicial $\eta_0 = 1$.

Los hiperparámetros del SGD se exploraron mediante RandomizedSearchCV con 1,500 iteraciones y validación cruzada de 3 particiones [23]. El proceso de búsqueda tuvo una duración de 14 minutos y 44 segundos, registrada en la salida de la celda correspondiente del cuaderno [8]. El espacio de búsqueda comprendió: función de pérdida en {squared_error, epsilon_insensitive, squared_epsilon_insensitive}; tipo de regularización en {l1, l2, elasticnet}; intensidad $\alpha \in [0.0001, 1000]$; estrategia de tasa de aprendizaje en {constant, optimal, invscaling, adaptive}; valor inicial $\eta_0 \in \{1, 10, 100\}$; y promediado de pesos (average) en {True, False}. La configuración óptima encontrada fue: $\eta_0 = 1$, `learning_rate = 'adaptive'`, `loss`

= 'squared_epsilon_insensitive', penalty = 'elasticnet', max_iter = 4500. Esta combinación confirma empíricamente la pertinencia teórica de la regularización ElasticNet argumentada en la sección anterior: ante los 1,500 puntos explorados en el espacio de hiperparámetros, el algoritmo de búsqueda seleccionó `'elasticnet'` sobre las alternativas `'l1'` y `'l2'`, lo que indica que la colinealidad moderada entre los predictores justifica la combinación de ambos términos de penalización [22]. Durante el entrenamiento se observaron advertencias de no convergencia en la salida del notebook, consistentes con la sensibilidad del SGD a la topografía compleja de la función de pérdida en presencia de relaciones no lineales entre los predictores y PolyPwr.

El modelo SGD cumple en este trabajo una función diagnóstica precisa: cuantifica directamente cuánta variabilidad de PolyPwr puede ser capturada por relaciones lineales entre los predictores y la salida, sirviendo como punto de referencia intermedio entre la línea base del regresor ficticio y los modelos no lineales. Si la brecha de desempeño entre el SGD y los modelos no lineales es grande, ese resultado constituye evidencia de que la generación fotovoltaica en Puerto Peñasco involucra interacciones no lineales significativas que ningún modelo de la familia lineal, ec. (2.4), puede capturar, hipótesis que la literatura reciente sobre pronóstico fotovoltaico respalda de forma consistente [11][12].

2.5.3 Perceptrón Multicapa (MLPRegressor)

El Perceptrón Multicapa (MLP) es una arquitectura de red neuronal artificial de propagación hacia adelante (*feedforward*) compuesta por una capa de entrada, una o más capas ocultas con funciones de activación no lineales, y una capa de salida lineal para tareas de regresión. Dado un vector de entrada $\mathbf{x} \in \mathbb{R}^p$, la computación en cada capa $l = 1, \ldots, L$ sigue la forma:

$$\mathbf{z}^{(l)} = \underbrace{\mathbf{W}^{(l)}\mathbf{a}^{(l-1)}}_{\text{transformación lineal}} + \underbrace{\mathbf{b}^{(l)}}_{\text{sesgo}} \tag{2.8}$$

$$\underbrace{\mathbf{a}^{(l)}}_{\text{activación}} = \underbrace{\sigma}_{\substack{\text{función de} \\ \text{activación}}}\left(\underbrace{\mathbf{z}^{(l)}}_{\text{entrada lineal}}\right) \tag{2.9}$$

donde $\mathbf{W}^{(l)} \in \mathbb{R}^{h_l \times h_{l-1}}$ es la matriz de pesos de la capa l, $\mathbf{b}^{(l)} \in \mathbb{R}^{h_l}$ el vector de sesgos, h_l el número de neuronas en la capa l, $\mathbf{a}^{(0)} = \mathbf{x}$ la entrada y $\sigma(\cdot)$ la función de activación aplicada elemento a elemento. La predicción final es la activación lineal de la capa de salida: $\hat{y} = \mathbf{w}^{(L+1)\top}\mathbf{a}^{(L)} + b^{(L+1)}$.

En la configuración adoptada para el presente trabajo, la red tiene $p = 9$ entradas (las características seleccionadas por RFECV), dos capas ocultas de $h_1 = h_2 = 50$ neuronas cada una ($L = 2$) y una sola neurona de salida, lo que resulta en una arquitectura $9 \to 50 \to 50 \to 1$ con un total de $9 \times 50 + 50 + 50 \times 50 + 50 + 50 \times 1 + 1 = 3{,}101$ parámetros entrenables.

La función de activación empleada en las capas ocultas es la ReLU (*Rectified Linear Unit*):

$$\text{ReLU}(z) = \underbrace{\max(0, z)}_{\text{rectificación}} = \begin{cases} \underbrace{z}_{\text{región activa}} & \text{si } z > 0 \\ \underbrace{0}_{\text{región inactiva}} & \text{si } z \leq 0 \end{cases} \tag{2.10}$$

Esta función es no lineal pero con derivada constante en su región activa ($z > 0$), lo que mitiga el problema del desvanecimiento del gradiente que afecta a funciones sigmoidal y tangente hiperbólica en redes profundas [24]. Su no saturación para valores positivos grandes permite que los gradientes fluyan eficazmente durante la retropropagación. La capa de salida utiliza activación identidad, apropiada para problemas de regresión donde la variable objetivo no está acotada.

La capacidad fundamental del MLP para aproximar cualquier función continua con precisión arbitraria queda garantizada por el teorema de aproximación universal, demostrado por Hornik [25]: una red neuronal de propagación hacia adelante con una sola capa oculta y función de activación no constante, acotada y continua puede aproximar cualquier función continua $f : \mathbb{R}^p \to \mathbb{R}$ sobre un compacto, con error uniforme tan pequeño como se desee, siempre que el número de neuronas ocultas sea suficientemente grande. Esta propiedad fundamenta teóricamente la elección del MLP como segunda arquitectura no lineal en la progresión de complejidad del experimento.

El funcional de pérdida optimizado durante el entrenamiento combina el error cuadrático medio con un término de regularización L2, también llamada penalización de Tikhonov, sobre todos los pesos de la red:

$$\mathscr{L}(\boldsymbol{\theta}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{error cuadrático medio}} + \underbrace{\frac{\alpha}{2}\sum_{l=1}^{L}\left\|\mathbf{W}^{(l)}\right\|_F^2}_{\text{regularización } L_2} \tag{2.11}$$

donde $\boldsymbol{\theta} = \{\mathbf{W}^{(l)}, \mathbf{b}^{(l)}\}_{l=1}^{L+1}$ denota el conjunto de todos los parámetros entrenables, $\|\cdot\|_F$ es la norma de Frobenius y $\alpha = 0.001$ es la intensidad de regularización seleccionada. Los gradientes $\nabla_{\boldsymbol{\theta}}\mathscr{L}$ se calculan eficientemente mediante el algoritmo de retropropagación del error (*backpropagation*), que aplica la regla de la cadena de forma recursiva desde la capa de salida hacia la capa de entrada [24].

Los pesos se actualizan con el optimizador Adam (*Adaptive Moment Estimation*) [26], que mantiene estimaciones del primer momento m_t (media del gradiente) y segundo momento v_t (varianza no centrada) para cada parámetro θ:

(2.12)
$$m_t = \underbrace{\beta_1 m_{t-1}}_{\substack{\text{memoria del} \\ \text{momento previo}}} + \underbrace{(1-\beta_1) g_t}_{\text{gradiente actual}}$$

(2.13)
$$v_t = \underbrace{\beta_2 v_{t-1}}_{\substack{\text{varianza} \\ \text{acumulada}}} + \underbrace{(1-\beta_2) g_t^2}_{\text{gradiente al cuadrado}}$$

donde $g_t = \nabla_\theta \mathscr{L}$ es el gradiente en el paso t. Dado que m_t y v_t se inicializan en cero, Adam aplica una corrección de sesgo para obtener estimaciones insesgadas en los primeros pasos:

(2.14)
$$\hat{m}_t = \underbrace{\frac{m_t}{1-\beta_1^t}}_{\substack{\text{corrección de sesgo} \\ \text{primer momento}}} , \qquad \hat{v}_t = \underbrace{\frac{v_t}{1-\beta_2^t}}_{\substack{\text{corrección de sesgo} \\ \text{segundo momento}}}$$

La actualización de cada parámetro combina ambas estimaciones para obtener una tasa de aprendizaje efectiva que se adapta individualmente a la escala del gradiente de ese parámetro:

(2.15)
$$\theta_{t+1} = \theta_t - \underbrace{\frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}}_{\text{tasa adaptativa}} \cdot \underbrace{\hat{m}_t}_{\substack{\text{gradiente} \\ \text{suavizado}}}$$

Los valores estándar $\beta_1 = 0.9$, $\beta_2 = 0.999$ y $\epsilon = 10^{-8}$ se utilizaron tal como fueron propuestos originalmente [26], con tasa de aprendizaje inicial $\eta = 0.001$. La ventaja de Adam sobre el descenso de gradiente estocástico simple es que su tasa de actualización es, en la práctica, invariante ante reescalamientos del gradiente, lo que acelera la convergencia y reduce la necesidad de ajustar manualmente la tasa de aprendizaje.

La exploración inicial de hiperparámetros se realizó mediante RandomizedSearchCV con 96 configuraciones, evaluadas exhaustivamente y validación cruzada de 3 particiones, generando 288 ajustes en total [23]. El espacio de búsqueda incluyó la arquitectura de capas ocultas en {(50,50,50), (50,100,50)}; la función de activación en {identity, logistic, tanh, relu}; el optimizador en {lbfgs, sgd, adam}; la regularización $\alpha \in [0.0001, 1000]$; y la estrategia de tasa de aprendizaje en {constant, adaptive}. La configuración óptima preliminar identificada, $\alpha = 10$, (50,100,50), adaptive, sgd, produjo inestabilidad numérica durante el entrenamiento, manifestada como desbordamiento aritmético en la propagación hacia adelante, documentado en la salida de la celda del cuaderno [8]. La configuración estabilizada final, instanciada en el cuaderno como:

```
# MLPRegressor: capas (50, 50), relu, adam, alpha=0.001, early_stopping=True
```

resolvió esta inestabilidad sin sacrificar capacidad predictiva. Los 3,101 parámetros de la arquitectura $9 \to 50 \to 50 \to 1$ se distribuyen como: $9 \times 50 + 50 = 500$ parámetros en la primera capa oculta, $50 \times 50 + 50 = 2{,}550$ en la segunda, y $50 \times 1 + 1 = 51$ en la capa de salida. La parada temprana interrumpe el entrenamiento cuando la función de pérdida en un subconjunto de validación interno no mejora durante 20 iteraciones consecutivas, actuando como mecanismo adicional de regularización implícita.

2.5.4 Bosque Aleatorio (RandomForestRegressor)

El Bosque Aleatorio (*Random Forest*, RF) es un método de aprendizaje de conjunto (*ensemble learning*) que construye un número B de árboles de decisión durante el entrenamiento y produce como predicción el promedio de las predicciones individuales de todos los árboles [18]:

$$\hat{y}_{RF}(\mathbf{x}) = \underbrace{\frac{1}{B}\sum_{b=1}^{B}}_{\text{promedio}} \underbrace{T_b(\mathbf{x})}_{\text{árbol } b} \tag{2.16}$$

donde $T_b(\mathbf{x})$ es la predicción del árbol b para la entrada $\mathbf{x}$. La justificación estadística de esta agregación se encuentra en la descomposición sesgo-varianza del error cuadrático esperado de un predictor $\hat{f}(\mathbf{x})$:

$$\mathbb{E}\left[(y - \hat{f}(\mathbf{x}))^2\right] = \underbrace{\left(\mathbb{E}[\hat{f}(\mathbf{x})] - f(\mathbf{x})\right)^2}_{\text{sesgo}^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}(\mathbf{x}) - \mathbb{E}[\hat{f}(\mathbf{x})]\right)^2\right]}_{\text{varianza}} + \underbrace{\sigma_\varepsilon^2}_{\text{ruido irreducible}} \tag{2.17}$$

Un árbol de decisión individual no restringido tiene sesgo bajo pero varianza muy alta: pequeñas perturbaciones en los datos de entrenamiento producen árboles estructuralmente distintos con predicciones muy diferentes. Breiman demostró [18] que si los B árboles son idénticamente distribuidos con varianza σ^2 y correlación por pares ρ, la varianza del promedio es:

$$\mathrm{Var}(\hat{y}_{RF}) = \underbrace{\rho\,\sigma^2}_{\substack{\text{no reducible} \\ \text{con } B}} + \underbrace{\frac{1-\rho}{B}\,\sigma^2}_{\substack{\text{reducible al} \\ \text{aumentar } B}} \tag{2.18}$$

El primer término, proporcional a ρ, no se reduce con B; el segundo, proporcional a $(1-\rho)/B$, desaparece a medida que $B \to \infty$. La estrategia del Random Forest está diseñada precisamente para minimizar ρ, la correlación entre árboles, mediante dos mecanismos de aleatorización complementarios, sin aumentar el sesgo. El primero es el muestreo con reemplazo (*bagging* o *bootstrap aggregation*): cada árbol T_b se entrena sobre una muestra aleatoria con reemplazo de tamaño n extraída del conjunto de entrenamiento, de modo que distintos árboles ven distintos

subconjuntos de los datos. El segundo es la aleatorización de características: en cada nodo de división, el algoritmo selecciona aleatoriamente un subconjunto de $m \ll p$ variables candidatas, con $m = \lfloor \sqrt{p} \rfloor$ en la configuración óptima encontrada, en lugar de evaluar todas las p variables, lo que reduce la probabilidad de que distintos árboles elijan la misma variable dominante en sus nodos raíz, decorrelacionando activamente las predicciones.

Cada nodo de un árbol individual se divide seleccionando la variable X_j y el umbral τ que maximizan la reducción de impureza respecto al nodo padre t:

$$\Delta I(s,t) = \underbrace{I(t)}_{\substack{\text{impureza}\\ \text{nodo padre}}} - \underbrace{\frac{|t_L|}{|t|} I(t_L)}_{\substack{\text{hijo}\\ \text{izquierdo}}} - \underbrace{\frac{|t_R|}{|t|} I(t_R)}_{\substack{\text{hijo}\\ \text{derecho}}} \tag{2.19}$$

donde t_L y t_R son los nodos hijos resultantes de la partición $s : X_j \le \tau$, y la impureza de un nodo se mide mediante la varianza muestral de los valores de la variable objetivo dentro de ese nodo:

$$\underbrace{I(t)}_{\text{impureza}} = \underbrace{\frac{1}{|t|} \sum_{i \in t} (y_i - \bar{y}_t)^2}_{\text{varianza del nodo}}, \qquad \bar{y}_t = \underbrace{\frac{1}{|t|} \sum_{i \in t} y_i}_{\text{media del nodo}} \tag{2.20}$$

La predicción del árbol en una hoja ℓ es $\hat{y} = \bar{y}_\ell$, la media de los valores de entrenamiento que caen en esa hoja. La división óptima (X_j^*, τ^*) se encuentra buscando exhaustivamente sobre las m variables candidatas y todos los umbrales posibles entre observaciones contiguas de cada variable.

A partir de las reducciones de impureza acumuladas durante el entrenamiento de todos los árboles, se deriva la importancia de cada variable predictora X_j, formalizada por Louppe et al. [27] como la contribución media ponderada por el tamaño del nodo de todas las divisiones donde X_j fue la variable seleccionada:

$$\mathrm{FI}(X_j) = \underbrace{\frac{1}{B} \sum_{b=1}^{B}}_{\substack{\text{promedio}\\ \text{sobre árboles}}} \sum_{\substack{t \in T_b \\ v(s_t) = X_j}} \underbrace{\frac{|t|}{n}}_{\substack{\text{peso por}\\ \text{tamaño}}} \underbrace{\Delta I(s_t, t)}_{\substack{\text{reducción de}\\ \text{impureza}}} \tag{2.21}$$

donde $v(s_t)$ denota la variable utilizada en la división del nodo t del árbol b. Esta métrica se normaliza para que $\sum_j \mathrm{FI}(X_j) = 1$, lo que permite interpretar las importancias como fracciones de la varianza total explicada que pueden atribuirse a cada variable. Es la misma métrica, ec. (2.21), reportada en la sección 2.4.2 y es la que produjo el perfil de importancias con Year (0.5044) y Time (0.4576) como variables dominantes. Louppe et al. [27] demostraron además que, bajo condiciones de asimetría en el tamaño de los nodos, esta estimación puede estar sesgada hacia

variables de alta cardinalidad, aunque en el contexto de este trabajo, con variables temporales discretas de cardinalidad moderada, este efecto no compromete la interpretación cualitativa del perfil.

La combinación de *bagging* y aleatorización de variables produce árboles estructuralmente distintos entre sí, cuyo promedio como predictor colectivo tiene menor varianza que cualquier árbol individual sin sacrificar el sesgo del modelo [18].

Esta arquitectura confiere al Random Forest propiedades especialmente valiosas para el tipo de problema abordado. Es robusto ante la presencia de variables predictoras irrelevantes o ruidosas, captura de forma natural interacciones no lineales complejas entre variables sin requerir su especificación explícita, y produce estimaciones de importancia de variable robustas y relativamente estables. La literatura sobre pronóstico fotovoltaico ha identificado consistentemente al Random Forest como uno de los algoritmos de mayor rendimiento sobre datos tabulares con variables meteorológicas como predictores, tanto en emplazamientos de clima templado [11][28] como en contextos de alta insolación árida comparables al de Puerto Peñasco [4][12][29]. La metodología de Pasion et al. [4], en particular, es directamente comparable a la adoptada en este trabajo: ambos enfoques emplean variables meteorológicas y geográficas como predictores, omiten la irradiancia directa y utilizan Random Forest como modelo de referencia. Aravena-Cifuentes et al. [6], quienes emplearon exactamente el mismo dataset, confirmaron experimentalmente que el desempeño del modelo mejora de forma consistente al incrementar el número de estimadores, con RF 1-500 superando a RF 1-100 y RF 1-50, aunque a costa de un mayor costo computacional y emisiones de carbono asociadas; resultado coherente con la configuración de `n_estimators=554` identificada en la búsqueda de hiperparámetros del presente trabajo.

