Universidad Autónoma del Estado De Hidalgo

Instituto de Ciencias Básicas e Ingeniería

Área Académica de Computación y Electrónica

Modelado y predicción inteligente de producción fotovoltaica

TESIS

Para obtener el título de

Licenciatura en Ciencias Computacionales

PRESENTA

Gándara Ambriz Carlos Alejandro

DIRECTORES DE TESIS

Dr. Virgilio López Morales

Dr. Aldo Marquez Grajales

Ciudad del Conocimiento, Pachuca de Soto, 2025

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences

Mineral de la Reforma, Hgo., a 20 de junio de 2025

Número de control: ICBI-D/1017/2025
Asunto: Autorización de impresión.

MTRA. OJUKY DEL ROCÍO ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

Con Título Quinto, Capítulo II, Capítulo V, Artículo 51 Fracción IX del Estatuto General de nuestra Institución, por este medio, le comunico que el Jurado asignado del egresado de la Licenciatura en Ciencias Computacionales **Carlos Alejandro Gándara Ambriz,** quien presenta el trabajo de titulación **"Modelado y predicción inteligente de producción fotovoltaica",** ha decidido, después de revisar fundamento en lo dispuesto en el Título Tercero, Capítulo I, Artículo 18 Fracción IV; dicho trabajo en la reunión de sinodales, **autorizar la impresión del mismo**, una vez realizadas las correcciones acordadas.

A continuación, firman de conformidad los integrantes del Jurado:

Presidente: Dra. Anilú Franco Árcega

Secretario: Dr. Manuel Alejandro Ojeda Misses

Vocal: Dr. Virgilio López Morales

Suplente: Dr. Aldo Márquez Grajales

Sin otro particular por el momento, reciba un cordial saludo.

Atentamente
"Amor, Orden y Progreso"

Mtro. Gabriel Vergara Rodríguez
Director del ICBI

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
ICBI
DIRECCIÓN

GVR/YCC

Ciudad del Conocimiento, Carretera Pachuca-Tulancingo Km. 4.5 Colonia Carboneras, Mineral de la Reforma, Hidalgo, México. C.P. 42184
Teléfono: 771 71 720 00 Ext. 40001
direccion_icbi@uaeh.edu.mx, vergarar@uaeh.edu.mx

Resumen

La presente tesis aborda el problema de la variabilidad en la producción de energía fotovoltaica, causada por factores climáticos impredecibles como lo son radiación solar, temperatura, humedad, velocidad, dirección del viento, nubosidad, entre otros. Para mejorar la eficiencia en la administración de la energía generada, se propone el desarrollo de un sistema de predicción basado en modelos de inteligencia artificial. Se emplea la metodología SEMMA(Sample, Explore, Modify, Model, Assess) para la recolección, preprocesamiento y análisis de los datos, utilizando información proveniente de la NASA y de una planta fotovoltaica.

Se implementan y comparan tres modelos de predicción: Bosques aleatorios (RF, por sus siglas en inglés), Support Vector Machines (SVM) y Long Short-Term Memory (LSTM). La evaluación se realiza mediante la métrica de Error Cuadrático Medio (MSE), determinando que el modelo LSTM presenta un mejor desempeño en la predicción de series temporales.

This thesis addresses the problem of variability in photovoltaic energy production, caused by unpredictable climatic factors such as solar radiation, temperature, and humidity to improve efficiency in energy management, the development of a prediction system based on artificial intelligence models is proposed. The SEMMA methodology is employed for data collection, preprocessing, and analysis, using information from NASA and a photovoltaic plant.

Three prediction models are implemented and compared: Random Forest (RF), Support Vector Machines (SVM), and Long Short-Term Memory (LSTM). The evaluation is conducted using the Mean Squared Error (MSE) metric, determining that the LSTM model performs best in time series prediction.

Agradecimientos

Gracias a todos los profesores que compartieron sus enseñanzas conmigo durante estos años y a mi asesor que compartió sus conocimientos y guió a través de los estudios necesarios para hacer esto posible.

Gracias a mis familiares por su apoyo, mis abuelos Jorge Ambriz Román, Claudia Catalina Sánchez Silva y Elvira Balderas Severino que siempre tuvieron palabras de aliento, mis padres y hermano Carlos Gándara Balderas, Claudia Ambriz Sánchez y Juda Israel Gándara Ambriz que mostraron siempre su paciencia y apoyo durante la realización de este proyecto y a mi pareja Lesly Fernanda Salinas Palencia quien creyó en mi incluso cuando yo mismo dudaba de mis capacidades y me impulso a seguir sin importar las adversidades, siempre estuvo reconfortándome en momentos difíciles además de estar siempre alentándome en cada uno de mis avances, inspirándome de las maneras más tiernas y cálidas que jamás habría podido imaginar y mostrándose como uno de mis grandes pilares, no solo en ámbito académico si no también personal.

Finalmente hacer una mención para mis pequeños apoyos peludos, que me acompañaron innumerables veces en desvelos, frustraciones y alegrías durante este proyecto, Kiba, Kuro, Kirara y Kaida.

Sin toda su ayuda, apoyo, paciencia, conocimientos y consejos nada de esto sería posible.

Eternamente gracias.

“El conocimiento no es propiedad de nadie, sino un bien que debe compartirse”

Dedicatoria

Para mi amor eterno y futura familia,

Que el amor y la felicidad sean siempre el motor de sus vidas.

Índice

INTRODUCCIÓN 1

1.1 Planteamiento del problema 4
1.3 Objetivos 4
1.3.1 Objetivo general 5
1.3.1 Objetivos específicos 5
1.4 Hipótesis 6
1.5 Estructura y desarrollo del documento 6

MARCO TEÓRICO 8

2.1 Marco conceptual 8
2.1.1 Energía fotovoltaica y su relación con factores climatológicos 8
2.1.2 Modelos inteligentes aplicados 10
2.2 Marco metodológico 25
2.2.1 Metodologías para minería de datos 26
2.2.2 Métrica de error cuadrático medio 34
2.2.3 Normalización de datos 35
2.3 Marco tecnológico 37

ESTADO DEL ARTE 38

3.1 Predicciones para dar la bienvenida a la era solar. 38
3.2 Predicción de la energía solar a través del aprendizaje automático. 39
3.3 Modelos de Aprendizaje Profundo para la Predicción de Energía Fotovoltaica 40
3.4 Un Modelo Híbrido de Aprendizaje Automático para la Predicción de Energía Solar 44

IMPLEMENTACIÓN 45

4.1 Configuración del experimento. 45
4.2 Arquitectura del sistema 46
4.3 Tratamiento de datos 49
4.4.1 Definición y clasificación de datos 49
4.4.1 Procesamiento de los datos para su estandarización 51
4.4.2 Normalización 52
4.5 Aplicación de modelos 54
4.5.1 Implementación de Bosques Aleatorios 55

4.5.2 Implementación de SVM 57
4.5.3 Implementación de LSTM 59
4.6 Comparación de resultados 61
4.7 Análisis 76

CONCLUSIONES 78

5.1 Conclusiones 78
5.2 Perspectivas del trabajo 80
5.3 Trabajos futuros 81

REFERENCIAS 82

Capítulo I

Introducción

Las energías limpias se han convertido en los últimos años en una de las fuentes de generación eléctrica más importantes alrededor del mundo. Esto se debe a una combinación de factores clave: la gran inversión destinada a proyectos de infraestructura energética sostenible, los incentivos y subsidios gubernamentales que han fomentado su adopción, y los compromisos internacionales para mitigar el cambio climático, como los establecidos en el Acuerdo de París el cual fue aprobado en la COP21 en París en 2015 y entró en vigor en 2016 (Organizacion de las Naciones Unidas [Informe], 2021). Estas iniciativas han hecho que su uso sea cada vez más extendido en distintos sectores, desde la industria y el transporte hasta el uso doméstico.

Además, el avance tecnológico ha desempeñado un papel fundamental al permitir una mayor eficiencia en cuanto al almacenamiento, conversión y distribución de la energía generada. Innovaciones como las baterías de última generación, los sistemas inteligentes de gestión de redes eléctricas y las mejoras en la eficiencia de los paneles solares y turbinas eólicas han hecho que estas tecnologías sean más accesibles y viables económicamente. Todo esto, como se describe en (Visser, 2023), ha llevado a una transformación significativa en el panorama energético global, consolidando a las energías limpias como una alternativa cada vez más competitiva frente a los combustibles fósiles (Visser, 2023).