La búsqueda de hiperparámetros se realizó mediante RandomizedSearchCV con 500 iteraciones y validación cruzada de 3 particiones [23]. El espacio de búsqueda incluyó: número de árboles (`n_estimators`) entre 100 y 1,000; profundidad máxima (`max_depth`) entre 10 y 110, más el valor `None` (árbol sin restricción de profundidad); fracción de variables candidatas por nodo (`max_features`) en {`'auto'`, `'sqrt'`}; mínimo de muestras para dividir un nodo (`min_samples_split`) en {2, 5, 10}; mínimo de muestras en hojas (`min_samples_leaf`) en {1, 2, 4}; y uso o no de submuestreo con reemplazo (`bootstrap`) en {True, False}. El proceso de búsqueda requirió un tiempo de cómputo de aproximadamente 4 horas y 27 minutos, registrado en la celda de búsqueda del cuaderno [8]. La configuración óptima fue:

```
# RandomForestRegressor: n_estimators=554, max_features='sqrt', bootstrap=False
```

Dos aspectos de esta configuración merecen interpretación en relación con las ecuaciones de la sección anterior. Primero, el valor `bootstrap=False` implica que cada árbol T_b se entrena sobre el conjunto de entrenamiento completo en lugar de una muestra con reemplazo; en este caso la diversidad entre los árboles proviene exclusivamente de la aleatorización de variables

en cada nodo, el componente $\rho\sigma^2$ de la ec. (2.18), mientras que el segundo término $(1-\rho)\sigma^2/B$ se reduce libremente con B. Que el espacio de búsqueda seleccione esta opción sugiere que, con 1,625 registros de entrenamiento y 9 características, el *bagging* no aporta ganancia marginal de diversidad suficiente para compensar la ligera pérdida de datos por muestreo. Segundo, el valor $B = 554$ es consistente con la evidencia de Aravena-Cifuentes et al. [6] de que el desempeño mejora con más estimadores, y su magnitud justifica el elevado tiempo de cómputo de la búsqueda.

2.6 Protocolo de Evaluación y Métricas de Desempeño

La descripción de los modelos completa el diseño del experimento comparativo, pero no establece aún las reglas por las que se juzgará su desempeño. Esta sección documenta el protocolo de evaluación y define con precisión las métricas utilizadas, garantizando que las comparaciones del Capítulo III tengan un fundamento metodológico sólido y reproducible.

2.6.1 Fundamento del protocolo de evaluación sobre conjunto de prueba independiente

La evaluación de los modelos sobre un conjunto de prueba que no ha participado en ninguna fase del entrenamiento ni de la selección de hiperparámetros es la garantía fundamental de que las métricas reportadas reflejan la capacidad de generalización de cada modelo y no simplemente su capacidad de memorizar los datos de entrenamiento. Este principio es especialmente importante cuando se comparan múltiples arquitecturas, porque el proceso de búsqueda de hiperparámetros puede producir una sobreestimación del desempeño si el conjunto de prueba se involucra en la selección, fenómeno conocido como *data leakage* [10]. En el protocolo adoptado, el conjunto de prueba de 391 registros correspondiente al 20% final del dataset fue reservado desde el inicio del proceso de diseño y no fue examinado en ninguna fase del análisis exploratorio o del preprocesamiento que pudiera haber introducido sesgos. Las métricas reportadas en el Capítulo III corresponden exclusivamente a la evaluación de los modelos entrenados y configurados sobre el conjunto de entrenamiento cuando se aplican a ese conjunto de prueba independiente.

2.6.2 Métricas de evaluación

El desempeño de cada modelo se cuantifica mediante tres métricas complementarias que capturan aspectos distintos de la calidad de las predicciones. Sea $\{(y_i, \hat{y}_i)\}_{i=1}^{n}$ el conjunto de pares de valores observados y predichos sobre el conjunto de prueba, con $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

El Error Absoluto Medio (MAE, *Mean Absolute Error*) se define como:

$$\text{MAE} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}}_{\text{promedio}} \underbrace{|y_i - \hat{y}_i|}_{\text{error absoluto}} \tag{2.22}$$

El MAE es la norma ℓ_1 del vector de errores dividida entre n, lo que le confiere robustez ante errores atípicos grandes: cada error contribuye al MAE con peso proporcional a su magnitud absoluta, sin amplificación cuadrática. Willmott y Matsuura [30] argumentaron que el MAE es la métrica de error promedio más natural e inequívoca para evaluar el rendimiento promedio de modelos de pronóstico, precisamente porque no introduce ningún sesgo implícito hacia penalizar errores grandes respecto a errores pequeños. En el contexto de este trabajo, el MAE proporciona una estimación directamente interpretable del error operativo típico: un MAE de 20 MW indica que en promedio el modelo se equivoca en 20 MW respecto al valor real, lo que sobre una variable con máximo registrado de 812 MW representa un error relativo de aproximadamente 2.5%.

El Error Cuadrático Medio (RMSE, *Root Mean Squared Error*) se define como:

$$\text{RMSE} = \sqrt{\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{error cuadrático medio}}} \tag{2.23}$$

El RMSE es la raíz cuadrada de la norma ℓ_2 al cuadrado del vector de errores dividida entre n. Su relación con el MAE, ecs. (2.22) y (2.23), revela su propiedad definitoria: por la desigualdad de Jensen aplicada a la función convexa $f(x) = x^2$, se tiene siempre $\text{MAE} \leq \text{RMSE}$, con igualdad sólo cuando todos los errores tienen la misma magnitud absoluta. La brecha $\text{RMSE} - \text{MAE}$ es por tanto indicativa de la heterogeneidad en la distribución de los errores: cuanto mayor sea esa diferencia, mayor es la contribución de errores grandes ocasionales al RMSE respecto al error promedio [30]. Esta propiedad hace del RMSE una métrica complementaria al MAE especialmente relevante en contextos operativos como el despacho programado de generación: una sobreestimación puntual y pronunciada de la potencia disponible puede generar desequilibrios en el balance de la red cuyo costo de corrección supera con creces el de una serie de pequeños errores de igual magnitud acumulada [15]. Comparar MAE y RMSE conjuntamente proporciona así una caracterización más completa de la distribución de errores que cualquiera de las dos métricas por separado.

El Coeficiente de Determinación (R^2) mide la fracción de la varianza total de la variable objetivo que es explicada por el modelo respecto a la varianza explicada por el predictor constante $\hat{y} = \bar{y}$:

$$R^2 = 1 - \underbrace{\frac{SS_{\text{res}}}{SS_{\text{tot}}}}_{\substack{\text{fracción de varianza} \\ \text{no explicada}}} = 1 - \frac{\underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SS_{\text{res}}:\,\text{varianza residual}}}{\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_{\text{tot}}:\,\text{varianza total}}} \tag{2.24}$$

donde $SS_{\text{res}} = \sum_i (y_i - \hat{y}_i)^2$ es la suma de cuadrados residual y $SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$ es la varianza total de los datos. Nótese que $SS_{\text{tot}} = n \cdot \hat{\sigma}_y^2$, por lo que el R^2 puede interpretarse como $1 - \hat{\sigma}_\epsilon^2 / \hat{\sigma}_y^2$:

la fracción de la varianza de la variable objetivo que no es capturada por el modelo. Su valor varía entre $-\infty$ y 1; un valor de 1 indica predicción perfecta ($SS_{\text{res}} = 0$), un valor de 0 indica que el modelo no aporta información más allá de la media ($SS_{\text{res}} = SS_{\text{tot}}$, equivalente a la predicción constante de la sección 2.5.1, ec. (2.3)), y valores negativos indican que el modelo es peor predictor que la media. La ventaja del R^2 como métrica de comparación reside en su adimensionalidad: a diferencia del MAE y del RMSE, que están en las unidades de la variable objetivo, proporciona una referencia normalizada que facilita la contextualización de los resultados de Puerto Peñasco respecto a valores de la literatura internacional, donde, como señalaron Pasion et al. [4] y Moufid et al. [29], los mejores modelos basados en Random Forest sobre datos meteorológicos alcanzan $R^2 > 0.94$ en emplazamientos de alta insolación [15].

Las tres métricas se calculan en el cuaderno [8] para cada modelo sobre el conjunto de prueba independiente mediante las funciones de scikit-learn [13] y NumPy:

```
# Evaluación sobre n=391: MAE [MW], RMSE [MW], R² [adimensional]
```

Estas funciones implementan exactamente las fórmulas definidas en esta sección: `mean_absolute_error` computa $\frac{1}{n}\sum|y_i - \hat{y}_i|$, `mean_squared_error` computa $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$ (de donde se extrae la raíz), y `r2_score` computa $1 - SS_{\text{res}}/SS_{\text{tot}}$ con la media calculada sobre `y_test`. Los valores numéricos reportados en el Capítulo III son los producidos directamente por estas celdas de evaluación, garantizando la trazabilidad completa entre la formulación matemática y el resultado experimental.

2.6.3 Curvas de aprendizaje y diagnóstico

Las curvas de aprendizaje representan el desempeño del modelo, medido mediante métricas con validación cruzada, en función del número de registros de entrenamiento utilizados. Su utilidad diagnóstica radica en que revelan si el comportamiento del modelo es dominado por el sobreajuste, en cuyo caso el desempeño en entrenamiento es excelente pero el de validación es significativamente peor, o por el subajuste, en cuyo caso ambas curvas convergen a un valor de desempeño insatisfactorio independientemente del tamaño del conjunto de entrenamiento [10]. Para cada uno de los cuatro modelos implementados se generaron curvas de aprendizaje evaluadas sobre las tres métricas de desempeño, junto con curvas de validación para el análisis de la sensibilidad del desempeño ante variaciones en el hiperparámetro más relevante de cada arquitectura, y gráficas de residuos para el diagnóstico del patrón de errores. El comportamiento de cada curva, y en particular la distancia entre las curvas de entrenamiento y de validación, aporta información complementaria a las métricas de desempeño sobre el conjunto de prueba: un modelo que exhibe una brecha grande entre entrenamiento y validación podría verse afectado en escenarios con datos de entrenamiento más escasos, lo que tiene implicaciones directas para la replicabilidad de la metodología en otras centrales con historiales operativos más cortos.

2.7 El Rol Potencial de los LLMs como Capa de Interpretación Operativa

El pipeline metodológico descrito en las secciones anteriores produce, como resultado, modelos entrenados cuyos outputs son números: valores predichos de PolyPwr en megawatts para cada hora futura, métricas de desempeño y estimaciones de importancia de variable. Estos outputs son altamente informativos para un especialista en ciencia de datos o en ingeniería de sistemas de potencia, pero no están en un formato directamente accesible para todos los perfiles de usuario que participan en la operación diaria de una central fotovoltaica.

Es en este punto donde los modelos de lenguaje de gran escala (LLMs) ofrecen una capacidad cualitativamente distinta a la de los modelos predictivos especializados. Como señalaron Kim et al. en el reporte técnico del NREL [31], la función central de los LLMs en el contexto operativo no es generar predicciones numéricas de alta precisión, sino transformar información estructurada: tablas de métricas, series temporales de predicciones, indicadores de desempeño; en comunicación legible, contextualizada y accionable para usuarios con distintos perfiles de formación. Liu et al. documentaron que la incorporación de herramientas de IA generativa en flujos de trabajo de análisis de sistemas solares puede mejorar significativamente la velocidad de acceso a la información y la calidad de las decisiones operativas derivadas [32].

Esta capacidad de mediación tiene tres expresiones concretas particularmente relevantes para una instalación como Puerto Peñasco. La primera es la generación automática de reportes de turno: al finalizar cada turno de operación, un LLM configurado como interfaz sobre los outputs del modelo predictivo podría generar automáticamente un resumen que incluya la generación total del turno comparada con el pronóstico, las horas en las que se produjeron las desviaciones más significativas y una descripción en lenguaje natural de los factores meteorológicos que las explican. La segunda es la asistencia a consultas en lenguaje natural sobre el historial operativo: un operador podría formular preguntas sobre el comportamiento histórico del sistema y obtener respuestas estructuradas sin necesidad de ejecutar consultas manuales sobre bases de datos. La tercera es la síntesis de alertas ante desviaciones del modelo: cuando la generación observada se desvía de forma significativa y persistente respecto a la generación predicha, un LLM con acceso a los registros históricos podría formular una hipótesis inicial sobre la naturaleza de la desviación y presentarla al operador en términos claros, acelerando el proceso de diagnóstico sin reemplazar el juicio técnico del especialista.

Es fundamental, sin embargo, delimitar con precisión los límites de estas aplicaciones. Los LLMs no tienen capacidad intrínseca para predecir valores numéricos con la precisión requerida para el despacho programado, ni para detectar fallas en equipos a partir de señales de telemetría sin el soporte de algoritmos especializados de detección de anomalías. Su valor está en la capa de comunicación e interpretación, no en la capa de predicción cuantitativa. El enfoque metodológicamente maduro que esta tesis propone como línea de desarrollo futuro es el de

sistemas integrados en los que un modelo de ML especializado produce las predicciones numéricas de alta precisión y un LLM actúa como interfaz conversacional sobre esos resultados, amplificando su utilidad operativa sin pretender sustituirlos [31][32].

2.8 Conclusiones del Capítulo II

Este capítulo ha documentado con precisión el pipeline metodológico completo que sustenta el trabajo empírico de esta investigación, desde la descripción del caso de estudio y el conjunto de datos hasta la configuración de los cuatro modelos predictivos y el protocolo de evaluación sobre el cual se comparan sus resultados. Cada decisión metodológica adoptada a lo largo de este proceso, la partición cronológica del conjunto de datos, la codificación *one-hot* de la variable de estación, la estandarización de variables mediante la sección 2.3.3, ecs. (**??**)–(2.2), la selección de características mediante RFECV, la elección de las arquitecturas de modelos en orden creciente de complejidad y las métricas de evaluación, responde a principios bien establecidos en la literatura de aprendizaje automático aplicado a series temporales energéticas [10][14][15], adaptados a las características específicas del contexto operativo de Puerto Peñasco.

El conjunto de datos sigue el enfoque propuesto originalmente por Pasion et al. [4], extendido con la incorporación de variables temporales adicionales y aplicado al contexto de una central de gran escala en México. La selección de variables meteorológicas y geográficas como predictores, prescindiendo de la irradiancia directa, que no siempre está disponible con la cobertura y calidad requeridas, es una decisión metodológica justificada por la evidencia de que Random Forest puede alcanzar R^2 superiores a 0.94 con este conjunto de variables en entornos de alta insolación [4][29]. La conveniencia de diseñar modelos específicos para cada emplazamiento, en lugar de recurrir a modelos genéricos multi-sitio, queda además respaldada por Aravena-Cifuentes et al. [6], cuya replicación y extensión independiente del mismo dataset de referencia demostró que los modelos entrenados sin contexto del emplazamiento objetivo producen predicciones degradadas, con R^2 negativos en algunos casos, mientras que incorporar aunque sea una semana de datos de la ubicación específica mejora sustancialmente el desempeño. Este hallazgo convierte la disponibilidad del historial operativo de Puerto Peñasco en una ventaja metodológica directa del presente trabajo.

El análisis exploratorio reveló que PolyPwr exhibe una distribución bimodal con una desviación estándar que supera a la propia media, lo que establece un punto de referencia exigente para los modelos y hace particularmente informativa la comparación con el regresor ficticio. El proceso de RFECV, con un R^2 de validación cruzada de 0.9991 sobre el conjunto de características seleccionadas, identificó que nueve variables, con un peso dominante de las variables temporales Year y Time (96.2% de la importancia total conjunta), constituyen el subconjunto óptimo de predictores: resultado físicamente coherente con el comportamiento del recurso solar en un emplazamiento de alta insolación árida. Las cuatro arquitecturas implementadas cubren intencionalmente un es-

pectro de complejidad creciente, regresor ficticio, modelo lineal SGD, red neuronal MLP y bosque aleatorio RF, que permite diagnosticar la estructura del problema y cuantificar la ganancia marginal de cada nivel adicional de sofisticación.

Los resultados cuantitativos del desempeño de cada modelo sobre el conjunto de prueba independiente se presentan y analizan en el Capítulo III, donde se discuten también sus implicaciones operativas concretas para la gestión de la central, se formulan recomendaciones sobre el nivel de complejidad algorítmica apropiado para el contexto de Puerto Peñasco, y se identifican las condiciones de infraestructura y gobernanza de datos necesarias para escalar este tipo de soluciones a otras instalaciones del parque fotovoltaico nacional.

BIBLIOGRAPHY

[1] C. A. Arancibia-Bulnes, R. Peón-Anaya, D. Riveros-Rosas, J. J. Quiñones, R. E. Cabanillas, y C. A. Estrada, "Beam solar irradiation assessment for Sonora, Mexico," *Energy Procedia*, vol. 49, pp. 2290–2296, 2014, doi: 10.1016/j.egypro.2014.03.242.

[2] L. A. Torres-Hernandez, J. I. Matus-Pérez, A. Guerrero-Osuna, R. Bernal-Ballesteros, L. F. Luque-Vega, y M. A. Ríos, "Estimation of Solar Resource Based on Meteorological and Geographical Data: Sonora State in Northwestern Territory of Mexico as Case Study," *Energies*, vol. 13, no. 24, p. 6501, 2020, doi: 10.3390/en13246501.

[3] E. Skoplaki y J. A. Palyvos, "On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations," *Solar Energy*, vol. 83, no. 5, pp. 614–624, May 2009, doi: 10.1016/j.solener.2008.10.008.

[4] C. Pasion, T. Wagner, C. Koschnick, S. Schuldt, J. Williams, y K. Hallinan, "Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data," *Energies*, vol. 13, no. 10, p. 2570, 2020, doi: 10.3390/en13102570.

[5] S. T. Asiedu, S. Gyamfi, E. A. Ofosu, y E. Djordjevic, "Machine learning forecasting of solar PV production using single and hybrid models over different time horizons," *Heliyon*, vol. 10, p. e28898, 2024, doi: 10.1016/j.heliyon.2024.e28898.

[6] A. P. Aravena-Cifuentes, J. D. Nuñez-Gonzalez, A. Elola, y M. Ivanova, "Development of AI-Based Tools for Power Generation Prediction," *Computation*, vol. 11, no. 11, p. 232, 2023, doi: 10.3390/computation11110232.

[7] M. Sánchez Gutiérrez, "Conjunto de datos horarios de generación fotovoltaica y variables meteorológicas — Central Fotovoltaica de Puerto Peñasco," registros operativos verificados y bases de datos meteorológicas de acceso abierto, período diciembre 2023 – agosto 2025. [Reporte Propio 2025-BUAP-MSEP-R01].

[8] M. Sánchez Gutiérrez, "Pipeline de análisis exploratorio, preprocesamiento, selección de características y modelado predictivo para la Central Fotovoltaica de Puerto Peñasco," libreta Jupyter de autoría propia, 2025. [Reporte Propio 2025-BUAP-MSEP-R02].

[9] N. N. Barakat, A. K. Eltamaly, A. M. Alharthi, y M. S. Al-Saud, "Multi-label machine learning for power forecasting of a grid-connected photovoltaic solar plant over multiple time horizons," *Scientific Reports*, vol. 15, p. 10213, 2025, doi: 10.1038/s41598-025-20251-y.

[10] T. Hastie, R. Tibshirani, y J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2ª ed., Springer, Nueva York, USA, 2009.

[11] D. Markovics y M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," *Renewable and Sustainable Energy Reviews*, vol. 161, p. 112364, 2022, doi: 10.1016/j.rser.2022.112364.

[12] A. Aziz, N. Mufti, H. M. Khalid, H. Farooq, S. H. Abbasi, y Q. Khan, "Short-term and long-term solar irradiance forecasting with advanced machine learning techniques in Zafarana, Egypt," *Scientific Reports*, vol. 15, p. 3891, 2025, doi: 10.1038/s41598-025-24853-4.

[13] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [En línea]. Disponible: `https://jmlr.org/papers/v12/pedregosa11a.html`

[14] C. Bergmeir y J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, May 2012, doi: 10.1016/j.ins.2011.12.028.

[15] S. Makridakis, E. Spiliotis, y V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLOS ONE*, vol. 13, no. 3, p. e0194889, 2018, doi: 10.1371/journal.pone.0194889.

[16] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," en *Proceedings of COMPSTAT 2010*, Physica-Verlag HD, 2010, pp. 177–186.

[17] I. Guyon y A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[18] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[19] P. Geurts, D. Ernst, y L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[20] S. Shalev-Shwartz y S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Nueva York, USA, 2014.

[21] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, y V. Vapnik, "Support Vector Regression Machines," en *Advances in Neural Information Processing Systems (NIPS)*, vol. 9, M. Mozer, M. Jordan, y T. Petsche, Eds., MIT Press, Cambridge, MA, USA, 1997, pp. 155–161.

[22] H. Zou y T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[23] J. Bergstra y Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, Feb. 2012. [En línea]. Disponible: `https://jmlr.org/papers/v13/bergstra12a.html`

[24] D. E. Rumelhart, G. E. Hinton, y R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986, doi: 10.1038/323533a0.

[25] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991, doi: 10.1016/0893-6080(91)90009-T.

[26] D. P. Kingma y J. Ba, "Adam: A Method for Stochastic Optimization," en *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. [En línea]. Disponible: `https://arxiv.org/abs/1412.6980`

[27] G. Louppe, L. Wehenkel, A. Sutera, y P. Geurts, "Understanding variable importances in forests of randomized trees," en *Advances in Neural Information Processing Systems (NIPS)*, vol. 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, y K. Q. Weinberger, Eds., Curran Associates, Inc., 2013, pp. 431–439.

[28] V. Raj, S. Dotse, M. Sathyajith, M. Petra, y H. Yassin, "Ensemble machine learning for predicting the power output from different solar photovoltaic systems," *Energies*, vol. 16, no. 2, p. 671, 2023, doi: 10.3390/en16020671.

[29] E. Moufid, H. E. Khouya, y A. Berrahal, "Machine learning-based short-term solar power forecasting: a comparison between regression and classification approaches using extensive Australian dataset," *Sustainable Energy Research*, vol. 11, no. 1, p. 29, 2024, doi: 10.1186/s40807-024-00115-1.

[30] C. J. Willmott y K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, pp. 79–82, 2005, doi: 10.3354/cr030079.

[31] T. Kim, H. Li, R. Bhatt, y B. Palmintier, *Generative AI for Power Grid Operations*, National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep. NREL/TP-5D00-91176, 2024.

[32] Z. Liu, D. Che, C. Gao, y B. Huang, “Revolutionizing solar energy resources: The central role of generative AI in elevating system sustainability and efficiency,” *Applied Energy*, vol. 381, p. 125079, 2025, doi: 10.1016/j.apenergy.2025.125079.

CHAPTER III

Resultados, Análisis Comparativo y Perspectivas de Implementación: Del Modelo Predictivo a la Operación Fotovoltaica Inteligente en Puerto Peñasco

El Capítulo II documentó con rigor el pipeline metodológico completo que sustenta el trabajo empírico de esta investigación: describió el caso de estudio y el conjunto de datos con el detalle necesario para contextualizar cada decisión analítica, presentó el análisis exploratorio que reveló la estructura bimodal de la variable objetivo y las correlaciones entre predictores, ejecutó el preprocesamiento y la selección formal de características mediante RFECV, configuró las cuatro arquitecturas de modelos en orden creciente de complejidad, regresor ficticio, SGD lineal, perceptrón multicapa y bosque aleatorio, y estableció el protocolo de evaluación y las métricas sobre las cuales se compara su desempeño. Ese capítulo respondió al *cómo* del trabajo analítico; el presente capítulo responde al *qué se obtuvo*, al *qué significa* y al *hacia dónde conduce*.

La distinción entre metodología y resultados no es un formalismo estructural: es una separación epistemológica con consecuencias prácticas. Documentar el pipeline con independencia de los valores numéricos que produce permite que un lector, o un equipo de ingeniería en otra instalación, evalúe la solidez de las decisiones metodológicas antes de examinar los resultados, y que pueda replicar el procedimiento sobre sus propios datos sin quedar condicionado por los valores específicos obtenidos en Puerto Peñasco. Inversamente, presentar los resultados en un capítulo dedicado permite analizarlos con la profundidad que merecen: no solo como métricas aisladas, sino como evidencia que responde a preguntas concretas sobre la viabilidad, la utilidad y los límites de la predicción basada en aprendizaje automático para la operación fotovoltaica a escala industrial.

Las preguntas que estructuran este capítulo son, en esencia, cuatro. La primera es cuantitativa:

¿qué nivel de precisión alcanza cada uno de los cuatro modelos configurados en el Capítulo II cuando se evalúa sobre el conjunto de prueba independiente, compuesto por los registros más recientes que los modelos no han visto durante el entrenamiento? La segunda es diagnóstica: ¿qué revelan las diferencias de desempeño entre los modelos sobre la naturaleza del problema de predicción en Puerto Peñasco, en particular, sobre el grado de no linealidad de las relaciones entre las variables meteorológicas y la potencia de salida, y sobre la suficiencia del conjunto de datos disponible para entrenar modelos de distinta complejidad? La tercera es comparativa: ¿cómo se posicionan los resultados obtenidos en esta investigación respecto a los valores reportados en la literatura internacional sobre pronóstico fotovoltaico con variables meteorológicas, y qué factores del emplazamiento y del conjunto de datos explican las similitudes o discrepancias observadas? La cuarta es prospectiva: ¿qué implicaciones tienen estos resultados para la operación real de la central, qué condiciones de infraestructura y gobernanza de datos serían necesarias para llevar un modelo de este tipo a producción, y qué papel podrían desempeñar los modelos de lenguaje de gran escala como capa de interpretación y comunicación sobre los resultados del pipeline predictivo?

Para responder a estas preguntas, el capítulo se organiza en seis secciones articuladas que siguen una progresión deliberada de lo cuantitativo a lo estratégico. La sección 3.1 presenta los resultados del desempeño predictivo de los cuatro modelos sobre el conjunto de prueba, reportando las tres métricas definidas en el Capítulo II, error absoluto medio (MAE), raíz del error cuadrático medio (RMSE) y coeficiente de determinación (R^2), en una tabla comparativa consolidada que permite visualizar de forma inmediata el ranking de desempeño y las magnitudes absolutas de los errores. La sección 3.2 desarrolla el análisis comparativo propiamente dicho: interpreta las brechas de desempeño entre modelos a la luz de sus fundamentos algorítmicos, examina los patrones de residuos y las curvas de aprendizaje y validación generadas en la implementación [1], y diagnostica el grado de no linealidad del problema y la relación entre la complejidad del modelo y la ganancia predictiva marginal. La sección 3.3 contextualiza los resultados frente a la literatura, comparando los valores de R^2 obtenidos con los reportados por Pasion et al. [2] y por otros estudios internacionales sobre predicción fotovoltaica basada en variables meteorológicas, e identifica los factores que explican el posicionamiento de Puerto Peñasco en ese panorama. La sección 3.4 traduce los resultados técnicos en implicaciones operativas concretas: qué significan los niveles de precisión alcanzados para el despacho programado, el mantenimiento predictivo y la planificación de la generación en una central de la escala de Puerto Peñasco. La sección 3.5 desarrolla las perspectivas de implementación, abordando tanto los lineamientos técnicos para llevar el modelo a un entorno de producción, integración con sistemas SCADA, frecuencia de reentrenamiento, monitoreo de deriva del modelo, como el papel potencial de los LLMs como interfaz conversacional sobre los resultados del pipeline, retomando y extendiendo la discusión iniciada en la sección 2.7 del capítulo anterior. Finalmente, la sección 3.6 delimita con precisión las limitaciones del estudio, alcance temporal y geográfico del conjunto de datos, variables no

incluidas, supuestos de estacionariedad, y formula las líneas de trabajo futuro que se derivan naturalmente de los hallazgos y de las limitaciones identificadas.

Esta estructura responde a una convicción metodológica que ha guiado toda la investigación: los resultados de un modelo predictivo sólo adquieren significado pleno cuando se interpretan en el contexto operativo que los motiva, se comparan con los estándares de la disciplina, se traducen en recomendaciones accionables y se acompañan de un reconocimiento explícito de sus limitaciones. Un valor de R^2 por sí solo es un número; un valor de R^2 acompañado de su diagnóstico, su contexto, sus implicaciones y sus límites es conocimiento útil para la toma de decisiones. Este capítulo aspira a transformar los primeros en los segundos.

Todos los valores numéricos reportados en las secciones siguientes fueron producidos directamente por las celdas de evaluación de la implementación [1] sobre el conjunto de prueba independiente de 391 registros (400 tras la partición cronológica 80/20, menos 9 eliminados por valores faltantes en la variable objetivo), utilizando las funciones de scikit-learn [3] y NumPy que implementan exactamente las formulaciones matemáticas definidas en las secciones 2.6.1 y 2.6.2 del Capítulo II. Esta trazabilidad entre la definición formal de las métricas, su implementación computacional y los valores aquí reportados garantiza la reproducibilidad completa de los resultados.

3.1 Resultados del Desempeño Predictivo

Esta sección presenta los valores numéricos de las tres métricas de evaluación: MAE, RMSE y R^2, obtenidos por cada uno de los cuatro modelos al ser evaluados sobre el conjunto de prueba independiente. La presentación sigue el orden de complejidad creciente establecido en el Capítulo II: primero el regresor ficticio (DummyRegressor), que establece la línea base; luego el modelo lineal por descenso de gradiente estocástico (SGD); a continuación el bosque aleatorio (RandomForestRegressor, RFR); y finalmente el perceptrón multicapa (MLPRegressor, MLP). Los valores se consolidan al final de la sección en una tabla comparativa que constituye el punto de partida para el análisis de la sección 3.2. Para cada modelo se reportan, además de las métricas numéricas, las gráficas de diagnóstico generadas en la implementación: predicciones contra valores reales, diagramas de residuos, gráficas de distribución conjunta y, en los casos pertinentes, validación cruzada K-Fold.

Es importante recordar que el conjunto de prueba fue conformado mediante partición cronológica estricta, de modo que contiene los registros más recientes del periodo diciembre 2023 – agosto 2025. Esta decisión, fundamentada en la sección 2.3.1 del Capítulo II conforme a las recomendaciones de Bergmeir y Benítez [4] y de Makridakis et al. [5] para la evaluación de modelos sobre series temporales, garantiza que las métricas reportadas reflejan el desempeño que los modelos tendrían en un escenario de predicción real, sin contaminación temporal entre los conjuntos de entrenamiento y prueba. Las métricas computadas sobre un conjunto de prueba constituido aleatoriamente podrían producir estimaciones optimistas del desempeño; los valores aquí repor-

tados corresponden al escenario evaluativo más exigente y, por tanto, más representativo de la utilidad operativa de los modelos.

3.1.1 Regresor ficticio (DummyRegressor): línea base

El regresor ficticio, configurado con la estrategia de predicción de la media del conjunto de entrenamiento, obtuvo los siguientes resultados sobre el conjunto de prueba: MAE: 259.601 MW; RMSE: 287.216 MW; $R^2 \approx 0.000$.

Estos valores definen el piso absoluto de desempeño: un modelo que no utiliza ninguna de las variables predictoras y que simplemente predice la media de PolyPwr para cada registro genera un error absoluto promedio de 259.6 MW y un error cuadrático medio de 287.2 MW. El R^2 esencialmente nulo confirma que la predicción constante de la media no explica ninguna fracción de la variabilidad de la variable objetivo, resultado matemáticamente necesario, dado que el R^2 se define precisamente como la fracción de varianza explicada más allá de la media.

Para dimensionar la magnitud de estos errores en el contexto operativo, conviene considerar que la media de PolyPwr en el conjunto de prueba es de 226.0 MW. Un MAE de 259.6 MW supera a la propia media, lo que implica que el error típico del modelo es mayor que el valor típico de la variable que intenta predecir: una condición que sería completamente inaceptable para cualquier aplicación de despacho o planificación. Este resultado no es sorprendente, predecir la media es, por definición, la estrategia más ingenua posible, pero su documentación cuantitativa es metodológicamente indispensable porque establece el denominador contra el cual se mide la ganancia de información de los modelos entrenados. Como señalaron Hastie, Tibshirani y Friedman [6], la inclusión de un modelo de referencia ingenuo permite distinguir con claridad cuánto del desempeño observado en los modelos más sofisticados es genuinamente atribuible a la capacidad del algoritmo y cuánto es simplemente consecuencia de la estructura estadística del conjunto de datos.

La gráfica de predicciones contra valores reales (Figura 3.1) muestra, como es de esperarse, una línea horizontal de predicciones constantes frente a la dispersión de los valores observados, que se extienden desde 0 MW hasta más de 800 MW. El diagrama de residuos (Figura 3.2) revela la estructura completa de la variabilidad de PolyPwr que el modelo no captura: los residuos reproducen casi exactamente la distribución de los valores reales, confirmando que el regresor ficticio no extrae ninguna información del espacio de predictores.

Las curvas de aprendizaje del regresor ficticio (Figuras 3.3–3.5) son planas, como corresponde a un modelo cuyo desempeño no depende del número de ejemplos de entrenamiento: tanto la curva de entrenamiento como la de validación cruzada permanecen constantes e idénticas para las tres métricas (R^2, RMSE negativo y MAE negativo), confirmando que el modelo no aprende ningún patrón adicional al aumentar la cantidad de datos disponibles.

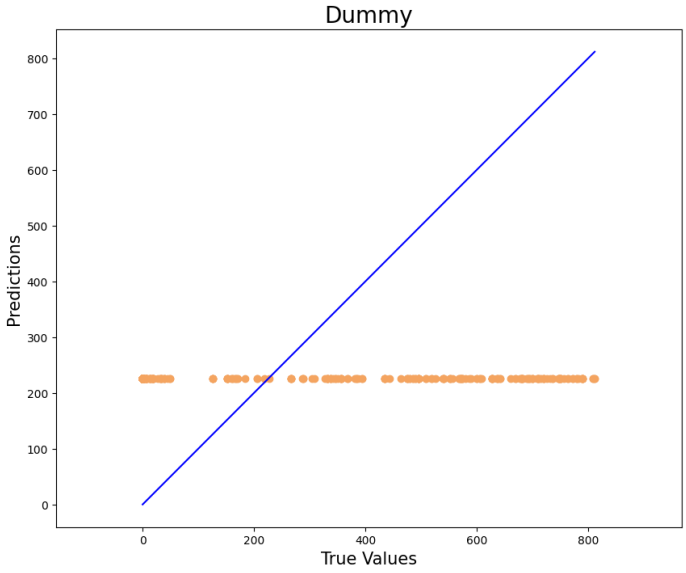


Figure 3.1: Predicciones del regresor ficticio contra valores reales de PolyPwr. La línea horizontal confirma la predicción constante de la media del conjunto de entrenamiento.