Éstas iniciativas no solo contribuyen a reducir las emisiones de gases de efecto invernadero, sino que también son una herramienta clave para la electrificación de zonas rurales y remotas, donde el acceso a la red eléctrica convencional es limitado o inexistente y las plantas de energías alternativas y los sistemas de microrredes pueden

proporcionar una fuente confiable y sostenible de energía. El acceso a servicios básicos como la iluminación, la refrigeración de alimentos y medicamentos, y la operación de herramientas y equipos esenciales, influye directamente en la mejora de la calidad de vida de los usuarios.

Desde la perspectiva de la sustentabilidad, las energías limpias se alinean con los Objetivos de Desarrollo Sostenible (ODS) de la Organización de las Naciones Unidas (ONU), especialmente con el ODS 7, que busca garantizar el acceso a una energía asequible, segura, sostenible y moderna para todos (Organizacion de las Naciones Unidas [Informe], 2021). Asimismo, su implementación en zonas rurales promueve el desarrollo económico y social de comunidades marginadas, reduciendo la brecha de desigualdad energética.

Entre las múltiples opciones con las que se cuenta para la generación de energía limpia se destaca la generación de energía fotovoltaica, en la cual se centrará el trabajo descrito en este documento. Ésta es un tipo de generación eléctrica emergente, que desde 2021 ha tenido un aumento significativo en su uso gracias a la introducción de nuevas tecnologías y la reducción en sus costes de producción y mantenimiento (IEA PVPS, 2022).

Con base a los datos recopilados en (IEA PVPS, 2022) se destaca que a nivel mundial ha aumentado la capacidad de instalación de plantas fotovoltaica en un 22 %, donde China, Europa y USA, son principales países con mayor instalación de plantas por año, como se observa en la Figura 1. En consecuencia, dichos países son los de mayor producción por año y se vislumbra una apertura mayor en la instalación de estas en Latinoamérica en países como: México, Brasil, Chile y Honduras.

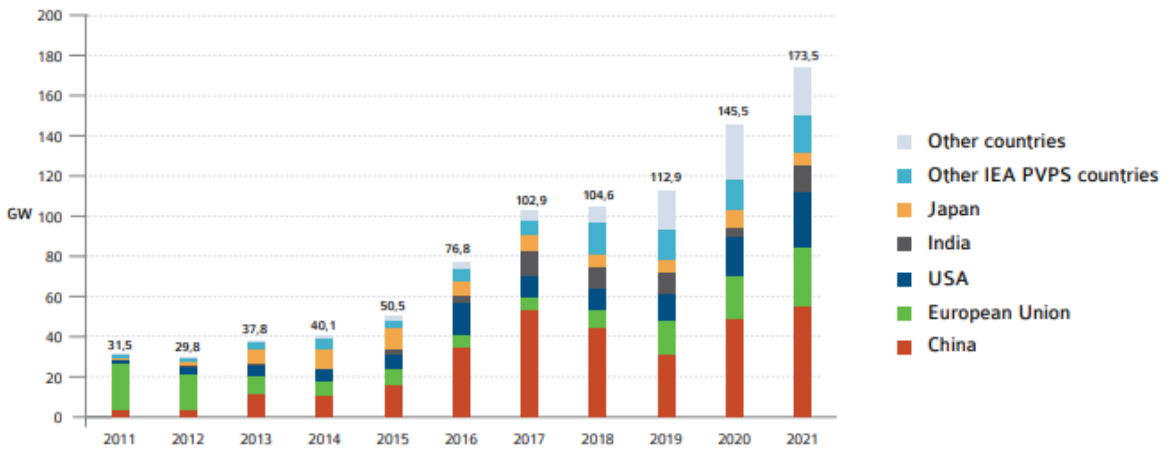


Figura 1 Evolución anual de PV instalados. Obtenida de (IEA PVPS, 2022)

Pero … ¿Cómo funciona la energía fotovoltaica?

Ésta se basa en la extracción de energía de la radiación solar a través de celdas fotovoltaicas, la productividad de estas células irá variando respecto a diversos factores climáticos tales como la nubosidad, la temperatura o la humedad entre otras que se tratarán más adelante.

Estas variables físicas son dinámicas y la productividad de una planta fotovoltaica puede verse afectada con una producción poco fiable, impidiendo la correcta administración de energía. Este problema tiene implicaciones significativas tanto económicas como ambientales. Una producción energética ineficiente puede provocar sobrecostos al no maximizar el aprovechamiento de los recursos disponibles y, a nivel ambiental, puede limitar el impacto positivo de una transición completa hacia energías limpias, al seguir dependiendo de fuentes no renovables para cubrir la demanda energética.

Es por esto por lo que la implementación de sistemas inteligentes capaces de realizar predicciones y administrar la energía producida basándose en las variables anteriormente descritas tal como se muestra en (Visser, 2023).

En este trabajo se proponen un conjunto de métodos de aprendizaje automático para modelar la predicción de la producción de energía de una planta fotovoltaica, al

evaluar el desempeño de los diversos modelos para encontrar el óptimo en las predicciones de la producción de energía de la planta fotovoltaica.

1.1 Planteamiento del problema

La variabilidad en la producción energética de plantas fotovoltaicas representa un desafío significativo para su integración eficiente en el sistema eléctrico. Esta variabilidad se debe a factores climáticos impredecibles como nubosidad, temperatura y velocidad del viento, que afectan directamente la generación de energía. Para abordar este problema, se ha propuesto implementar sistemas inteligentes que permitan predecir de manera más precisa la producción energética, optimizando así su administración.

Por lo tanto, el identificar uno o varios modelos inteligentes que mejor se adapte a este propósito es crucial, ya que un sistema eficiente y fiable en sus predicciones no solo mejorará la administración de la energía, sino que también proporcionará un avance innovador al sector energético, aumentando la viabilidad de esta tecnología y su penetración en el mercado.

1.3 Objetivos

La presente investigación tiene como finalidad la implantación de modelos inteligentes que permitan predecir con mayor precisión la producción energética de una planta fotovoltaica, considerando variables climatológicas. Para ello, se establece un objetivo general que orienta el enfoque principal del estudio, acompañado de objetivos específicos que desglosan las acciones necesarias para alcanzar dicha meta.

1.3.1 Objetivo general

Implementar un conjunto de modelos inteligentes entrenados para realizar predicciones precisas sobre el volumen de producción de energía de una planta fotovoltaica, asegurando su óptimo desempeño a través de técnicas de minería de datos y aprendizaje automático

1.3.1 Objetivos específicos

I. Recopilar las bases de datos relevantes provenientes de fuentes confiables, buscando en fuentes confiables que sigan criterios de selección para garantizar la representatividad y calidad de los datos.

II. Analizar los datos recopilados, identificando patrones, tendencias, correlaciones y posibles anomalías, para identificar las variables más relevantes que influyen en la producción fotovoltaica.

III. Preprocesar los datos con procesos de normalización de datos para preparar adecuadamente las bases de datos para su uso en los modelos predictivos.

IV. Implementar modelos inteligentes de predicción, utilizando los datos ya procesados para estimar la producción energética de la planta.

V. Evaluar el desempeño de los modelos desarrollados mediante métricas de evaluación, comparando resultados y determinando cuál de los modelos ofrece la mayor precisión y eficiencia en la predicción para identificar el modelo óptimo para la predicción de energía fotovoltaica.

Una vez establecidos los objetivos específicos que guiarán el desarrollo del presente trabajo, es importante contextualizar la fundamentación teórica que respalda el uso de modelos inteligentes para la predicción de la producción energética. En particular, los

modelos inteligentes basados en datos, como "*Long Short-Term Memory*" (LSTM), Maquinas de Soporte Vectorial (SVM) y Bosques aleatorios (RF), son herramientas efectivas para, en este caso, poder predecir la producción de energía de una planta fotovoltaica al aprovechar datos meteorológicos y de producción previamente recolectados. Esto se fundamenta en estudios previos que demuestran la capacidad de los modelos basados en aprendizaje automático para identificar patrones complejos en datos no lineales y secuenciales.

Con esta información presente se puede plantear la siguiente hipótesis:

1.4 Hipótesis

Al implementar algunas técnicas de análisis y minería de datos y los algoritmos de aprendizaje automático se encontrará un modelo que ayude en la gestión de una planta fotovoltaica al realizar predicciones precisas sobre el volumen de producción de energía fotovoltaica, y en consecuencia facilitar la mejor administración de una red de suministro de energías limpias.