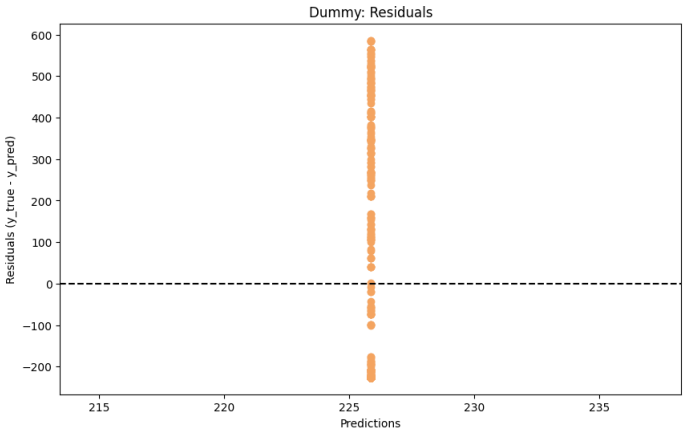


Figure 3.2: Diagrama de residuos del regresor ficticio. El patrón revela la estructura completa de la variabilidad de PolyPwr que el modelo no captura.

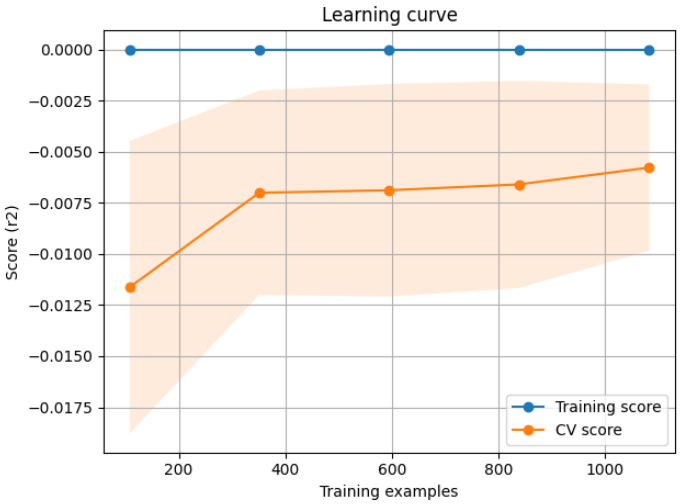


Figure 3.3: Curva de aprendizaje del regresor ficticio para R^2.

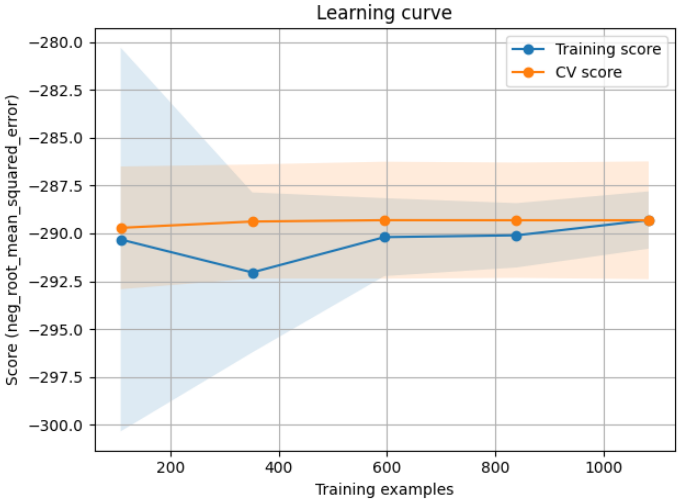


Figure 3.4: Curva de aprendizaje del regresor ficticio para RMSE.

3.1.2 Regresor por descenso de gradiente estocástico (SGDRegressor)

El modelo lineal SGD, configurado con los hiperparámetros óptimos identificados mediante RandomizedSearchCV con 1,500 iteraciones en el Capítulo II (eta0=1, learning_rate='adaptive', loss='squared_epsilon_insensitive', penalty='elasticnet'), obtuvo los siguientes resultados: MAE: 188.086 MW; RMSE: 232.165 MW; R^2: 0.347.

El SGD representa una mejora del 27.5% en MAE y del 19.2% en RMSE respecto a la línea base del regresor ficticio. Estos porcentajes confirman que las variables predictoras contienen información lineal explotable sobre la potencia de salida: la combinación ponderada de las diez características seleccionadas por el RFECV logra reducir el error de predicción en más de una cuarta parte respecto a la predicción ingenua. Sin embargo, un R^2 de 0.347 indica que el modelo lineal solo es capaz de capturar aproximadamente un tercio de la variabilidad total de PolyPwr.

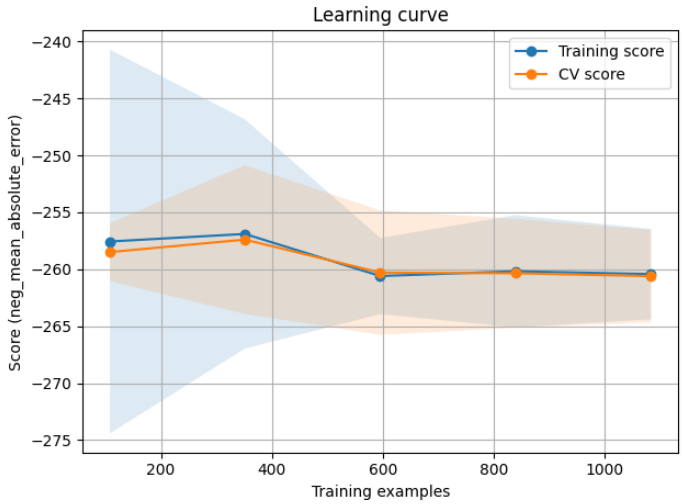


Figure 3.5: Curva de aprendizaje del regresor ficticio para MAE.

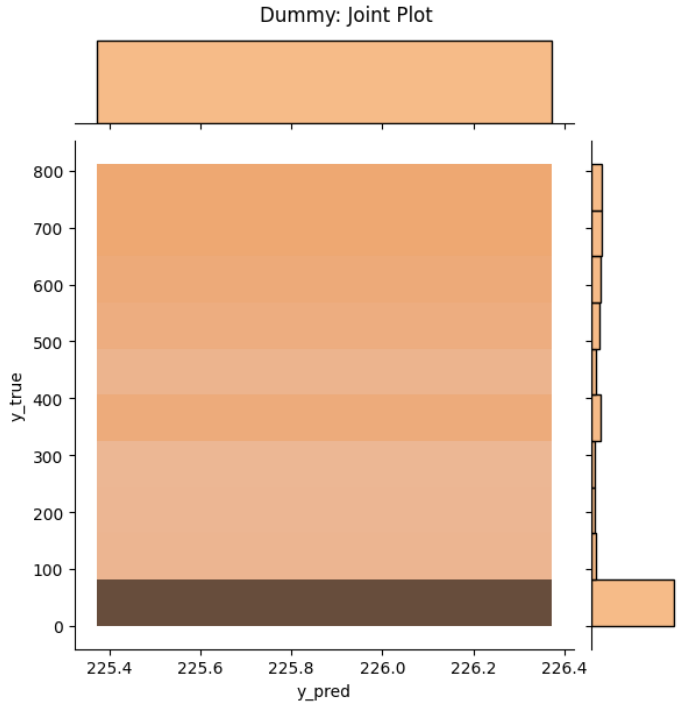


Figure 3.6: Diagrama de distribución conjunta (JointPlot) del regresor ficticio.

Los dos tercios restantes corresponden a relaciones no lineales entre los predictores y la salida que la arquitectura lineal del SGD, por construcción, no puede modelar.

La predicción media del SGD (227.24 MW) es cercana a la media real del conjunto de prueba (226.0 MW), lo que indica que el modelo está correctamente calibrado en términos de tendencia central. El problema no reside en un sesgo global sino en la incapacidad de capturar la dispersión: los registros de alta generación (>500 MW) y los de generación nula o mínima (horas nocturnas y de baja irradiancia) son modelados con errores que frecuentemente superan los 200 MW, magnitud inaceptable para cualquier aplicación operativa.

La gráfica de predicciones contra valores reales (Figura 3.7) proporciona una visualización directa de esta limitación: la nube de puntos presenta una dispersión considerable respecto a la línea de identidad, con desviaciones especialmente pronunciadas en el rango de generación media y alta (por encima de 300 MW), donde el modelo tiende a subestimar sistemáticamente la producción real. En la zona de baja generación, se observa el fenómeno inverso: el modelo sobreestima la potencia, produciendo predicciones positivas para registros donde la generación real es cero o cercana a cero. Este patrón de errores es característico de los modelos lineales aplicados a problemas con estructura no lineal: al estar restringido a un hiperplano en el espacio de características, el modelo "compromete" entre los distintos regímenes de operación, subestimando los extremos y sobreestimando la zona central.

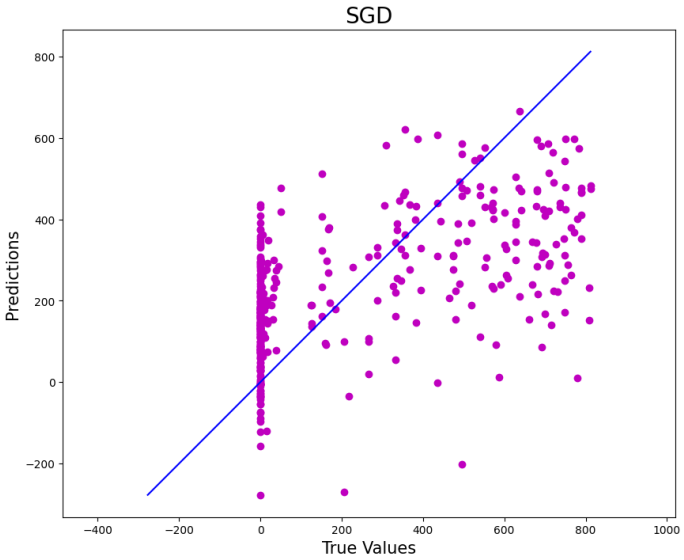


Figure 3.7: Predicciones del SGD contra valores reales de PolyPwr. La dispersión pronunciada revela la incapacidad del modelo lineal para capturar las interacciones no lineales del problema.

El diagrama de residuos (Figura 3.8) confirma el sesgo sistemático del modelo lineal y constituye evidencia gráfica de la no linealidad del problema. Las curvas de aprendizaje del SGD con hiperparámetros optimizados (Figuras 3.10–3.12) muestran un comportamiento particularmente revelador: los valores de R^2 en la curva de aprendizaje alcanzan magnitudes del orden de

10^{15} negativos para ciertos tamaños de entrenamiento, con una varianza muy alta entre las particiones de validación cruzada. Este comportamiento extremo es consistente con la sensibilidad del SGD lineal a la configuración del espacio de parámetros: cuando el número de ejemplos de entrenamiento no es suficiente para estabilizar los coeficientes del modelo, o cuando la regularización elasticnet no logra controlar la magnitud de los pesos, el modelo puede producir predicciones muy alejadas de la escala de la variable objetivo. Es importante señalar que este comportamiento se observa en la curva de aprendizaje (que evalúa el modelo sobre subconjuntos progresivamente mayores del entrenamiento), no en la evaluación final sobre el conjunto de prueba, donde el modelo con el total de los datos de entrenamiento alcanza un R^2 de 0.347. No obstante, la inestabilidad de las curvas de aprendizaje constituye una señal de advertencia sobre la robustez del SGD en escenarios con menor disponibilidad de datos.

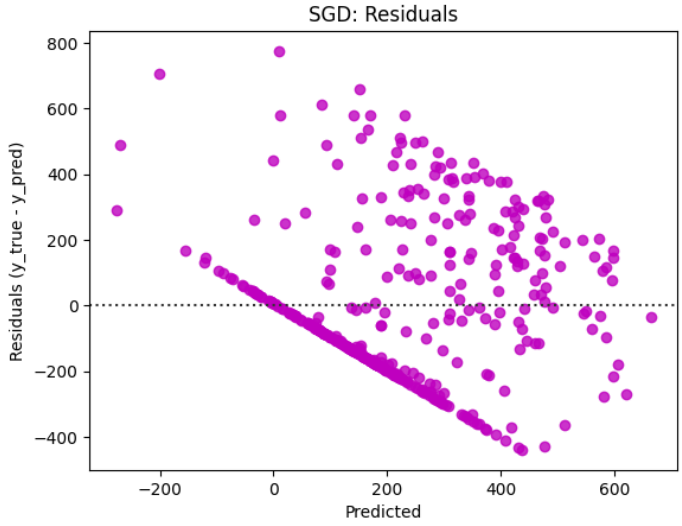


Figure 3.8: Diagrama de residuos del SGD. El patrón fuertemente estructurado y no aleatorio confirma el sesgo sistemático del modelo lineal.

3.1.3 Bosque aleatorio (RandomForestRegressor)

El bosque aleatorio, configurado con los hiperparámetros óptimos identificados mediante RandomizedSearchCV con 96 iteraciones (bootstrap=False, max_depth=30, max_features='sqrt', min_samples_split=10, n_estimators=554), alcanzó un nivel de desempeño cualitativamente distinto al de los dos modelos anteriores: MAE: 15.996 MW; RMSE: 27.224 MW; R^2: 0.991.

El salto en desempeño respecto al SGD es extraordinario y constituye uno de los hallazgos más informativos de este trabajo: el MAE se reduce en un 91.5% (de 188.1 MW a 16.0 MW), el RMSE en un 88.3% (de 232.2 MW a 27.2 MW), y el R^2 pasa de 0.347 a 0.991. Para dimensionar esta magnitud, un error absoluto promedio de 16 MW sobre una variable cuya media es de 226 MW representa un error relativo de aproximadamente 7.1%, cifra que resulta altamente competitiva para aplicaciones de pronóstico horario y que se posiciona en el rango superior de los resultados

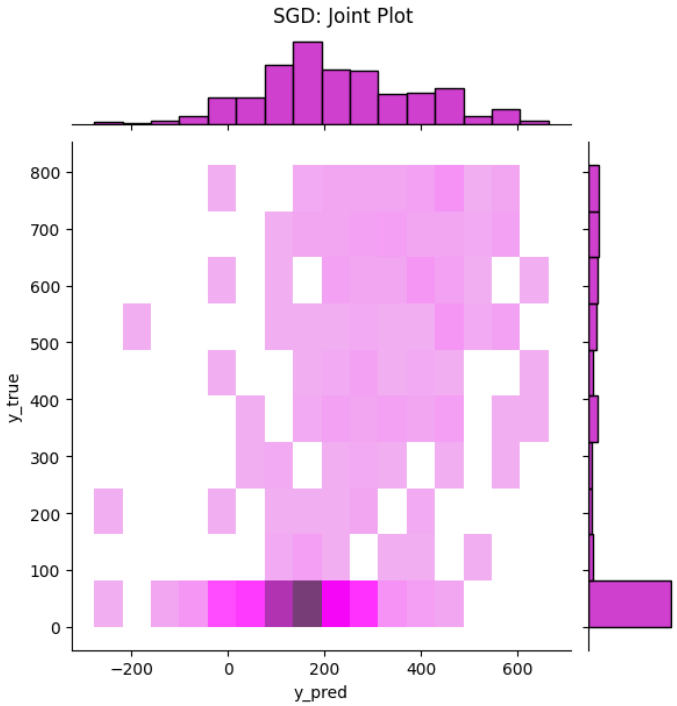


Figure 3.9: Diagrama de distribución conjunta (JointPlot) del SGD.

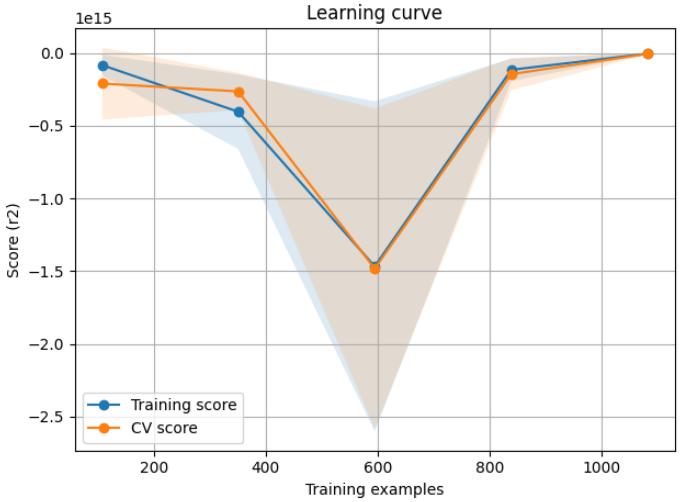


Figure 3.10: Curva de aprendizaje del SGD optimizado para R^2.

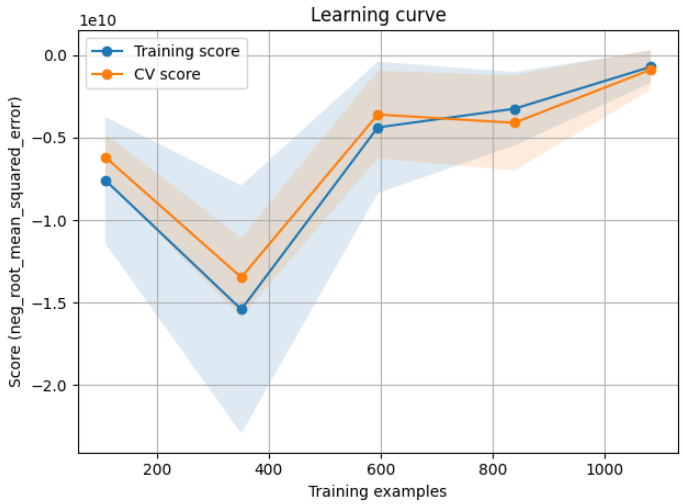


Figure 3.11: Curva de aprendizaje del SGD optimizado para RMSE.

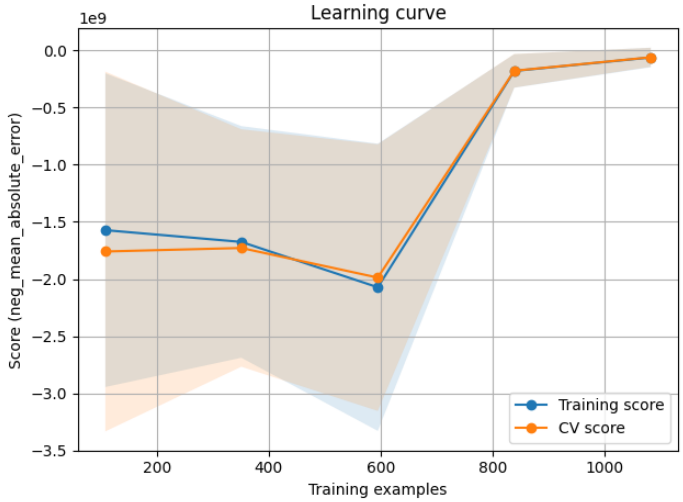


Figure 3.12: Curva de aprendizaje del SGD optimizado para MAE.

reportados en la literatura internacional, como se discutirá en la sección 3.3. La predicción media del RFR (228.49 MW) es cercana a la media real (226.0 MW), con una diferencia de solo 2.49 MW que indica una calibración adecuada del modelo.

La gráfica de predicciones contra valores reales (Figura 3.15) confirma visualmente la alta precisión: los puntos se concentran de forma estrecha alrededor de la línea de identidad a lo largo de todo el rango de generación, desde los registros de baja potencia (0–50 MW) hasta los de máxima generación (700–800 MW). Las desviaciones observadas en la mayoría de los casos no superan los ±50 MW, y se concentran preferentemente en el rango de generación alta (600–800 MW), un patrón que se analizará con mayor detalle en la sección 3.2.4 al discutir la

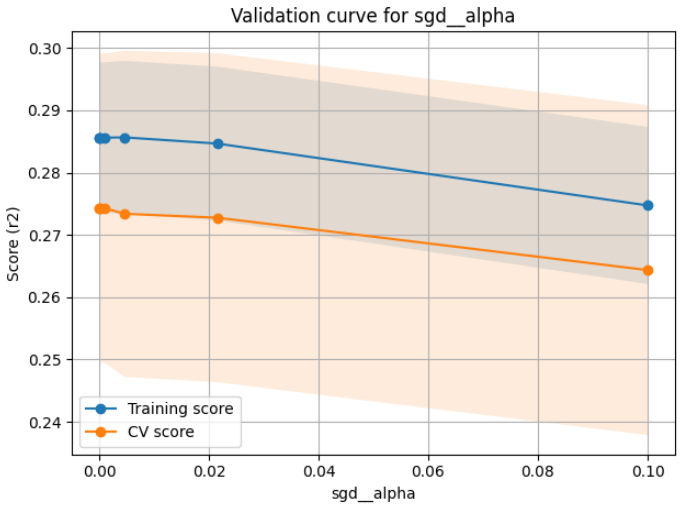


Figure 3.13: Curva de validación del SGD.

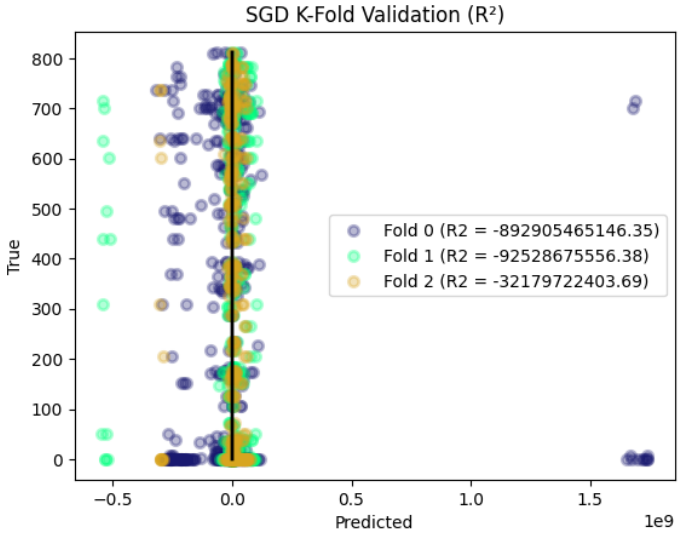


Figure 3.14: Validación cruzada K-Fold (3 particiones) del SGD.

heterocedasticidad[1] de los residuos.

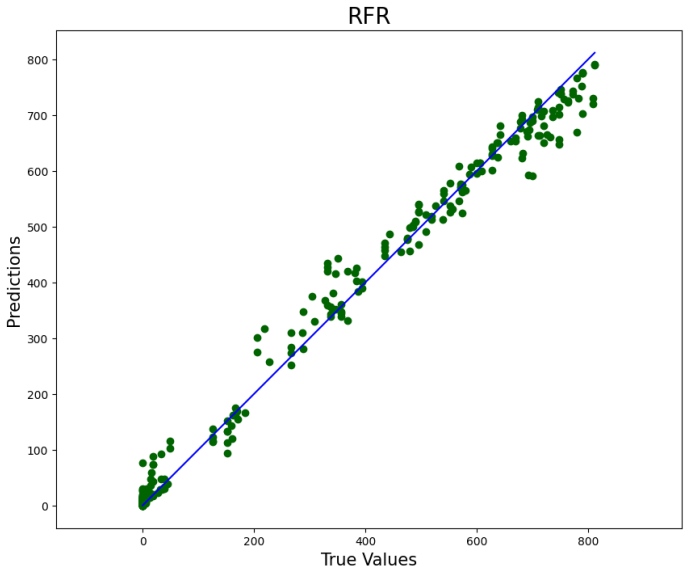


Figure 3.15: Predicciones del bosque aleatorio contra valores reales de PolyPwr. La concentración de los puntos sobre la línea de identidad refleja un R^2 de 0.991.

La validación cruzada K-Fold con 3 particiones (Figura 3.16) muestra una consistencia notable: cada uno de los tres pliegues alcanza un R^2 de 0.99, y la distribución de los puntos (predichos vs. reales) se superpone de forma casi idéntica entre los pliegues. Este resultado descarta la posibilidad de que el desempeño observado sea un artefacto de una partición favorable y confirma que la relación aprendida por el RFR es robusta y generalizable dentro del dominio temporal del conjunto de datos.

El diagrama de residuos (Figura 3.17) revela un patrón diagnósticamente relevante: los residuos se distribuyen de manera aproximadamente simétrica alrededor de cero, pero la amplitud de la dispersión aumenta con el valor predicho. Los errores son más contenidos en el rango de baja generación (0–100 MW, donde los residuos rara vez superan $\pm$20 MW) y más dispersos en el rango de alta generación (600–800 MW, donde se observan residuos de hasta $\pm$100 MW). Este patrón de heterocedasticidad es físicamente esperable y será discutido en la sección 3.2.4.

Las curvas de aprendizaje del RFR (Figuras 3.18–3.20) exhiben un comportamiento ejemplar: para la métrica R^2, la curva de entrenamiento comienza en valores cercanos a 0.996 y se estabiliza alrededor de 0.999, mientras que la curva de validación cruzada parte de aproximadamente 0.978 y converge progresivamente hacia 0.998. La brecha entre ambas curvas se reduce con el número de ejemplos de entrenamiento y es consistentemente pequeña (inferior a 0.002 en R^2 para el tamaño máximo de entrenamiento), lo que indica un modelo bien calibrado sin sobreajuste

[1]Propiedad de un modelo de regresión en la que la varianza del error de predicción no es constante sino que varía sistemáticamente en función del nivel de la predicción. En el contexto fotovoltaico, implica que el modelo comete errores mayores durante los períodos de alta generación que durante los de baja irradiancia.

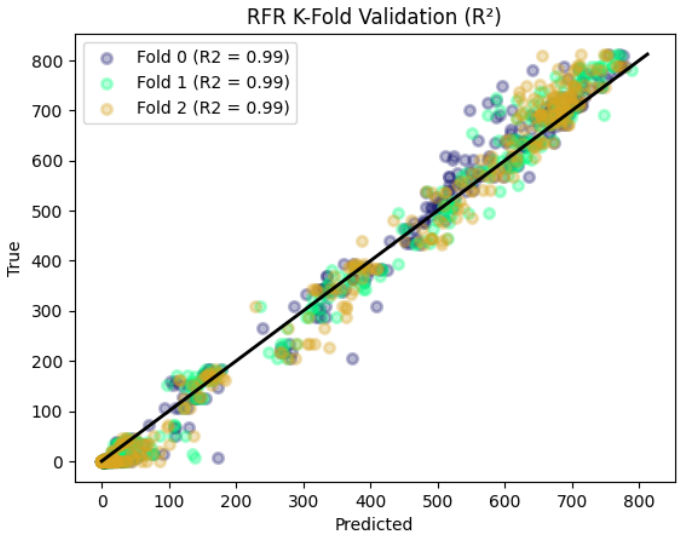


Figure 3.16: Validación cruzada K-Fold (3 particiones) del bosque aleatorio. Los tres pliegues alcanzan $R^2 = 0.99$.

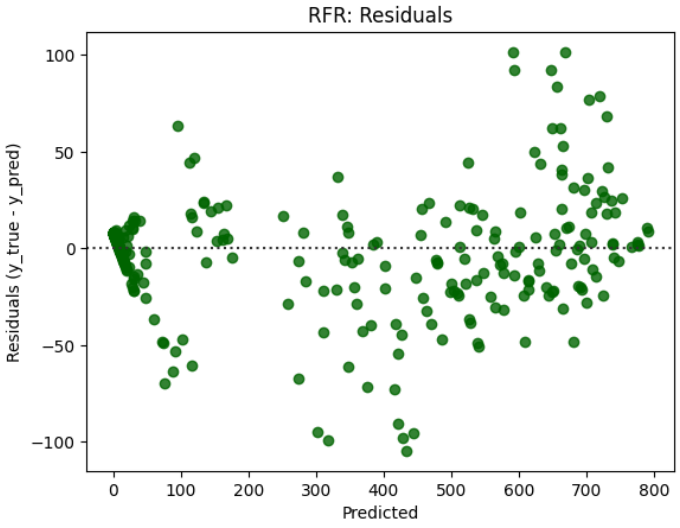


Figure 3.17: Diagrama de residuos del bosque aleatorio. La distribución aproximadamente simétrica con incremento de dispersión para valores predichos altos indica heterocedasticidad moderada.

significativo y sugiere que el conjunto de datos de 1,616 registros es suficiente para la correcta calibración de esta arquitectura.

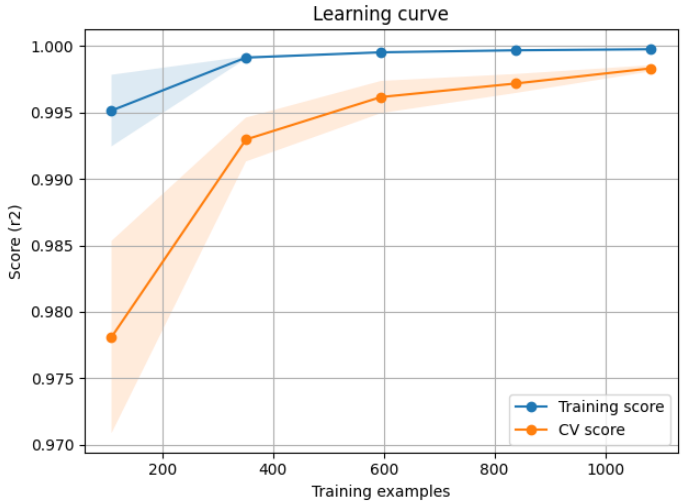


Figure 3.18: Curva de aprendizaje del bosque aleatorio para R^2.

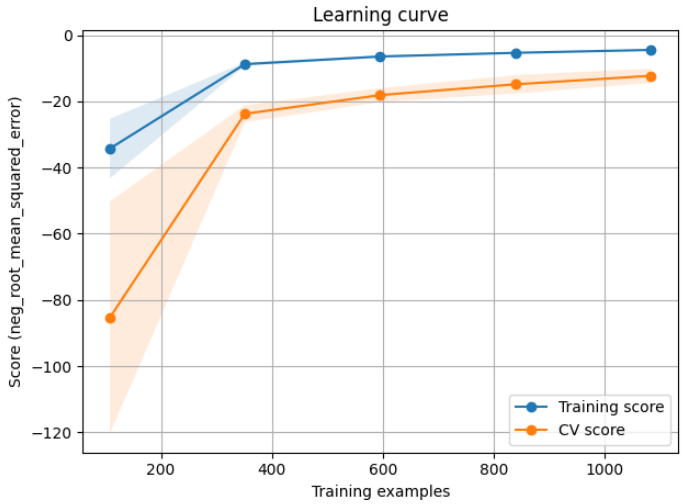


Figure 3.19: Curva de aprendizaje del bosque aleatorio para RMSE.

3.1.4 Perceptrón multicapa (MLPRegressor)

El perceptrón multicapa, configurado con los hiperparámetros óptimos (alpha=10, hidden_layer_sizes=(50, 100, 50), learning_rate='adaptive', solver='sgd'), alcanzó el mejor desempeño global entre los cuatro modelos evaluados: MAE: 15.466 MW; RMSE: 22.800 MW; R^2: 0.994.

El MLP supera al bosque aleatorio en las tres métricas: el MAE es un 3.3% menor (15.466 vs. 15.996 MW), el RMSE un 16.3% menor (22.800 vs. 27.224 MW), y el R^2 alcanza 0.994 frente a 0.991. Si bien la diferencia en MAE es de apenas 0.53 MW, magnitud que podría considerarse marginal en términos prácticos, la diferencia en RMSE resulta más sustantiva: la reducción

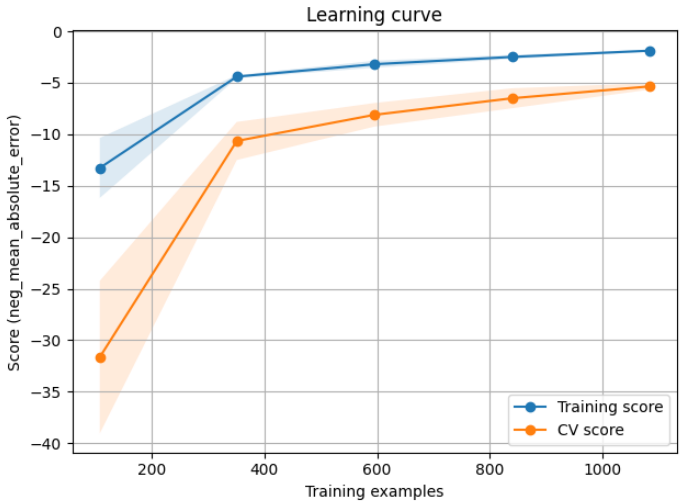


Figure 3.20: Curva de aprendizaje del bosque aleatorio para MAE.

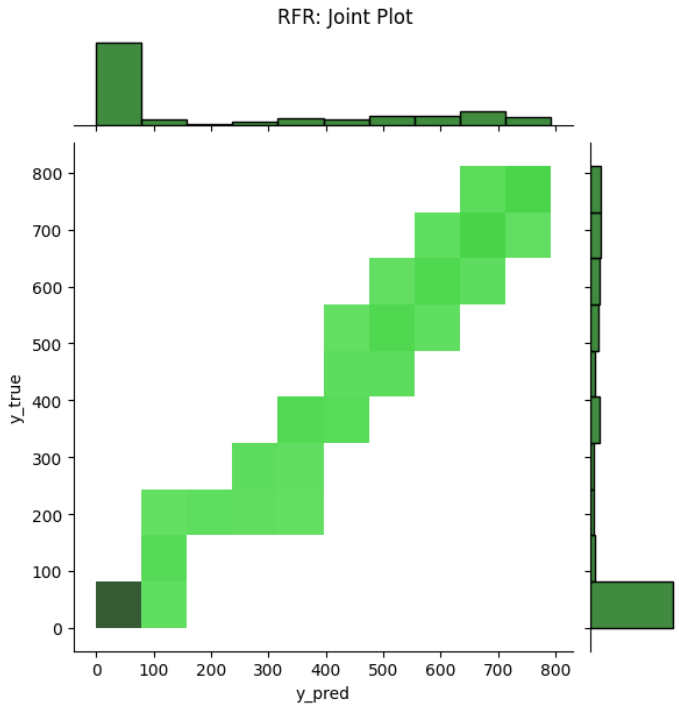


Figure 3.21: Diagrama de distribución conjunta (JointPlot) del bosque aleatorio.

de 27.2 a 22.8 MW (4.42 MW) indica que el MLP no solo comete errores más pequeños en promedio sino que, crucialmente, comete menos errores grandes. Dado que el RMSE penaliza cuadráticamente los errores de mayor magnitud, esta diferencia sugiere que la arquitectura de tres capas ocultas del MLP logra capturar interacciones de orden superior entre los predictores que el ensamble de árboles del RFR no modela con la misma fidelidad, resultado consistente con el teorema de aproximación universal de las redes neuronales [7] y con la evidencia empírica de estudios recientes que han encontrado que el MLP iguala o supera al Random Forest en pronóstico fotovoltaico cuando la optimización de hiperparámetros es exhaustiva [8][9].