1.5 Estructura y desarrollo del documento

El problema planteado será abordado mediante el diseño, implementación y evaluación de un sistema capaz de procesar datos climatológicos y de producción energética. A través de este sistema, se implementarán diversos modelos de aprendizaje automático, como LSTM, SVM y Bosques aleatorios, para realizar predicciones de producción fotovoltaica. Los datos serán preprocesados, normalizados y alimentados en los modelos para entrenarlos y evaluar su precisión, utilizando métricas específicas como el Error Cuadrático Medio (MSE).

A lo largo de esta tesis, se espera establecer una solución eficiente y basada en datos para la predicción de la producción fotovoltaica. Al final del documento, se

identificarán las ventajas y limitaciones de los modelos inteligentes estudiados, con el objetivo de elegir el más adecuado y contribuir al desarrollo de tecnologías aplicadas al sector energético.

Este documento se organiza en cinco capítulos principales:

- **Capítulo I: Introducción.** Se presentan el contexto, el planteamiento del problema, los objetivos y la hipótesis del trabajo.
- **Capítulo II: Estado del arte.** revisión a las principales influencias para este proyecto y el contexto actual de los temas que atañen a este.
- **Capitulo III: Marco teórico.** Se describen las bases conceptuales, metodologías y tecnológicas utilizada en este trabajo.
- **Capítulo IV: Implementación.** Se detallan los requerimientos, la arquitectura del sistema y el proceso de implementación, así como las funcionalidades clave de la aplicación desarrollada.
- **Capítulo V: Análisis y Conclusiones.** Se analizan los resultados obtenidos al evaluar los modelos, se establecen conclusiones y se plantean perspectivas futuras para mejorar el sistema y expandir su alcance.

Capítulo II

Marco teórico

El presente capítulo tiene como objetivo establecer los fundamentos conceptuales, técnicos y metodológicos que sustentan el desarrollo del sistema de predicción de energía fotovoltaica. Se exploran los principios del funcionamiento de las plantas solares, así como los factores ambientales que influyen en su eficiencia. Asimismo, se realiza una revisión de los principales modelos de aprendizaje automático utilizados en la predicción energética, incluyendo Máquinas de soporte vectorial (SVM), Bosques aleatorios y Redes neuronales. Esta base teórica permite comprender las herramientas empleadas y justifica su selección para abordar el problema de predicción abordado en esta investigación.

2.1 Marco conceptual

En este apartado se definen los conceptos clave que sustentan el desarrollo de este trabajo, tales como la energía fotovoltaica, el efecto de los factores climatológicos sobre su rendimiento, y los modelos de inteligencia artificial aplicados a la predicción energética. La comprensión clara de estos conceptos es esencial para contextualizar el enfoque metodológico adoptado en esta investigación.

2.1.1 Energía fotovoltaica y su relación con factores climatológicos

Para poder encontrar una solución que mejore la eficiencia de las plantas fotovoltaicas, primero se debe comprender cómo estas plantas producen energía eléctrica y cómo se ven afectadas por los factores climatológicos.

La energía fotovoltaica es una forma de generación eléctrica que utiliza la radiación solar como fuente primaria. Este tipo de energía se obtiene a través de dispositivos llamados paneles solares, los cuales convierten directamente la luz solar en electricidad mediante un proceso basado en el efecto fotoeléctrico. Este fenómeno pudo ser aprovechado tras décadas de avances tecnológicos que permitieron hacer de esta una alternativa viable y competitiva para la generación de energía (IEA PVPS, 2022).

Los paneles solares están compuestos por múltiples celdas fotovoltaicas, generalmente fabricadas a partir de materiales semiconductores como el silicio. Estas celdas capturan los fotones de la luz solar, generando un flujo de electrones que produce una corriente eléctrica. Existen diferentes tipos de paneles solares, entre los que destacan los monocristalinos, policristalinos y de película delgada, cada uno con características específicas en términos de eficiencia, durabilidad y costos de fabricación (Visser, 2023).

Una de las principales ventajas de la tecnología fotovoltaica es su modularidad, ya que los paneles pueden instalarse en una amplia variedad de ubicaciones, desde pequeñas viviendas hasta grandes plantas solares. Sin embargo, su productividad está influenciada por diversos factores ambientales, como la radiación solar disponible, la temperatura ambiente, la acumulación de suciedad en las celdas y las condiciones climáticas locales (Visser, 2023).

Actualmente, el desarrollo de tecnologías avanzadas ha mejorado tanto la eficiencia como la accesibilidad de los paneles solares. Innovaciones como los paneles bifaciales, que capturan la radiación en ambas caras, y el uso de nuevos materiales, han incrementado significativamente la capacidad de generación (IEA PVPS, 2022). Estas mejoras han permitido que la energía fotovoltaica desempeñe un papel fundamental en la transición energética global, tanto en grandes instalaciones como en aplicaciones urbanas.

2.1.2 Modelos inteligentes aplicados

Los modelos inteligentes han cobrado relevancia en múltiples disciplinas debido a su capacidad para resolver problemas complejos mediante el análisis de grandes volúmenes de datos. En esta sección se abordan diversas aplicaciones de algoritmos como Máquinas de Soporte Vectorial, Bosques Aleatorios y Redes Neuronales en contextos que van desde el sector energético hasta el procesamiento del lenguaje natural y la clasificación de patrones.

2.1.2.1 Modelos de predicción basados en SVM

La máquina de soporte vectorial (SVM) es una técnica de aprendizaje supervisado utilizado para clasificación y regresión. Se enfoca en encontrar una separación en un plano de datos que sea capaz de clasificarlos de manera lineal como se muestra en Figura 2, esta separación es conocida como “hiperplano”.

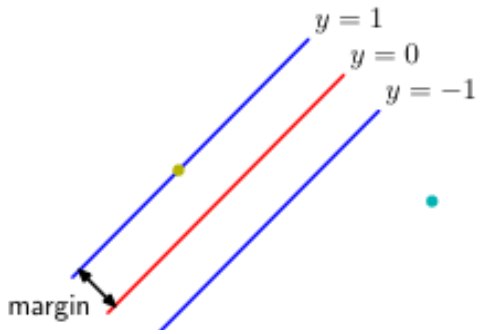


Figura 2 Imagen de un hiperplano resaltando los Vectores de soporte y su margen. obtenida de (Bishop, 2006)

El hiperplano es una frontera de decisión que divide el espacio de datos en regiones que corresponden a diferentes clases, que podemos representar de la siguiente forma:

Dado un conjunto de datos $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, donde x_n son los vectores de entrada y $y_n \in \{-1,1\}$ son las etiquetas de clase, el objetivo es encontrar un hiperplano que separe las clases de manera óptima. El hiperplano se define con la Ecuación (8).

(8)

$$w * x + b = 0$$

Dónde w es el vector normal al hiperplano, b es el término de sesgo que determina la posición del hiperplano en el espacio y x es el vector de características de los datos.

El criterio fundamental para seleccionar el hiperplano óptimo es que maximice el margen, es decir, la distancia mínima entre el hiperplano y los vectores de soporte, definido con la Ecuación (9).

$$M = \frac{1}{\|w\|} \qquad (9)$$

Dónde M es el margen, w es la magnitud del vector.

Los vectores de soporte son los puntos de datos más cercanos al hiperplano, y son fundamentales porque determinan su posición y orientación se encuentran a través de la Ecuación (10). Este concepto fue formalizado por Vapnik y Lerner en 1963 (Betancourt, 2005), y su importancia radica en que un mayor margen generalmente implica un mejor rendimiento de generalización del modelo, reduciendo la probabilidad de error en datos no vistos previamente (Bishop, 2006).

$$\frac{1}{2}\|w\|^2 \qquad (10)$$

Dónde w es la magnitud del vector.

Como se podrá pensar, este modelo parece sumamente efectivo en casos linealmente separables , donde los datos son fácilmente clasificables, sin embargo, en situaciones donde los datos no son linealmente separables en su espacio original, las SVM pueden

seguir siendo efectivas mediante el uso de una técnica conocida como "*Kernel Trick*". Esta técnica permite transformar el espacio original de los datos en un espacio de mayor dimensionalidad, donde es más probable que los datos sean linealmente separables (León E. C., 2016).