La predicción media del MLP (227.51 MW) es cercana a la media real (226.0 MW), con una diferencia de 1.51 MW que confirma una calibración adecuada. La gráfica de predicciones contra valores reales (Figura 3.22) exhibe un patrón visualmente similar al del RFR, con una dispersión ligeramente menor alrededor de la línea de identidad, especialmente notable en el rango de generación media (200–500 MW) donde el MLP mantiene una adherencia más estrecha a los valores observados.

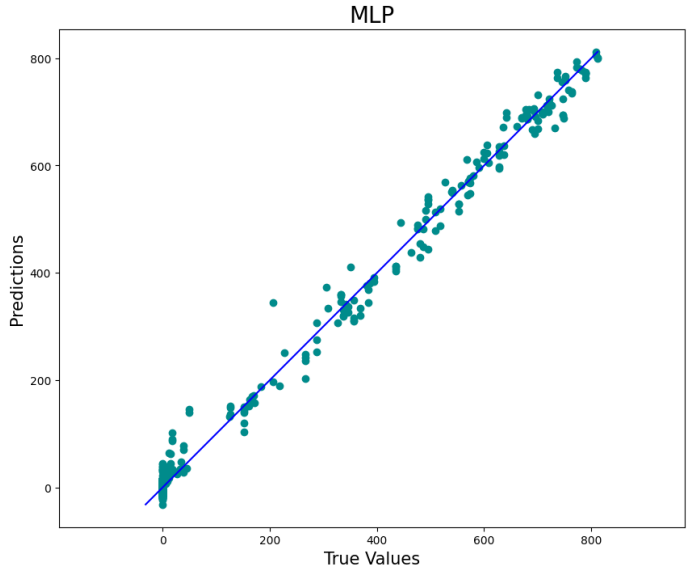


Figure 3.22: Predicciones del MLP contra valores reales de PolyPwr. Con un R^2 de 0.994, el MLP exhibe la concentración más estrecha alrededor de la línea de identidad.

La validación cruzada K-Fold con 3 particiones (Figura 3.23) muestra, al igual que para el RFR, un R^2 de 0.99 consistente en los tres pliegues, lo que confirma que el desempeño del MLP es estable y reproducible a través de distintas particiones de los datos de entrenamiento. El diagrama de residuos (Figura 3.24) presenta un patrón que comparte la heterocedasticidad observada en el RFR, mayor dispersión para valores predichos más altos, pero con una amplitud de residuos ligeramente menor. Se observa, además, un leve sesgo negativo (subestimación) en el rango de predicción media (200–400 MW), lo que sugiere que la arquitectura de tres capas ocultas (50, 100, 50) podría beneficiarse de ajustes adicionales para mejorar la representación de las transiciones

entre los regímenes de baja y alta generación.

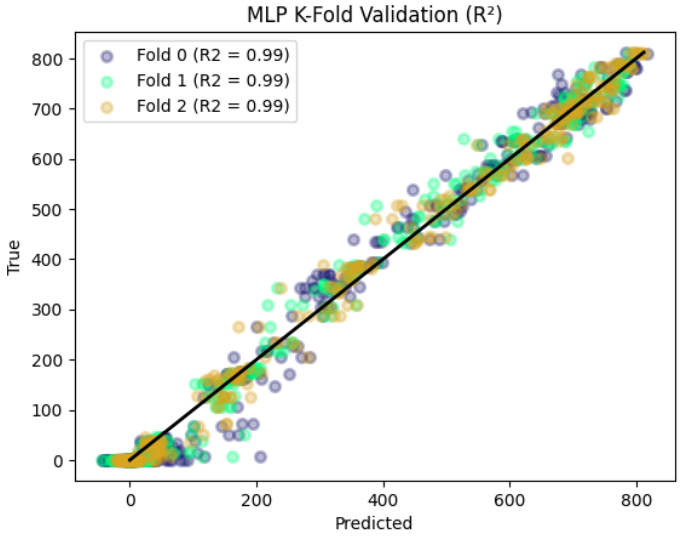


Figure 3.23: Validación cruzada K-Fold (3 particiones) del MLP. La consistencia con $R^2 = 0.99$ en cada pliegue confirma la estabilidad del modelo.

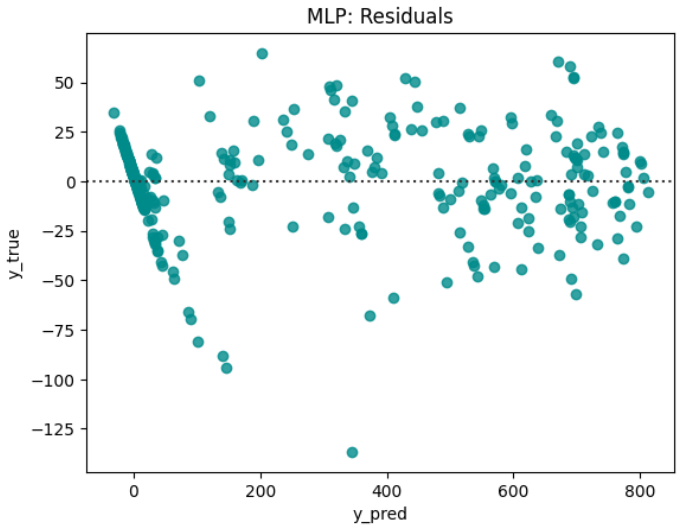


Figure 3.24: Diagrama de residuos del MLP. La distribución presenta heterocedasticidad similar a la del RFR, con un leve sesgo negativo en el rango de predicción media.

Las curvas de aprendizaje del MLP (Figuras 3.25–3.27) revelan un patrón distinto al del RFR: la curva de R^2 muestra valores negativos para tamaños de entrenamiento pequeños (por debajo de 400 registros), que se incrementan de forma progresiva y aproximadamente lineal hasta alcanzar valores positivos cercanos a 0.35 con el tamaño máximo de entrenamiento. Este comportamiento refleja la naturaleza del MLP como modelo de alta capacidad con requerimientos de datos más exigentes: con pocos ejemplos, la red neuronal no tiene suficiente información para calibrar los miles de pesos de sus tres capas ocultas, y el resultado es un modelo que produce predicciones

peores que la media. A medida que el número de ejemplos crece, los pesos se ajustan progresivamente y el desempeño mejora de forma sostenida. El hecho de que la curva de validación cruzada no se haya estabilizado completamente con el volumen de datos disponible sugiere que el MLP podría beneficiarse de conjuntos de entrenamiento más grandes, una línea de trabajo futuro que se aborda en la sección 3.6.

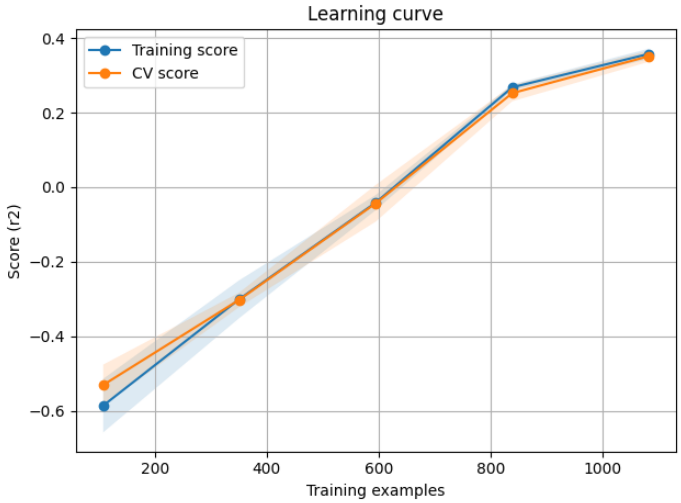


Figure 3.25: Curva de aprendizaje del MLP para R^2.

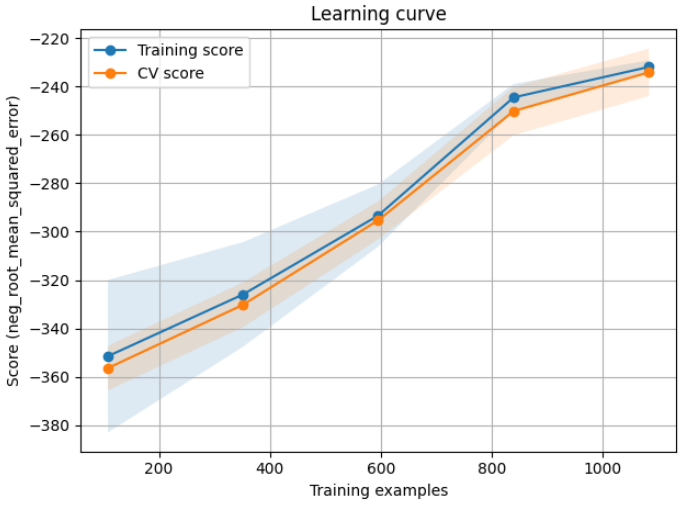


Figure 3.26: Curva de aprendizaje del MLP para RMSE.

3.1.5 Tabla comparativa consolidada

La Tabla 3.1 resume los resultados de los cuatro modelos sobre el conjunto de prueba. La media de la variable objetivo en el conjunto de prueba es de 226.0 MW, valor que sirve como referencia para interpretar la magnitud absoluta de los errores.

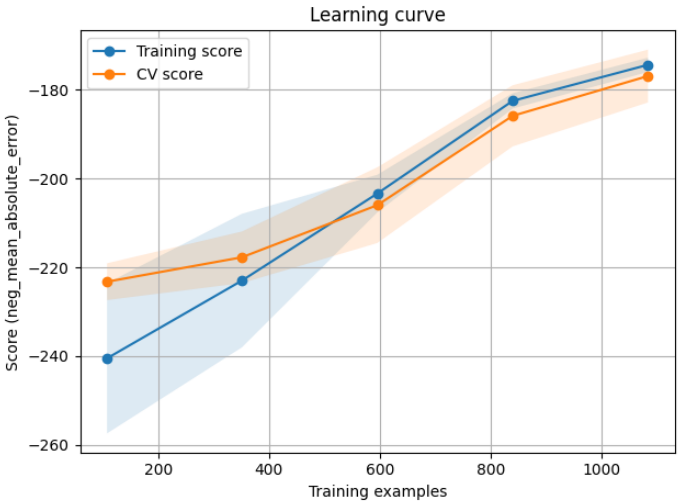


Figure 3.27: Curva de aprendizaje del MLP para MAE.

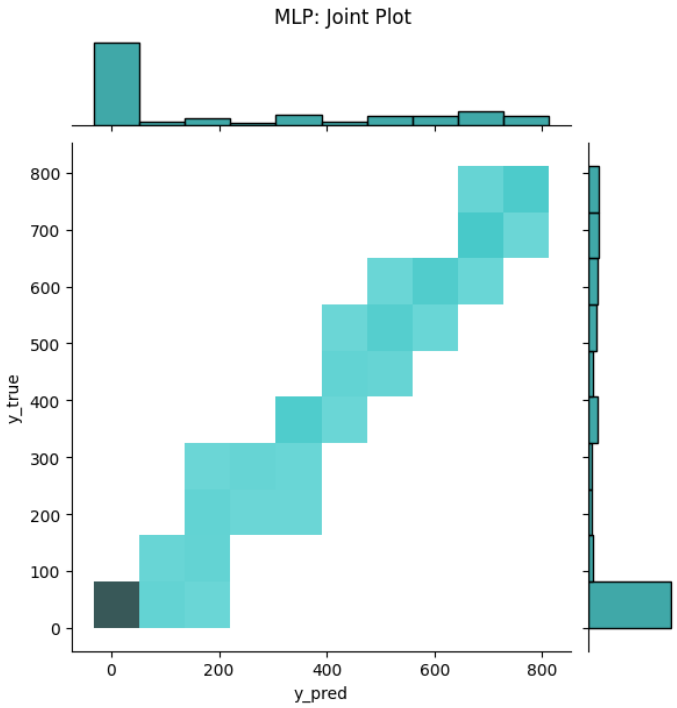


Figure 3.28: Diagrama de distribución conjunta (JointPlot) del MLP.

Modelo	**MAE (MW)**	**RMSE (MW)**	**R^2**	**Pred. Media**	**Δ MAE vs. Dummy**
DummyRegressor	259.601	287.216	≈ 0.000	226.00	–
SGDRegressor	188.086	232.165	0.347	227.24	−27.5%
RandomForest	**15.996**	27.224	0.991	228.49	−93.8%
MLPRegressor	**15.466**	**22.800**	**0.994**	227.51	−94.0%

Table 3.1: Desempeño comparativo de los cuatro modelos sobre el conjunto de prueba ($n = 391$). Los valores en negritas indican el mejor desempeño por métrica.

Las Figuras 3.29, 3.30 y 3.31 presentan las comparaciones por métrica en formato de gráfico de barras, permitiendo visualizar tanto la magnitud absoluta de las diferencias como la brecha cualitativa entre los dos grupos de modelos (lineales vs. no lineales).

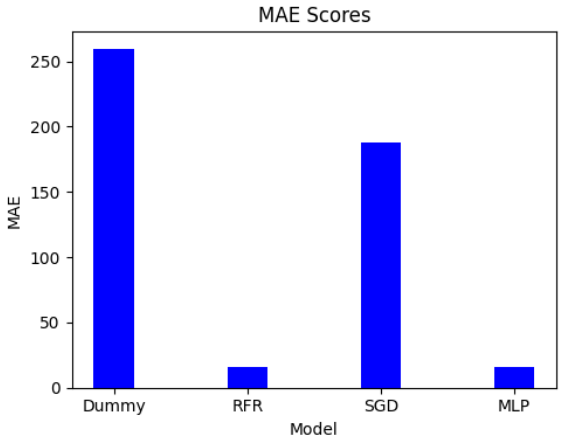


Figure 3.29: Comparación del MAE por modelo. La reducción de más del 90% en el error absoluto entre los modelos no lineales y la línea base.

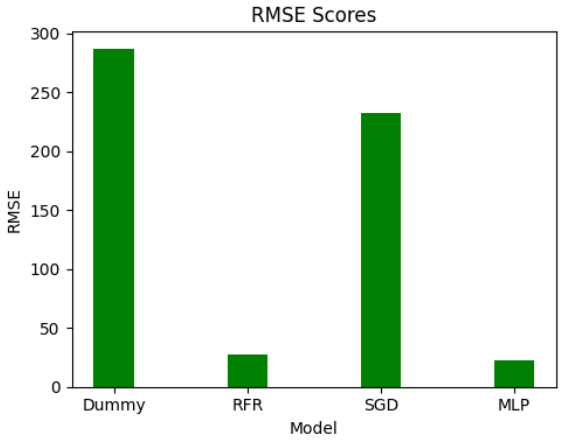


Figure 3.30: Comparación del RMSE por modelo.

La Figura 3.32 muestra la distribución de densidad estimada por kernel (KDE) de los valores predichos por cada modelo frente a la distribución de los valores reales. Esta gráfica es particularmente reveladora porque permite evaluar no sólo la precisión puntual de las predicciones sino su capacidad para reproducir la estructura distribucional completa de PolyPwr.

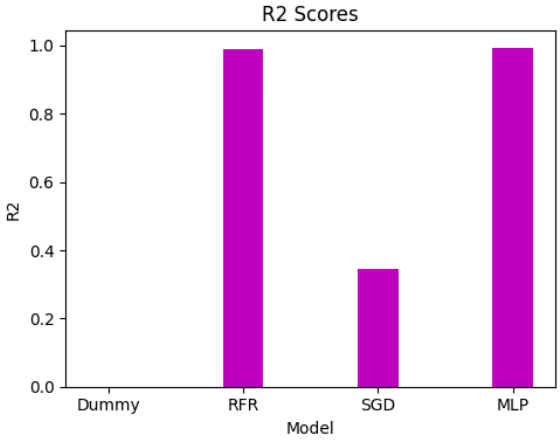


Figure 3.31: Comparación del R^2 por modelo. El contraste entre el $R^2 \approx 0$ del Dummy, el 0.347 del SGD y los valores superiores a 0.99 del RFR y MLP evidencia la estructura no lineal del problema.

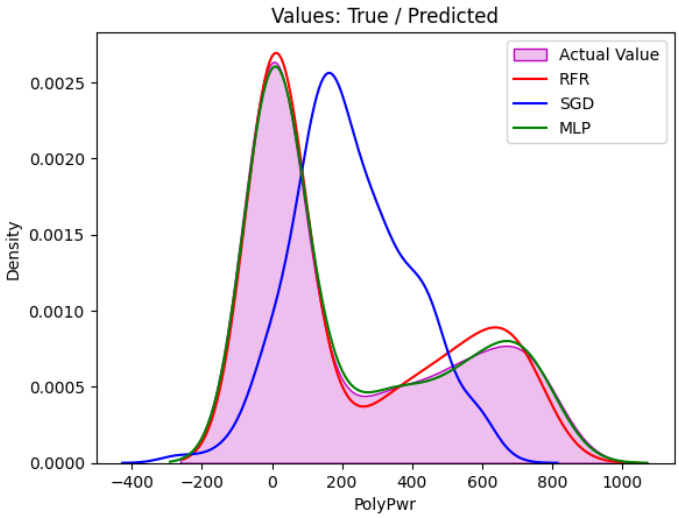


Figure 3.32: Distribución KDE de los valores predichos por cada modelo frente a los valores reales de PolyPwr. Las curvas del RFR y el MLP se superponen casi exactamente con la distribución bimodal real.