Este modelo ha demostrado ser extremadamente útiles en una amplia variedad de aplicaciones, desde el reconocimiento de patrones y la bioinformática hasta la detección de intrusos y el diagnóstico médico. Por ejemplo, en el diagnóstico de cáncer de mama, las SVM pueden ser entrenadas para clasificar tumores como malignos o benignos, proporcionando una herramienta poderosa para los médicos en la toma de decisiones clínicas como se puede ver en (León E. C., 2016).

El diagrama de flujo (Figura 3) describe el proceso completo de un modelo SVM, desde la carga y preparación de los datos hasta la evaluación y predicción. Cada etapa es crucial para garantizar que el modelo sea preciso y generalice bien a nuevos datos. El SVM es un algoritmo poderoso para problemas de clasificación y regresión, especialmente cuando los datos no son linealmente separables, gracias al uso de kernels.

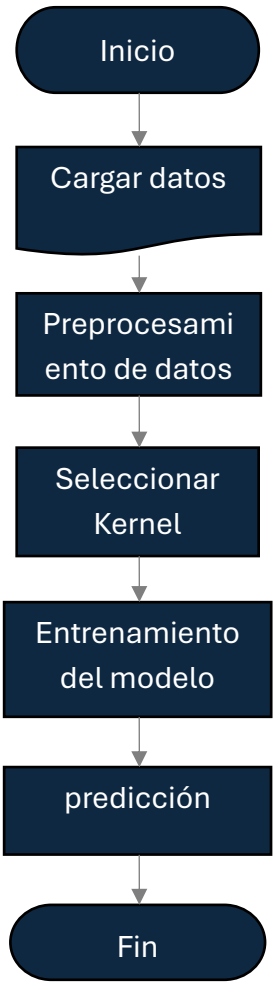


Figura 3 Diagrama de flujo para SVM

1. **Cargar datos:** comenzamos cargando el conjunto de datos que se utilizará para entrenar y evaluar el modelo. Este dataset debe contener características (variables independientes) y etiquetas (variable dependiente).

2. **Preprocesamiento de datos:** los datos se preparan para ser utilizados por el modelo. Esto incluye:
    - Limpieza (eliminar valores nulos o duplicados).
    - Normalización o estandarización (escalar las características para que tengan un rango similar).
    - Codificación de variables categóricas (si es necesario).
    - División del dataset en conjuntos de entrenamiento y prueba.

3. **Seleccionar el kernel:** se elige el tipo de kernel que se utilizará en el SVM. Los kernels más comunes son: lineal, polinomial, RBF (Radial Basis Function) y sigmoide. La elección del kernel depende de la naturaleza de los datos y del problema a resolver.

4. **Entrenar el modelo SVM:** el modelo se entrena utilizando el conjunto de datos de entrenamiento. Durante este proceso, el SVM busca encontrar el hiperplano óptimo que separa las clases (en clasificación) o ajusta los datos (en regresión). Se configuran hiperparámetros como el parámetro C (controla el trade-off entre maximizar el margen y minimizar el error) y gamma (afecta la influencia de cada punto de datos).

5. **Predicción:** finalmente, el modelo entrenado y validado se utiliza para hacer predicciones sobre nuevos datos. Esto es especialmente útil en aplicaciones prácticas donde se necesita clasificar o predecir resultados.

2.1.2.1.1 Funciones Kernel

Como se mencionó anteriormente el modelo de máquina de soporte vectorial tiene problemas ante casos con soluciones no lineales, para estos casos se busca redimensionar el plano donde se encuentran los datos a uno de mayor dimensión así logrando una separación de estos y así conseguir una solución lineal.

Para realizar esta redimensión del plano es necesario utilizar funciones "*Kernel*", siendo la elección de la función crucial en el funcionamiento del modelo, estas permiten calcular el producto interno de un espacio de características de alta dimensión sin necesidad de hacer una transformación explicita como lo muestra (Bishop, 2006).

Una función *Kernel* es esencialmente una técnica matemática que calcula el producto interno entre las imágenes de los datos en este nuevo espacio de alta dimensionalidad, sin necesidad de realizar explícitamente la transformación a dicho espacio. Esta capacidad de trabajar de manera implícita con el espacio de características permite que las SVM manejen eficientemente conjuntos de datos complejos sin el costo computacional tan alto que implicaría calcular directamente las coordenadas en un espacio de mayor dimensión (Betancourt, 2005).

Entre las funciones *Kernel* más utilizadas se encuentran:

- **Kernel Lineal**: para datos que son aproximadamente linealmente separables en su forma original. Calcula el producto interno estándar y es útil cuando el número de características es muy grande, como se muestra en la Ecuación (11).

$$K(x, xi) = \sum (x * xi) \qquad (11)$$

  Dónde $K(x, x_i)$ es la función kernel lineal, x es el vector de características de entrada, x_i es el vector de características del conjunto de entrenamiento, y la suma representa el producto punto que mide la similitud entre ambos vectores

- **Kernel Polinómico**: define la separación entre clases mediante un polinomio de grado *n*, lo que permite capturar relaciones no lineales más complejas, esto se muestra en la Ecuación (12).

$$K(x, xi) = 1 + \sum (x * xi)^d \qquad (12)$$

- **Kernel Gaussiano o Radial (*RBF*)**: es uno de los más populares debido a su capacidad para manejar una amplia variedad de datos. Este *Kernel* mide la

similitud entre dos puntos, lo que permite capturar relaciones no lineales muy complejas en los datos originales mostrado en la Ecuación (13).

$$K(x, xi) = exp\ (-gamma * \sum (x - xi)^2) \quad (13)$$

El uso adecuado del *Kernel* no solo depende de la naturaleza de los datos, sino también de una selección cuidadosa de sus parámetros.

Una vez se tiene e hiperplano más eficiente se entrena el modelo y este será capaz de realizar la predicción de un nuevo dato evaluándolo con el signo de la funcion de decisión, la cual consiste en una combinación lineal de los productos "*Kernel*" entre el nuevo dato y los vectores de soporte.

2.1.2.2 Modelos de Bosques aleatorios (Random forest)

El modelo Bosques aleatorios, introducido por (Breiman, 2001), es una extensión de los árboles de decisión que utiliza algoritmos de aprendizaje supervisado. Su objetivo principal es mejorar la precisión de las predicciones y reducir la varianza, problemas comunes en los árboles de decisión individuales. Este modelo es ampliamente utilizado tanto para tareas de clasificación como de regresión, destacándose por su robustez, precisión y capacidad para manejar conjuntos de datos complejos.

El método se basa en la creación de múltiples árboles de decisión independientes durante el proceso de entrenamiento, como se ilustra en la Ecuación (4). Cada árbol se construye utilizando un subconjunto diferente del conjunto de datos original, generado mediante una técnica conocida como "*Bootstrap Aggregating*" o "*Bagging*". Este enfoque, descrito por Breiman en 1996 (Breiman, 2001), implica el muestreo con reemplazo, lo que significa que algunos datos pueden aparecer más de una vez en una muestra, mientras que otros pueden no aparecer en absoluto (León M. G., 2018).

$$T = \{Árbol_1, Árbol_2, \dots, Árbol_k\} \quad (4)$$

Adicionalmente, en cada nodo de un árbol de decisión, el modelo selecciona un subconjunto aleatorio de características (variables) para determinar las divisiones. Este procedimiento tiene dos objetivos clave:

- Reducir la correlación entre los árboles individuales.
- Evitar el sobreajuste, mejorando así la capacidad de generalización del modelo.

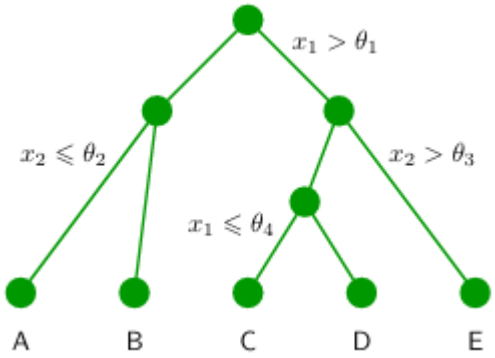


Figura 4 Diagrama de árbol binario. obtenida de (Bishop, 2006)

Cada árbol se crea a través de una serie de particiones llamadas "*splits*" como se ve en la Figura 4, estas particiones darán origen a diferentes nodos que irán dividiendo el espacio en dos regiones dependiendo de las variables, buscando su clasificación y al finalizar el proceso se tendrá un árbol completo. Los "*splits*", están determinados por la Ecuación (5) (León M. G., 2018).

(5)

$$x_i \ll \theta_k$$

Dónde de x_i son las variables de entrada y θ_k es el valor umbral.

La construcción final del modelo implica el entrenamiento de tantos árboles de decisión como sea necesario. En general, un mayor número de árboles mejora la precisión del modelo, pero a costa de un mayor tiempo de entrenamiento y un incremento en los recursos computacionales requeridos.

Bosques aleatorios

En los Bosques aleatorios, las predicciones finales se obtienen combinando los resultados de los árboles individuales:

- En problemas de clasificación, se realiza una votación mayoritaria entre los árboles donde se utiliza la Ecuación (6).

$$\hat{y} = argmax_c \sum_{i=1}^{N} I(T_{i1}(x) = c) \quad (6)$$

Dónde $\hat{y}$ es la clase predicha por el modelo, c es cada una de las posibles clases, $T_i(x)$ es la predicción del árbol i para la entrada x, I es la función indicadora que devuelve 1 si $T_i(x) = c$, y 0 en caso contrario y N es el número total de árboles en el bosque

- En problemas de regresión, se calcula el promedio de las predicciones individuales donde se aplica la Ecuación (7).

(7)

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_{i1}(x)$$

Dónde $\hat{y}$ es un valor predicho por el modelo, $T_i(x)$ es la predicción del árbol i para la entrada x y N es el número total de árboles en el bosque

Este enfoque de combinación reduce la varianza y mejora la estabilidad de las predicciones, abordando uno de los principales problemas de los árboles de decisión individuales. En la Figura 5 se muestra un diagrama de flujo sobre el funcionamiento del modelo de Bosques aleatorios y su tratamiento de datos.

El flujo dentro del diagrama de Bosques aleatorios describe cómo los datos se trabajan y avanzan por las diferentes etapas del modelo para llegar a una predicción final. A continuación, se detallan estas etapas específicas en el contexto del diagrama de flujo:

1. **Inicio**: se cargan los datos necesarios para el modelo.
2. **Procesamiento de los datos**: se preparan los datos, eliminando ruido y asegurando consistencia.
3. **Creación de subconjuntos (Bootstrap)**: se generan múltiples subconjuntos aleatorios de los datos originales mediante muestreo con reemplazo.
4. **Entrenamiento de árboles de decisión**: cada subconjunto entrena un árbol de decisión independiente.
5. **Predicciones individuales**: todos los árboles generan su propia predicción.
6. **Combinación de predicciones**: obtenemos los promedios de las predicciones (en regresión) o se aplica el voto mayoritario (en clasificación).
7. **Resultado final**: finalmente se obtiene la predicción final combinada.

Este proceso asegura que el modelo aproveche la diversidad de los árboles para mejorar la precisión y reducir el sobreajuste.

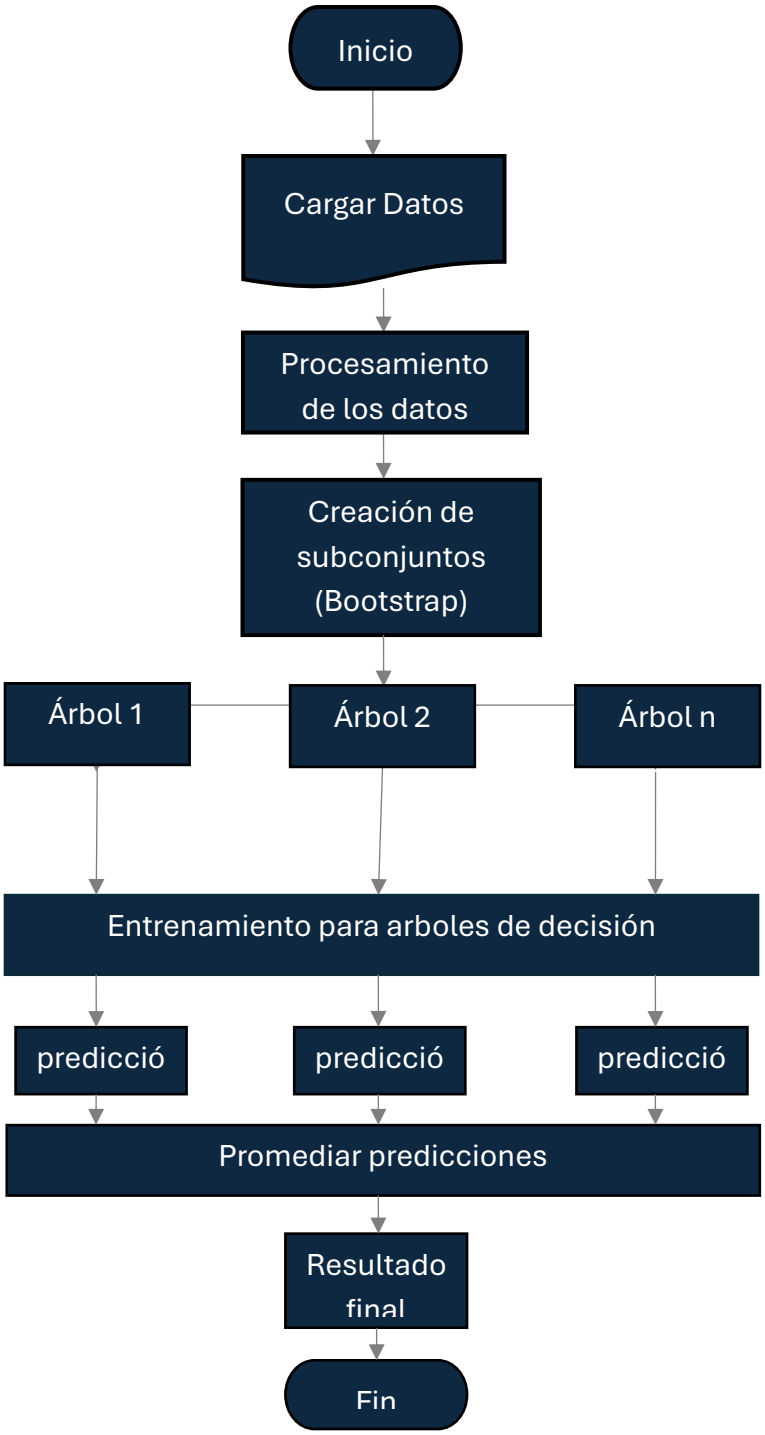


Figura 5 Diagrama de flujo sobre Bosques aleatorios

En el documento (Bueno, 2017) se puede ver el uso del modelo “Bosques aleatorios” para la predicción de los precios de la electricidad en el mercado, este tema cuenta con una gran volatilidad y complejidad. En la investigación, este modelo fue seleccionado debido a su capacidad para manejar grandes conjuntos de datos y múltiples variables explicativas, lo cual fue particularmente efectivo en este contexto porque permitió capturar las relaciones no lineales y complejas entre las variables, algo que sería difícil de lograr con modelos lineales tradicionales. Además, al utilizar un gran número de árboles en el bosque se pudo mejorar la estabilidad de las predicciones y minimizar el impacto de las fluctuaciones tan cambiantes en los datos de entrada.

2.1.2.3 Redes neuronales aplicadas

El modelo Long Short-Term Memory (LSTM) es un tipo especializado de red neuronal recurrente, este fue introducido por Sepp Hochreiter y Jürgen Schmidhuber en 1997, buscando solucionar el problema de desvanecimiento del gradiente que se encontraba presente en las redes neuronales recurrentes (RNN), este modelo está diseñado para modelar secuencias de datos donde se requiera el aprendizaje de dependencia a corto y largo plazo siendo esta su mayor característica y virtud respecto a otras redes neuronales recurrentes (Suárez, López, & Ramos, 2009).

Esto se debe a su estructura interna que permite mantener y actualizar estados de memoria a largo plazo, esta se divide en múltiples “*gates”*.