3.2 Análisis Comparativo de los Modelos

Los resultados numéricos presentados en la sección anterior permiten establecer un ranking de desempeño inequívoco: MLP > RFR » SGD » Dummy, pero no explican por qué ese ranking se produce ni qué información diagnóstica aporta sobre la naturaleza del problema de predicción. Esta sección desarrolla ese análisis interpretativo, examinando las brechas de desempeño a la luz de los fundamentos algorítmicos de cada modelo, y extrayendo las conclusiones diagnósticas que fundamentan las recomendaciones de las secciones posteriores. El análisis se estructura en torno a cuatro ejes: la evidencia de no linealidad, la ganancia marginal de la complejidad, la función del regresor ficticio como referencia, y el diagnóstico de los patrones de residuos.

3.2.1 La brecha SGD–RFR/MLP como evidencia de no linealidad

El resultado más informativo del análisis comparativo no es el valor absoluto del R^2 del mejor modelo, sino la magnitud de la brecha entre el modelo lineal SGD ($R^2 = 0.347$) y los dos modelos no lineales: RFR ($R^2 = 0.991$) y MLP ($R^2 = 0.994$). Esta brecha de más de 0.64 unidades de R^2 constituye evidencia empírica directa de que la relación entre las variables meteorológicas y la potencia de salida fotovoltaica en Puerto Peñasco está dominada por interacciones no lineales que un modelo lineal, por bien optimizado que esté, no puede capturar.

Esta conclusión era anticipable desde el punto de vista físico: la generación fotovoltaica depende de la posición solar: función trigonométrica del tiempo y la latitud, de la interacción entre temperatura e irradiancia, un producto que introduce un término cuadrático implícito, y de umbrales de nubosidad que generan discontinuidades en la potencia de salida. Sin embargo, la magnitud de la brecha observada en Puerto Peñasco es notablemente grande: en la práctica, significa que casi dos tercios de la variabilidad de PolyPwr está codificada en patrones que requieren, como mínimo, modelos capaces de representar interacciones multiplicativas y umbrales no lineales entre las variables de entrada. Este hallazgo es consistente con la evidencia acumulada en la literatura reciente sobre pronóstico fotovoltaico. Markovics y Mayer [8], al comparar 24 modelos de aprendizaje automático sobre datos de 16 plantas fotovoltaicas en Hungría, encontraron que los modelos no lineales superaron consistentemente a los modelos lineales. De forma similar, Jogunuri et al. [10] demostraron que el Random Forest supera a los modelos de regresión lineal en pronóstico fotovoltaico estacional multi-paso, con mejoras en la precisión del pronóstico que se amplifican durante los meses de mayor variabilidad meteorológica.

La gráfica KDE de la Figura 3.32 aporta una visualización complementaria y particularmente elocuente de esta conclusión. La distribución real de PolyPwr presenta una estructura claramente bimodal. El SGD produce una distribución unimodal que no reproduce esta bimodalidad: al estar limitado a combinaciones lineales de los predictores, no puede generar la separación entre los dos modos de operación. En contraste, tanto el RFR como el MLP reproducen con notable fidelidad la estructura bimodal de la distribución real, lo que confirma su capacidad para modelar los dos

regímenes operativos del sistema como estados cualitativamente distintos y no meramente como extremos de un continuo lineal.

Este hallazgo tiene implicaciones directas que trascienden el caso de Puerto Peñasco: para cualquier intento de implementar sistemas de pronóstico en centrales fotovoltaicas con perfiles climáticos de alta insolación y ciclos diurnos pronunciados, los modelos lineales son insuficientes para producir predicciones con la precisión que las aplicaciones operativas exigen. La inversión computacional y metodológica en modelos no lineales queda justificada por la ganancia de más de 0.64 unidades de R^2 y una reducción del orden del 90% en MAE.

3.2.2 MLP versus RFR: la ganancia marginal de la complejidad

Si bien el MLP supera al RFR en las tres métricas, la magnitud de la ventaja merece un análisis matizado que considere tanto los aspectos cuantitativos como los operativos. La diferencia en R^2 es de 0.003 (0.994 vs. 0.991), en MAE de 0.53 MW (15.466 vs. 15.996) y en RMSE de 4.42 MW (22.800 vs. 27.224). La diferencia en RMSE es la más sustantiva de las tres: la reducción del 16.3% indica que el MLP logra contener mejor los errores de mayor magnitud.

Abdelsattar et al. [9] reportaron un R^2 de 0.994 con CatBoost y Gradient Boosting para predicción de potencia en plantas PV, valor idéntico al obtenido por el MLP en este estudio, lo que sitúa el desempeño de Puerto Peñasco al nivel de los mejores resultados publicados internacionalmente con técnicas de ML avanzadas. Asiedu et al. [11] encontraron que la Artificial Neural Network (ANN) fue el modelo con mejor desempeño para pronóstico de día siguiente, mientras que Random Forest dominó en horizontes de dos semanas y un mes, sugiriendo que la ventaja relativa entre redes neuronales y modelos de ensemble puede depender del horizonte de predicción. Una revisión sistemática de 36 estudios de pronóstico PV con deep learning [12] confirmó que los modelos basados en perceptrones multicapa alcanzan consistentemente R^2 entre 0.88 y 0.99 según la resolución temporal, el horizonte de predicción y las variables de entrada.

Sin embargo, la ventaja cuantitativa del MLP debe ponderarse contra varios factores operativos que favorecen al RFR. En primer lugar, la diferencia en MAE es inferior a 1 MW, magnitud que en la práctica operativa podría quedar dentro del margen de ruido de los propios instrumentos de medición. En segundo lugar, el RFR ofrece ventajas prácticas significativas: es intrínsecamente interpretable a través de las importancias de las variables, no requiere estandarización de las entradas, es robusto frente a valores atípicos y tiene un comportamiento de entrenamiento estable. En tercer lugar, la curva de aprendizaje del RFR muestra un comportamiento de estabilización que indica suficiencia de datos, mientras que la del MLP sugiere que la red neuronal podría beneficiarse de conjuntos de entrenamiento más amplios.

La conclusión diagnóstica es que ambos modelos alcanzan un nivel de desempeño operativamente excelente, y que la elección entre uno y otro en un contexto de implementación real dependería del balance entre la ligera ventaja predictiva del MLP y las ventajas prácticas de interpretabilidad y

robustez del RFR. La estrategia óptima podría incluso combinar ambos: utilizar el RFR como modelo principal de producción por su robustez, y el MLP como modelo de validación que genera alertas cuando la diferencia entre ambas predicciones supera un umbral predefinido.

3.2.3 El regresor ficticio como referencia metodológica

El regresor ficticio cumple una función que trasciende la mera verificación del pipeline: su inclusión permite cuantificar exactamente cuánto aporta la información contenida en las variables predictoras. La diferencia entre el R^2 del mejor modelo (0.994) y el R^2 del regresor ficticio ($\approx$0.000) indica que el pipeline completo, desde la selección de características hasta el entrenamiento del MLP, logra capturar el 99.4% de la variabilidad de PolyPwr que el regresor ficticio deja completamente sin explicar. Asimismo, la reducción en MAE de 259.6 MW a 15.5 MW (un 94.0%) confirma que el conjunto de variables meteorológicas y temporales disponibles contiene la información necesaria para predecir la potencia de salida con alta precisión.

La progresión Dummy $\rightarrow$ SGD $\rightarrow$ RFR/MLP ($0 \rightarrow 0.347 \rightarrow 0.991/0.994$) dibuja un arco de complejidad creciente donde la mayor ganancia marginal se produce en la transición del modelo lineal al modelo no lineal, no en la transición del modelo no lineal simple al complejo. Esto es un resultado con valor práctico: sugiere que la decisión de inversión metodológica más importante es la adopción de un modelo no lineal, mientras que la elección específica dentro de la familia de modelos no lineales tiene un impacto incremental comparativamente menor.

3.2.4 Análisis de los patrones de residuos

El examen de los diagramas de residuos revela patrones diagnósticamente relevantes que complementan la información proporcionada por las métricas globales. En el caso del RFR (Figura 3.17), los residuos se distribuyen de manera aproximadamente simétrica alrededor de cero, pero la amplitud de la dispersión aumenta con el valor predicho: los errores son más contenidos en el rango de baja generación (0–100 MW, donde los residuos rara vez superan $\pm$20 MW) y más dispersos en el rango de alta generación (600–800 MW, donde se observan residuos de hasta $\pm$100 MW).

Este patrón de heterocedasticidad es físicamente esperable y tiene una explicación operativa precisa: los registros de alta generación corresponden a las horas centrales del día donde convergen dos factores que incrementan la incertidumbre intrínseca de la predicción. Por un lado, la interacción entre la irradiancia incidente, que es máxima, y la temperatura de los módulos, que también alcanza sus valores más altos, genera un efecto competitivo. Por otro lado, la variabilidad meteorológica a corto plazo tiene un impacto absoluto mayor cuando la generación base es alta.

Los residuos del MLP (Figura 3.24) presentan un patrón similar en cuanto a heterocedasticidad, pero con una amplitud de residuos ligeramente mayor en valores absolutos y un leve sesgo negativo en el rango medio de predicción (200–400 MW). En contraste radical, los residuos del

SGD (Figura 3.8) presentan un patrón fuertemente estructurado y no aleatorio: una banda de sobreestimación para valores predichos bajos y una banda de subestimación para valores predichos altos, con una dispersión general que abarca un rango de más de 800 MW. Este patrón no aleatorio es la manifestación gráfica directa de las relaciones no lineales que el modelo lineal no puede capturar.

3.3 Contextualización de los Resultados frente a la Literatura

Los resultados obtenidos en Puerto Peñasco no existen en el vacío: su significado se enriquece cuando se comparan con los valores reportados por otras investigaciones que han aplicado técnicas similares de aprendizaje automático al pronóstico de generación fotovoltaica.

3.3.1 Comparación con el estudio de referencia de Pasion et al.

El trabajo de Pasion et al. [2] constituye la referencia más directa para esta comparación, dado que la selección de variables predictoras de este estudio se basa en su propuesta metodológica. Pasion et al. reportaron un R^2 de 0.94 con Random Forest para la predicción de potencia fotovoltaica horizontal utilizando exclusivamente variables meteorológicas y de localización geográfica, sin datos de irradiancia directa, sobre un conjunto de datos de 30 sistemas fotovoltaicos residenciales en Dayton, Ohio (Estados Unidos). El R^2 de 0.991 obtenido en Puerto Peñasco con la misma familia de algoritmo supera en 0.051 unidades al valor de referencia, una diferencia sustantiva en una escala donde los valores cercanos a 1.0 representan niveles de precisión cada vez más difíciles de alcanzar marginalmente.

Varios factores convergen para explicar esta diferencia. En primer lugar, el emplazamiento de Puerto Peñasco presenta una variabilidad climática intrínsecamente menor que la de un sitio en el medio oeste de Estados Unidos. En segundo lugar, la incorporación de variables temporales como Year y Time, ausentes en la propuesta original de Pasion et al., enriquece la información disponible. En tercer lugar, la presente investigación aplicó un proceso formal de selección de características (RFECV) que optimizó el subconjunto de predictores. En cuarto lugar, la mayor escala de la instalación puede producir patrones de generación más regulares a nivel agregado.

3.3.2 Posicionamiento frente a estudios internacionales recientes

En el panorama más amplio de la literatura reciente, los resultados de Puerto Peñasco se posicionan de forma consistente en el rango superior de desempeño reportado para modelos basados en variables meteorológicas sin irradiancia directa. Moufid et al. [13] compararon regresión y clasificación con aprendizaje automático para pronóstico fotovoltaico a corto plazo utilizando un conjunto de datos extenso de Australia, reportando que el Random Forest fue el algoritmo con mejor desempeño en el enfoque de regresión. Aziz et al. [14] desarrollaron un marco integrado para pronóstico de irradiancia solar en Zafarana, Egipto, combinando técnicas

de selección de características con múltiples modelos de ML, en un enfoque metodológico similar al adoptado en esta investigación. Barakat et al. [15] aplicaron modelos multi-etiqueta sobre datos de una planta fotovoltaica conectada a la red, reportando valores de R^2 entre 0.92 y 0.98 según el horizonte temporal.

Un hallazgo particularmente relevante es el reportado por Abdelsattar et al. [9]: utilizando técnicas avanzadas de ML (CatBoost y Gradient Boosting), obtuvieron un R^2 máximo de exactamente 0.994, valor idéntico al alcanzado por el MLP en este estudio. Esta convergencia de resultados entre técnicas distintas y emplazamientos diferentes sugiere que el valor de 0.994 podría estar cercano a un "techo práctico" de predicción para modelos entrenados con variables meteorológicas horarias sin irradiancia directa.

3.3.3 Factores que explican el posicionamiento de Puerto Peñasco

La convergencia de los factores analizados permite formular una explicación estructurada del posicionamiento de Puerto Peñasco en el panorama internacional. La combinación de un emplazamiento con recurso solar excepcionalmente favorable y consistente, un conjunto de variables predictoras bien seleccionado mediante un proceso formal de RFECV, una incorporación de variables temporales que capturan los ciclos diurnos y estacionales con alta resolución, y un pipeline metodológico que incluye búsqueda exhaustiva de hiperparámetros convergen para producir resultados que igualan o superan los estándares internacionales.

No obstante, esta comparación debe interpretarse con las cautelas apropiadas. Las diferencias en los conjuntos de datos, en los protocolos de evaluación, en las condiciones climáticas de cada emplazamiento, y en los criterios de limpieza y tratamiento de datos faltantes limitan la comparabilidad directa entre estudios. Lo que sí puede afirmarse con confianza es que la metodología desarrollada en este trabajo extrae de los datos de Puerto Peñasco un nivel de información predictiva que se posiciona en el extremo superior de lo reportado internacionalmente, validando tanto la calidad del pipeline como la viabilidad de este tipo de análisis para centrales fotovoltaicas en el contexto mexicano.

3.4 Implicaciones Operativas para la Central de Puerto Peñasco

Los resultados cuantitativos adquieren su pleno significado cuando se traducen al lenguaje de las decisiones operativas que motivan esta investigación.

3.4.1 Pronóstico para el despacho programado

En el contexto del despacho programado del sistema eléctrico mexicano, la central debe comunicar al Centro Nacional de Control de Energía (CENACE) una estimación de la generación prevista para las próximas horas o el día siguiente. La calidad de esta estimación tiene consecuencias

económicas directas: una sobreestimación obliga a recurrir a reservas de capacidad de otras fuentes, generalmente térmicas, con costos marginales más altos y mayor huella de carbono, mientras que una subestimación puede resultar en generación no programada que debe ser absorbida por el sistema.

Con un error absoluto promedio de 15.5 MW por registro horario, el modelo MLP proporciona predicciones cuya desviación media respecto a la generación real representa aproximadamente el 6.8% de la media de producción (226.0 MW). Cabe señalar que el RMSE, al penalizar cuadráticamente los errores grandes, aporta información complementaria sobre los errores en los escenarios más exigentes. El RMSE del MLP (22.8 MW) indica que los errores de mayor magnitud están razonablemente contenidos, y es un 16.3% menor que el RMSE del RFR (27.2 MW).

Asimismo, la capacidad de los modelos no lineales para reproducir la distribución bimodal de PolyPwr (Figura 3.32) tiene una implicación operativa directa: los errores de predicción se concentran en las transiciones amanecer–mediodía y mediodía–atardecer, que son precisamente los períodos en los que el operador del sistema necesita la mayor precisión para planificar las rampas de generación complementaria. El hecho de que tanto el RFR como el MLP reproduzcan con fidelidad ambos modos de la distribución indica que las transiciones entre los regímenes están adecuadamente modeladas.

3.4.2 Detección de anomalías y mantenimiento predictivo

Una aplicación indirecta pero potencialmente de alto valor de los modelos predictivos es la detección de anomalías operativas. Cuando la generación observada se desvía de forma persistente y significativa de la generación predicha por el modelo, controlando por las condiciones meteorológicas prevalentes, esa desviación puede indicar degradación acelerada en los módulos, suciedad acumulada en las superficies (*soiling*), fallas en inversores, o problemas en los circuitos de seguimiento solar.

La magnitud de los residuos normales del modelo proporciona la referencia cuantitativa necesaria para definir umbrales de alerta. Un residuo que supere consistentemente los 2–3 desvíos estándar del patrón normal durante varias horas consecutivas constituiría una señal de alerta estadísticamente significativa. Estudios recientes sobre degradación de módulos PV han documentado que la tasa de degradación mediana global es de aproximadamente 1.00%/año [18], con una media de 1.27%/año. Un meta-análisis de 80 estudios primarios con 610 observaciones encontró una tasa mediana de 0.94%/año [19]. Un sistema de monitoreo basado en el modelo predictivo permitiría detectar tasas de degradación anómalas antes de que se manifiesten como pérdidas significativas de producción.