I. *Forget gate*: esta puerta decide qué información de la memoria a largo plazo debe ser olvidada o retenida. La decisión se basa en los datos de entrada actuales y la memoria previa. El valor de salida de esta puerta es un número entre 0 y 1, donde 1 significa "completamente mantener" y 0 significa "completamente olvidar". El objetivo de esta puerta es ayudar a la red a deshacerse de información irrelevante, se expresa con la Ecuación (14).

$$f_t = \sigma\left(W_f * [h_t - 1, x_t] + b_f\right) \qquad (14)$$

Dónde f_t es un vector de activación de la puerta de olvido en el instante t, σ es la función sigmoide, W_f es la matriz de pesos asociada a la puerta de olvido, $h_t - 1$ es el estado oculto del instante anterior, x_t es la entrada en el instante t y b_f es el sesgo de la puerta de olvido

II. *Input gate*: esta puerta regula qué nueva información será almacenada en la memoria a largo plazo. La información entrante se filtra y solo los datos considerados útiles, basados en la comparación con los datos almacenados previamente, se añaden a la memoria. Esta puerta también utiliza funciones de activación, como la sigmoide, para controlar el grado de actualización de la memoria, visualizada en Ecuación (15) y Ecuación (16).

$$i_t = \sigma(W_i * [h_t - 1, x_t] + b_i) \qquad (15)$$

$$\mathcal{C}_t = \sigma(W_C * [h_t - 1, x_t] + b_C) \qquad (16)$$

Dónde i_t es la activación de la puerta de entrada, $\mathcal{C}_t$ es la nueva información candidata a añadirse a la memoria, W_i, W_C son las matrices de pesos asociadas a la puerta de entrada y a la candidata respectivamente y b_i, b_C son los sesgos correspondientes.

III. *Output gate*: la puerta de salida se encarga de determinar qué partes de la memoria a largo plazo se utilizarán para realizar una predicción en el momento actual (Ecuación (17) y Ecuación (18)). Esta predicción es modulada por la salida de las otras dos puertas y se basa en el estado actual de la memoria.

(17)

$$o_t = \sigma(W_o * [h_t - 1, x_t] + b_o)$$

$$h_t = o_t \odot tanh\,(C_t) \qquad (18)$$

Dónde o_t es la activación de la puerta de salida, h_t es el estado oculto o salida final de la célula LSTM, W_o es la matriz de pesos de la puerta de salida, b_o es el sesgo correspondiente y C_t es el estado de la memoria

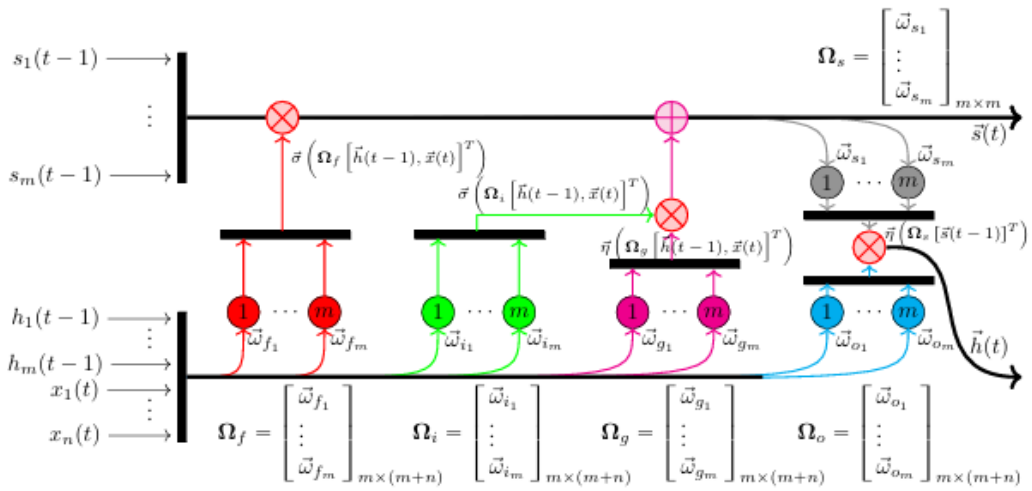


Figura 6 Diagrama de una celda LSTM. obtenida de (Suárez, López, & Ramos, 2009)

Todas las compuertas descritas en la Figura 6 se encuentran en cada una de las células LSTM que permitirán regular el flujo de la información, una vez construida la red neuronal LSTM se procede a su entrenamiento, alimentándola con nuestro conjunto de datos, durante este proceso se ajustarán los pesos de sus compuertas y de la memoria a largo plazo para minimizar los errores en las predicciones y los valores reales (Arna, 2021).

Una vez entrenada esta será capaz de predecir valores futuros en secuencias, clasificar secuencias o generar nuevas secuencias lo que hace de las redes neuronales LSTM una herramienta muy poderosa en el área del aprendizaje supervisado.

Las LSTM son especialmente poderosas en el campo del aprendizaje supervisado, donde se utilizan para tareas como la predicción de valores futuros en secuencias, la clasificación de secuencias, y la generación de nuevas secuencias. Su capacidad para aprender y recordar patrones a largo plazo las hace idóneas para aplicaciones como:

- Procesamiento del Lenguaje Natural (PLN): las LSTM son ampliamente utilizadas en modelos de PLN para tareas como la traducción automática y la generación de texto. su habilidad para recordar contextos lingüísticos a largo plazo es crucial en estas aplicaciones.
- Reconocimiento de Voz: en el reconocimiento de voz, las LSTM pueden capturar patrones temporales en señales de audio, permitiendo una transcripción más precisa del habla en texto.
- Predicción de Series Temporales: las LSTM se utilizan en la predicción de valores futuros en series temporales, como precios de acciones o datos meteorológicos, aprovechando su capacidad para manejar dependencias a largo plazo

El diagrama de flujo visto en la Figura 7 describe el proceso completo de un modelo LSTM, desde la carga y preparación de los datos hasta la evaluación y predicción. El LSTM es especialmente útil para problemas que involucran secuencias, como series de tiempo, procesamiento de lenguaje natural (NLP) o predicción de datos temporales. Cada etapa es crucial para garantizar que el modelo sea preciso y generalice bien a nuevos datos, a continuación, se describirán los procesos definidos en el anterior diagrama:

1. **Cargar datos:** se carga el conjunto de datos, que suele ser secuencial (series de tiempo, texto, etc.).
2. **Preprocesamiento de datos:** los datos se preparan para ser utilizados por el modelo. Esto incluye:
    - Limpieza y normalización.
    - Creación de secuencias o ventanas de tiempo.
    - División en conjuntos de entrenamiento, validación y prueba.

- Normalización de los datos.

3. **Entrenar el modelo:** el modelo se entrena utilizando los datos de entrenamiento y se valida con los datos de validación para evitar el sobreajuste.
4. **Predicción:** el modelo entrenado se utiliza para hacer predicciones sobre nuevos datos.

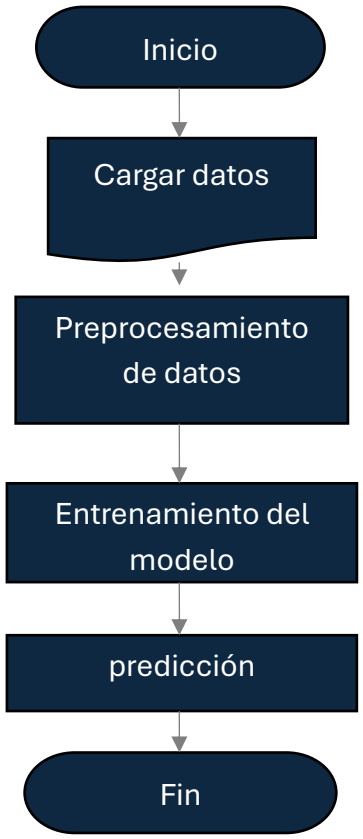


Figura 7 Diagrama de flujo para LSTM

2.2 Marco metodológico

Este apartado describe el enfoque metodológico adoptado para el desarrollo del sistema de predicción de energía fotovoltaica. Se detallan las técnicas, herramientas y procedimientos utilizados para la recolección, procesamiento y análisis de datos, así como la implementación y evaluación de los modelos de predicción. Además, se justifica la selección del método de investigación, se establecen los criterios para el

diseño experimental y se definen las métricas utilizadas para medir el rendimiento de los modelos aplicados. La metodología propuesta busca garantizar la validez y reproducibilidad de los resultados obtenidos.

2.2.1 Metodologías para minería de datos

En el desarrollo de proyectos de ciencia de datos, minería de datos y modelado predictivo, es fundamental seguir una metodología estructurada que guíe de forma sistemática cada una de las etapas del proceso analítico. Existen diversas metodologías reconocidas en este ámbito, siendo las más destacadas: KDD ("*Knowledge Discovery in Databases*"), CRISP-DM ("*Cross Industry Standard Process for Data Mining*") y SEMMA ("*Sample*", "*Explore*", "*Modify*", "*Model*", "*Assess*"). A continuación, se presenta una descripción detallada del funcionamiento de cada una, seguida de una comparación crítica entre estas.