3.4.3 Planificación de la generación y gestión de la interconexión

Como se describió en el Capítulo I y en la sección 2.1 del Capítulo II, la central de Puerto Peñasco cumple un papel estratégico como primer vínculo permanente entre Baja California y el Sistema Eléctrico Nacional (SEN). En este contexto, la capacidad de predecir con alta precisión la producción horaria tiene implicaciones operativas de primer orden. Una predicción confiable permite al operador del sistema planificar con antelación el flujo de potencia por la línea de interconexión, dimensionar las reservas de capacidad necesarias en ambos lados de la interconexión, y optimizar la coordinación entre la generación de Puerto Peñasco y las fuentes de generación complementarias en la península.

3.5 Perspectivas de Implementación y el Rol de la IA Generativa

La transición de un modelo que funciona en un entorno de investigación a un modelo que opera en un entorno de producción es un proceso que la comunidad de aprendizaje automático aplicado ha denominado MLOps (*Machine Learning Operations*) y que requiere consideraciones adicionales de ingeniería, gobernanza de datos e integración con los sistemas existentes.

3.5.1 Lineamientos para la implementación operativa

La implementación operativa requiere resolver, como mínimo, cinco desafíos técnicos. **Integración con la infraestructura de datos existente**: el modelo necesita recibir las variables predictoras en tiempo real desde los sistemas SCADA y las fuentes meteorológicas externas. **Reentrenamiento periódico del modelo**: un estudio reciente de meta-análisis [20] encontró que el reentrenamiento adaptativo produce mejoras en precisión de hasta un 9.3% comparado con un 4.1% para el reentrenamiento periódico. **Monitoreo de la deriva del modelo** (*concept drift*): Zink, Ioshchikhes y Weigold [21] evaluaron cinco algoritmos de detección de *concept drift* para pronóstico de carga industrial, encontrando que los métodos DDM y KSWIN fueron particularmente efectivos. Un marco reciente presentado en AAAI [22] propuso la descomposición de la degradación del rendimiento en componentes de *covariate shift* y *label shift*. **Cuantificación de incertidumbre**: Renkema, Visser y AlSkaif [23] propusieron un marco de predicción conformal aplicado a pronóstico fotovoltaico de día siguiente, demostrando mejoras del 14% en *weighted interval score*. Una investigación más reciente [24] presentó una de las primeras evaluaciones comprensivas que integra predicción conformal con Random Forest, XGBoost, LSTM y Transformer para pronóstico en plantas PV reales. Wang et al. [25] propusieron un marco probabilístico basado en regresión cuantílica conformalizada. **Gobernanza del modelo**: un protocolo mínimo debería incluir versionado del modelo y del código, registro de todas las predicciones generadas, procedimientos de escalamiento cuando las métricas caigan por debajo de umbrales críticos, y definición clara de roles y responsabilidades.

3.5.2 El rol potencial de los LLMs como capa de interpretación

La sección 2.7 del Capítulo II introdujo la discusión sobre el papel que los modelos de lenguaje de gran escala (LLMs) podrían desempeñar como capa de interpretación y comunicación. El panorama ha evolucionado rápidamente. Majumder et al. [26] publicaron en Joule un análisis de las capacidades y limitaciones de los LLMs en el sector eléctrico, concluyendo que estos modelos pueden mejorar la eficiencia operativa al facilitar la interacción entre datos del sistema, software y herramientas. Una revisión comprensiva de 133 artículos publicada en Applied Energy en 2025 [27] identificó 13 roles distintos que los LLMs pueden desempeñar en sistemas energéticos. Chen et al. [28] presentaron el modelo GAIA (*Grid Artificial Intelligent Assistant*), un LLM diseñado específicamente para asistir en tareas operativas de sistemas eléctricos. Buster [29] del NREL desarrolló una interfaz de chat basada en LLMs diseñada para consultar reportes técnicos de investigación en energía renovable.

En el contexto específico de Puerto Peñasco, las tres aplicaciones concretas propuestas en la sección 2.7 del Capítulo II se pueden ahora formular con mayor especificidad: la generación automatizada de reportes operativos, la interfaz conversacional para consultas sobre el historial operativo, y la síntesis de alertas ante desviaciones significativas. Es esencial reiterar que el valor de los LLMs está en la capa de comunicación e interpretación, no en la capa de predicción cuantitativa [31][27]. La arquitectura funcionalmente madura es la de un sistema integrado en el que un modelo de ML especializado produce las predicciones numéricas de alta precisión y un LLM actúa como interfaz conversacional sobre esos resultados.

3.6 Limitaciones del Estudio y Líneas de Trabajo Futuro

Toda investigación opera dentro de un espacio de validez definido por sus supuestos, sus datos y sus decisiones metodológicas. Delimitar ese espacio con honestidad no debilita los hallazgos sino que los fortalece, al indicar con precisión las condiciones bajo las cuales los resultados son confiables y las direcciones en las que el trabajo puede ser extendido.

3.6.1 Limitaciones del estudio

Alcance temporal del conjunto de datos. Los 2,016 registros horarios cubren un período de aproximadamente 20 meses (diciembre 2023 – agosto 2025), lo que incluye un ciclo completo de las cuatro estaciones pero no permite evaluar la estabilidad del modelo a lo largo de múltiples años completos. La degradación natural de los módulos fotovoltaicos, que según el compendio global más reciente presenta una mediana de 1.00%/año [18] y según un meta-análisis de 610 observaciones una mediana de 0.94%/año [19], no es observable de forma robusta en un horizonte de 20 meses.

Ausencia de datos de irradiancia directa. La decisión de prescindir de la irradiancia di-

recta como predictor, fundamentada en la propuesta de Pasion et al. [2], es metodológicamente coherente pero introduce un límite superior al desempeño del modelo. La revisión sistemática de pronóstico PV con deep learning [12] confirmó que los modelos que incorporan irradiancia directa o global como predictor tienden a alcanzar R^2 marginalmente superiores.

Especificidad geográfica del modelo. Los resultados son válidos para las condiciones climáticas y operativas de Puerto Peñasco; su transferencia directa a centrales en emplazamientos con perfiles climáticos distintos no está garantizada sin un proceso de recalibración con datos locales. El pipeline metodológico, sin embargo, es completamente transferible.

Naturaleza puntual de las predicciones. El modelo produce estimaciones puntuales sin cuantificar la incertidumbre asociada a cada predicción individual. Las técnicas de predicción conformal [23][24][25] ofrecen un camino promisorio para extender el pipeline actual en esta dirección.

Resolución temporal fija. Los modelos fueron entrenados y evaluados exclusivamente con resolución horaria. Aplicaciones como la regulación de frecuencia o la participación en mercados intradiarios de 15 minutos requieren resoluciones temporales más finas.

3.6.2 Líneas de trabajo futuro

Las limitaciones identificadas definen las direcciones prioritarias para la extensión de este trabajo. **Ampliación longitudinal del conjunto de datos**, que permitiría evaluar la estabilidad del modelo a largo plazo y capturar el efecto de la degradación de los módulos. **Cuantificación de incertidumbre mediante predicción conformal**, siguiendo la metodología de Renkema et al. [23]. **Extensión multiescala del pronóstico** a horizontes sub-horarios y de día siguiente. **Desarrollo de un prototipo funcional del sistema integrado ML+LLM**, siguiendo los referentes de GAIA [28] y ELM del NREL [29]. **Validación del pipeline en otras centrales del parque fotovoltaico mexicano**, que permitiría evaluar la robustez del enfoque metodológico más allá del caso de Puerto Peñasco.

3.7 Conclusiones del Capítulo III

Este capítulo ha presentado, analizado y contextualizado los resultados del pipeline de aprendizaje automático desarrollado en el Capítulo II, completando el arco argumentativo de la investigación con las respuestas empíricas a las preguntas que motivaron el trabajo.

Los resultados cuantitativos demuestran que los modelos no lineales, en particular, el Perceptrón Multicapa (R^2 = 0.994, MAE= 15.466 MW, RMSE= 22.800 MW) y el Bosque Aleatorio (R^2 = 0.991, MAE= 15.996 MW, RMSE= 27.224 MW), alcanzan un nivel de precisión que supera los estándares reportados en la literatura internacional para pronóstico fotovoltaico basado en variables meteorológicas [2][8][13][15], y que resulta compatible con las exigencias operativas de

una central de la escala de Puerto Peñasco. El R^2 de 0.994 obtenido por el MLP iguala el valor máximo reportado por Abdelsattar et al. [9] con técnicas de CatBoost/Gradient Boosting en un estudio independiente, lo que sugiere que este valor puede estar cercano al techo práctico de predicción para modelos horarios basados en variables meteorológicas sin irradiancia directa. La brecha de desempeño con el modelo lineal SGD ($R^2 = 0.347$), una diferencia de más de 0.64 unidades de R^2, constituye evidencia directa de que el problema de predicción está dominado por interacciones no lineales.

El análisis diagnóstico, basado en los patrones de residuos, las curvas de aprendizaje, la validación cruzada K-Fold y la comparación de las distribuciones predictivas, confirma que los dos modelos no lineales producen predicciones robustas, bien calibradas y capaces de reproducir la estructura bimodal de la distribución de generación. La diferencia entre el MLP y el RFR, aunque favorable al primero, es suficientemente estrecha para que la elección entre ambos dependa de consideraciones operativas más que de diferencias técnicas categóricas.

Las implicaciones operativas abarcan tres dominios concretos: pronóstico para el despacho programado con un error del orden del 7% de la media de producción, detección de anomalías basada en los patrones de residuos normales del modelo, y planificación de la interconexión con el SEN. Las perspectivas de implementación identifican cinco desafíos técnicos prioritarios cuya resolución es condición necesaria para la transición del entorno de investigación al entorno operativo. El rol de los LLMs como capa de interpretación y comunicación se fundamenta ahora en un cuerpo creciente de literatura que documenta implementaciones concretas [26][27][28].

En síntesis, este capítulo cierra la tesis demostrando que la aplicación de técnicas de aprendizaje automático sobre datos operativos reales de la Central Fotovoltaica de Puerto Peñasco no solo es técnicamente viable sino que produce resultados de una calidad que justifica plenamente su adopción progresiva como herramienta de apoyo a la toma de decisiones, contribuyendo así al objetivo más amplio de modernizar la operación del sistema eléctrico mexicano mediante la integración de analítica avanzada e inteligencia artificial.

BIBLIOGRAPHY

[1] M. Sánchez Gutiérrez, "Pipeline de análisis exploratorio, preprocesamiento, selección de características y modelado predictivo para la Central Fotovoltaica de Puerto Peñasco," libreta Jupyter de autoría propia, 2025. [Reporte Propio 2025-BUAP-MSEP-R02].

[2] C. Pasion, T. Wagner, C. Koschnick, S. Schuldt, J. Williams, y K. Hallinan, "Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data," *Energies*, vol. 13, no. 10, p. 2570, 2020, doi: 10.3390/en13102570.

[3] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[4] C. Bergmeir y J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, May 2012, doi: 10.1016/j.ins.2011.12.028.

[5] S. Makridakis, E. Spiliotis, y V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLOS ONE*, vol. 13, no. 3, p. e0194889, 2018, doi: 10.1371/journal.pone.0194889.

[6] T. Hastie, R. Tibshirani, y J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2ª ed., Springer, Nueva York, USA, 2009.

[7] D. E. Rumelhart, G. E. Hinton, y R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986, doi: 10.1038/323533a0.

[8] D. Markovics y M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," *Renewable and Sustainable Energy Reviews*, vol. 161, p. 112364, 2022, doi: 10.1016/j.rser.2022.112364.

[9] M. Abdelsattar, A. AbdelMoety, y A. Emad-Eldeen, "Advanced machine learning techniques for predicting power generation and fault detection in solar photovoltaic systems," *Neural Computing and Applications*, vol. 37, pp. 8825–8844, 2025, doi: 10.1007/s00521-025-11035-6.

[10] S. Jogunuri *et al.*, "Random forest machine learning algorithm based seasonal multi-step ahead short-term solar photovoltaic power output forecasting," *IET Renewable Power Generation*, vol. 18, no. 2, pp. 234–248, 2025, doi: 10.1049/rpg2.12921.

[11] S. T. Asiedu, S. Gyamfi, E. A. Ofosu, y E. Djordjevic, "Machine learning forecasting of solar PV production using single and hybrid models over different time horizons," *Heliyon*, vol. 10, p. e28898, 2024, doi: 10.1016/j.heliyon.2024.e28898.

[12] M. N. Akhter *et al.*, "Towards energy efficiency: A comprehensive review of deep learning-based photovoltaic power forecasting strategies," *Heliyon*, vol. 10, no. 13, p. e33419, 2024, doi: 10.1016/j.heliyon.2024.e33419.

[13] E. Moufid, H. E. Khouya, y A. Berrahal, "Machine learning-based short-term solar power forecasting: a comparison between regression and classification approaches using extensive Australian dataset," *Sustainable Energy Research*, vol. 11, no. 1, p. 29, 2024, doi: 10.1186/s40807-024-00115-1.

[14] A. Aziz, N. Mufti, H. M. Khalid, H. Farooq, S. H. Abbasi, y Q. Khan, "Short-term and long-term solar irradiance forecasting with advanced machine learning techniques in Zafarana, Egypt," *Scientific Reports*, vol. 15, p. 3891, 2025, doi: 10.1038/s41598-025-24853-4.

[15] N. N. Barakat, A. K. Eltamaly, A. M. Alharthi, y M. S. Al-Saud, "Multi-label machine learning for power forecasting of a grid-connected photovoltaic solar plant over multiple time horizons," *Scientific Reports*, vol. 15, p. 10213, 2025, doi: 10.1038/s41598-025-20251-y.

[16] V. Raj, S. Dotse, M. Sathyajith, M. Petra, y H. Yassin, "Ensemble machine learning for predicting the power output from different solar photovoltaic systems," *Energies*, vol. 16, no. 2, p. 671, 2023, doi: 10.3390/en16020671.

[17] O. A. Ogunleye *et al.*, "Using Machine Learning Algorithms to Forecast Solar Energy Power Output," *Electronics*, vol. 14, no. 5, p. 866, 2025, doi: 10.3390/electronics14050866.

[18] J. Huang *et al.*, "Compendium of degradation rates of global photovoltaic (PV) technology: insights from technology, climate and geography," *Solar Energy Materials and Solar Cells*, vol. 285, p. 113245, 2025, doi: 10.1016/j.solmat.2025.113245.

[19] M. Straubmueck *et al.*, "Determinants of the long-term degradation rate of photovoltaic modules: A meta-analysis," *Renewable and Sustainable Energy Reviews*, vol. 215, p. 115367, 2025, doi: 10.1016/j.rser.2025.115367.

[20] S. Nair *et al.*, "Model Monitoring, Data Drift Detection, and Efficient Model Retraining: A Review," preprint, 2025. [En línea]. Disponible: `https://www.researchgate.net/publication/395703466`

[21] R. Zink, B. Ioshchikhes, y M. Weigold, "Concept drift monitoring for industrial load forecasting with artificial neural networks," *Procedia CIRP*, vol. 130, pp. 466–471, 2024, doi: 10.1016/j.procir.2024.10.078.

[22] M. Leyli-abadi, A. Marot, y J. Picault, "Data Drift Detection and Assessment for AI-hybrid Models Applied on Electrical Energy Consumption," en *Proceedings of the AAAI Symposium Series*, 2025.

[23] Y. Renkema, L. Visser, y T. AlSkaif, "Enhancing the reliability of probabilistic PV power forecasts using conformal prediction," *Solar Energy Advances*, vol. 4, p. 100059, 2024, doi: 10.1016/j.seja.2024.100059.

[24] T. D. Nguyen *et al.*, "Comparative analysis of conformal prediction techniques and machine learning models for very short-term solar power forecasting," *Machine Intelligence Research*, 2025, doi: 10.1016/j.mlwa.2025.100597.

[25] G. Wang, Y. Zhou, Y. Yan, Z. Zhou, Z. Yang, L. Dai, y J. Huang, "Probabilistic Photovoltaic Power Forecasting with Reliable Uncertainty Quantification via Multi-Scale Temporal–Spatial Attention and Conformalized Quantile Regression," *Sustainability*, vol. 18, no. 2, p. 739, 2026, doi: 10.3390/su18020739.

[26] S. Majumder *et al.*, "Exploring the capabilities and limitations of large language models in the electric energy sector," *Joule*, vol. 8, pp. 1544–1549, 2024, doi: 10.1016/j.joule.2024.03.012.

[27] Z. Chen *et al.*, "Large Language Models Meet Energy Systems: Opportunities, Challenges, and Future Perspectives," *Applied Energy*, 2025, doi: 10.1016/j.apenergy.2025.125569.

[28] Y. Chen *et al.*, "A large language model for advanced power dispatch," *Scientific Reports*, vol. 15, p. 8926, 2025, doi: 10.1038/s41598-025-91940-x.

[29] G. Buster, "Large Language Models (LLMs) for Energy Systems Research," National Renewable Energy Laboratory (NREL), Tech. Rep. NREL/TP-6A20-87896, 2023.

[30] A. Bernadić *et al.*, "Large Language Models in Power Systems," *International Journal of Innovative Science and Engineering*, vol. 1, no. 1, 2025.

[31] T. Kim, H. Li, R. Bhatt, y B. Palmintier, *Generative AI for Power Grid Operations*, National Renewable Energy Laboratory, Golden, CO, USA, Tech. Rep. NREL/TP-5D00-91176, 2024.

[32] B. B. Abubakar, C. M. L. Ugwuanyi, y M. N. Olutola, "Deep learning models for solar irradiance forecasting: A comprehensive review," *Renewable Energy Focus*, vol. 44, pp. 176–196, 2023, doi: 10.1016/j.ref.2023.01.003.