2.2.1.1 KDD (Knowledge Discovery in Databases)

KDD es una metodología conceptual desarrollada en la década de 1990, y fue una de las primeras en definir el proceso completo de descubrimiento de conocimiento a partir de datos. Su enfoque es teórico y abarca desde la obtención de los datos hasta la interpretación de los resultados, las fases que la componen se pueden ver en la Figura 8.

Figura 8 Fases de la metodología KDD

1. Selección: Se identifican y seleccionan los conjuntos de datos relevantes desde diferentes fuentes de información.

2. Preprocesamiento: Se realiza una limpieza de los datos para eliminar valores faltantes, errores y ruidos que puedan afectar la calidad del análisis.

3. Transformación: Se aplican técnicas para convertir los datos a un formato adecuado para el modelado, como normalización, codificación o generación de variables derivadas.

4. Minería de datos: Es el núcleo del proceso, donde se aplican algoritmos de aprendizaje automático o estadísticos para extraer patrones, asociaciones o modelos predictivos.

5. Interpretación/Evaluación: Se interpretan los patrones o modelos obtenidos y se evalúa su utilidad y validez en relación con los objetivos planteados.

2.2.1.2 CRISP-DM (Cross Industry Standard Process for Data Mining)

CRISP-DM es una metodología desarrollada a finales de los años 90 como un estándar abierto para proyectos de minería de datos en entornos empresariales. Es ampliamente utilizada en la industria debido a su enfoque práctico y su estructura iterativa como se muestra en Figura 9, esta cuenta con las siguientes fases:

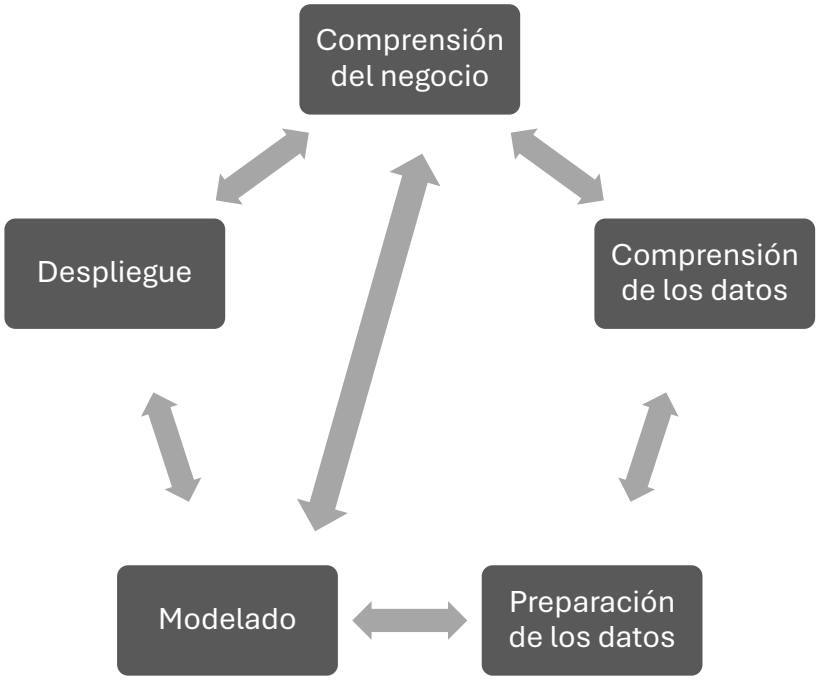


Figura 9 Fases de la metodología CRSP-DM

1. Comprensión del negocio: se definen los objetivos del proyecto desde la perspectiva del negocio y se establecen los requerimientos del problema a resolver.

2. Comprensión de los datos: se recopila información sobre los datos disponibles, su estructura, calidad y relevancia, generando una primera comprensión del contenido.

3. Preparación de los datos: incluye todas las tareas necesarias para construir el dataset final que se usará para el modelado, como limpieza, selección de atributos, integración y transformación de datos.

4. Modelado: se seleccionan y aplican técnicas de minería de datos o aprendizaje automático, ajustando los parámetros para mejorar el rendimiento.

5. Evaluación: se revisan los resultados desde la perspectiva del negocio y se determina si el modelo cumple con los objetivos planteados.

6. Despliegue: se implementa el modelo en el entorno real, generando reportes, aplicaciones o sistemas de soporte a la toma de decisiones.

2.2.1.2 SEMMA (Sample, Explore, Modify, Model, Assess)

La metodología SEMMA la cual se especializa en el análisis de datos o "*Data Mining", lo que significa que está se enfoca en "*la transformación de datos a conocimiento a través del análisis de información con técnicas y modelos.*"* (Berry & S.Linoff, 2004). Esta metodología aporta un marco estructurado en el que guiarnos a través de las etapas necesarias para un proyecto de minería de datos.

Esto es especialmente útil en este trabajo para mejorar la eficiencia del análisis de datos desde la selección de la muestra representativa permitiendo su modificación y simplificación de los datos, pasando por los modelos predictivos para concluir con una evaluación de los resultados obtenidos.

Fue desarrollada en 1990 por *"SAS institute"*, tratándose de una empresa estadounidense especializada en *software* de análisis estadístico, la metodología fue concebida para ser utilizada en su herramienta de minería de datos llamada "*SAS Enterprise Miner*". Debe su nombre a las cinco fases en las que se irán procesando los datos hasta llevarlos a su análisis final como se observa en la Figura 10 (Rodríguez Montequín, Álvarez Cabal, & Mesa Fernández, 2003).

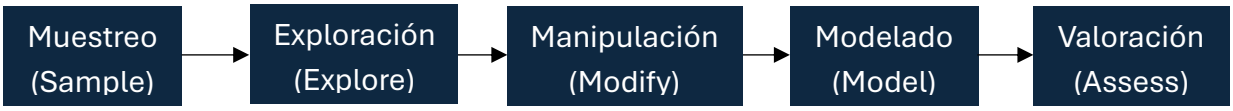


Figura 10 Fases de la metodología SEMMA. Imagen obtenida de (Rodríguez Montequín, Álvarez Cabal, & Mesa Fernández, 2003)

SEMMA se divide en 5 etapas que deberá atravesar la información para así ser procesada de manera eficiente y confiable para su análisis final.

I. Muestreo (Sample): la primera etapa de SEMMA consiste en seleccionar una muestra representativa de los datos originales. Este paso es fundamental, ya que una muestra adecuada permite obtener resultados precisos sin necesidad de trabajar con la totalidad de los datos, lo que reduce significativamente los tiempos de procesamiento y el uso de recursos computacionales. El muestreo se realiza utilizando técnicas estadísticas que garantizan que todos los elementos de la población tengan la misma probabilidad de ser seleccionados. Esto se puede lograr mediante métodos como el muestreo aleatorio simple o estratificado, dependiendo de las características de la población. Una muestra mal diseñada puede introducir sesgos, afectando negativamente la validez de los resultados.

   Para este proyecto se tomará una porción representativa de los datos climatológicos provenientes de la NASA y de la producción energética de la planta fotovoltaica correspondiente a un 80% de la población muestra total, siendo estos 3 meses de datos. Esto permitirá realizar análisis preliminares y pruebas iniciales con un menor costo computacional.

II. Exploración (Explore):en esta fase, se lleva a cabo un análisis exhaustivo de los datos para identificar patrones, relaciones entre variables y posibles

problemas, como valores atípicos o inconsistencias. La exploración ayuda a comprender mejor la naturaleza de los datos y establecer cuáles variables son relevantes para el problema.

El análisis exploratorio incluye la generación de estadísticas descriptivas, gráficos y visualizaciones que permiten detectar tendencias y correlaciones. Este paso es crucial para seleccionar las variables explicativas más importantes y eliminar aquellas que no aportan información útil o que podrían introducir ruido en los modelos.

Se analizarán variables como la radiación solar, la temperatura y la velocidad del viento, comparándolas con la producción energética de la planta fotovoltaica y se identificará el ruido o datos corruptos que puedan llegar a tener dichos datos.

III. Manipulación (Modify): la etapa de modificación se centra en la limpieza, transformación y preparación de los datos para garantizar su adecuación al análisis. Esto incluye la eliminación de ruido, el manejo de valores faltantes y la normalización de las variables.

Los datos raramente están listos para su análisis en bruto. Es necesario aplicar técnicas de preprocesamiento para mejorar la calidad de los datos. Por ejemplo:

- Eliminación de ruido: valores anómalos que no representan el comportamiento esperado de la población.
- Manejo de valores faltantes: sustitución por la media, mediana o valores interpolados.
- Normalización: escalado de las variables a un rango común, como 0 a 1, para evitar que una variable domine a las demás debido a diferencias de magnitud.

Se utilizará la técnica de normalización Min-Max para escalar variables a un rango entre 0-1 para mantener el mismo rango de magnitudes y así poder alimentar los modelos de manera eficiente. Asimismo, los valores faltantes en los datos de producción energética se sustituirán por la media de las observaciones. También se garantizará que todos los datos estén alineados temporalmente en intervalos de 1 hora para mantener la coherencia entre las series temporales.

IV. Modelado (Model): en esta fase, se seleccionan y entrenan los modelos predictivos que relacionarán las variables explicativas con la variable objetivo. Esta etapa incluye la elección del modelo más adecuado y el ajuste de sus hiperparámetros para maximizar su rendimiento.

El modelado implica el uso de técnicas estadísticas y algoritmos de aprendizaje automático para construir predicciones basadas en los datos procesados. Los modelos por utilizar dependen del tipo de datos y del problema a resolver. Algunos modelos comunes incluyen redes neuronales (LSTM), árboles de decisión (Bosques aleatorios) y métodos de clasificación como SVM.

Para esta etapa se entrenarán tres modelos inteligentes: LSTM, SVM y Bosques aleatorios. Cada modelo será alimentado con los datos preprocesados para predecir la producción energética de la planta fotovoltaica. Durante esta etapa, se realizarán ajustes en los hiperparámetros, como el número de árboles en Bosques aleatorios o el número de neuronas en LSTM, para optimizar su desempeño.

V. Valoración (Assess): la etapa final de SEMMA consiste en evaluar el desempeño de los modelos construidos utilizando métricas específicas. Este análisis permite determinar cuál modelo ofrece mejores resultados y es más adecuado para el problema planteado.

La evaluación incluye el uso de métricas como el Error Cuadrático Medio (MSE), que mide la magnitud promedio de los errores al cuadrado entre los valores predichos y los valores reales. Este tipo de métrica otorga mayor peso a los errores más grandes, lo cual es especialmente relevante en este trabajo.

Se utilizarán datos de prueba que no hayan sido utilizados durante el entrenamiento para evaluar el desempeño de los modelos. Por ejemplo, si el MSE de un modelo es significativamente menor que el de los otros, este modelo será considerado el más adecuado para predecir la producción energética. Los resultados se presentarán en forma de gráficos y tablas para facilitar su interpretación.

Al comparar estas tres metodologías, es evidente que cada una responde a diferentes necesidades como se muestra en la **¡Error! No se encuentra el origen de la referencia.**:

Tabla 1Comparativa de metodologías de minería de datos

Características	**KDD**	**CRISP-DM**	**SEMMA**
Enfoque principal	Conceptual/científico	Empresarial/practico	Técnico/analítico
Nivel de detalle técnico	Bajo	Medio	Alto
Aplicabilidad a investigación	Alta	Media	Alta
Orientación a negocio	Baja	Alta	Baja

.

En el contexto de este trabajo de tesis, centrado en el análisis e implementación de modelos inteligentes para la predicción de la producción fotovoltaica, la metodología SEMMA resulta especialmente adecuada debido a su enfoque técnico y operativo. A diferencia de otras metodologías orientadas al negocio, SEMMA está diseñada para entornos analíticos y científicos, lo cual se alinea con los objetivos de esta investigación, que se enfoca en la experimentación con modelos de aprendizaje profundo y la evaluación de su desempeño en la predicción de variables energéticas. Además, ofrece una estructura clara y precisa en etapas clave como la modificación de datos, el modelado predictivo y la evaluación cuantitativa, aspectos fundamentales cuando se trabaja con datos meteorológicos y eléctricos. Su alta compatibilidad con herramientas como Python y MATLAB facilita la implementación práctica de los modelos sin depender de objetivos comerciales. Finalmente, SEMMA prioriza la validación técnica rigurosa de los modelos desarrollados, lo cual garantiza la solidez y confiabilidad de los resultados obtenidos en contextos científicos y energéticos.

2.2.2 Métrica de error cuadrático medio

Para llevar a cabo un análisis útil entre los modelos es necesario utilizar una métrica de evaluación con la cual compararlas, la métrica que se estará utilizando con este propósito será la del error cuadrático medio (MSE por sus siglas en inglés), la cual es utilizada en el análisis de regresión para evaluar la precisión en la predicción de modelos por lo que la hace ideal para el análisis que llevaremos a cabo.

Esta métrica busca la diferencia del valor cuadrático esperado entre los resultados observados y los resultados predichos por el modelo, como se muestra en la Ecuación (1) (Runger & Montgomery, 2014):

(1)

$$MSE(\hat{\theta}) = E\left[\left\|\widehat{\theta - \theta}\right\|\right]^2$$

Esta misma ecuación puede ser separada en dos partes, la varianza y el sesgo al cuadrado, tal como se expresa en la Ecuación (2) (Runger & Montgomery, 2014):

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[E(\hat{\theta}) - \theta\right]^2 \qquad (2)$$

Siendo $Var(\hat{\theta})$ la representación de la varianza que es la dispersión entre datos del modelo y $\left[E(\hat{\theta}) - \theta\right]^2$ el sesgo al cuadrado que mide que tan lejos está el valor esperado del valor del modelo, entre menores sean ambos valores, mayor será la precisión y la exactitud respectivamente.

De esta forma se puede visualizar que tan fiable es la predicción del modelo bajo análisis y poder comparar la eficiencia de estos.

2.2.3 Normalización de datos

La mayoría de los algoritmos de aprendizaje automático tienen problemas para lidiar con atributos numéricos con diferentes valores de escalas, es por esto por lo que se recomienda la normalización de la escala (Géron, 2022).

Para la normalización de la escala se utilizará la técnica del escalado de característica o "*Min-Max escaler*", esta normaliza los valores a un rango especifico, por lo general para alimentar modelos de "*Machine Learning"* y "*Deep Learning"*, se utilizan rangos de entre [0,1], como se indica en la Ecuación (3):

(3)

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Dónde X es el valor original de la característica, X_{max} es el valor máximo de la característica, X_{min} es el valor mínimo de la característica y X_{scaled} es el valor de la característica escalado.

Con la aplicación de esta ecuación al conjunto de datos se puede normalizarlos a una escala de [0,1]. Para una mejor comprensión del uso de esta técnica aborde el siguiente ejemplo donde se muestra cómo normalizar un conjunto de datos (10, 15, 20, 25 y 30) utilizando la fórmula de escalado Min-Max, la cual transforma los valores al rango [0, 1] mediante $X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$, donde $X_{min} = 10$ y $X_{max} = 30$.

Se obtiene entonces una transformación que permite trabajar los datos en una misma escala.

Original
10
15
20
25
30

Xmin=10 Xmax=30

$$X_{scaled} = \frac{10-10}{30-10} = 0 \quad X_{scaled} = \frac{15-10}{30-10} = 0.25$$

$$X_{scaled} = \frac{20-10}{30-10} = 0.50$$

$$X_{scaled} = \frac{25-10}{30-10} = 0.75 \quad X_{scaled} = \frac{30-10}{30-10} =$$

Normalizado
0
0.25
0.50
0.75
1

2.3 Marco tecnológico

Para el desarrollo del presente proyecto de predicción de producción energética de una planta fotovoltaica, se utilizó el lenguaje de programación Python. Esta elección se debe a su amplia adopción en el campo del aprendizaje automático, así como a la gran disponibilidad de bibliotecas especializadas y documentación. Python ha demostrado ser una herramienta robusta y flexible para tareas de manipulación de datos, modelado y visualización, lo que lo convierte en un estándar de facto en investigaciones que implican inteligencia artificial y análisis de datos.

Entre las bibliotecas empleadas, se destacan las siguientes:

- NumPy: utilizada para la manipulación eficiente de arreglos multidimensionales, operaciones matemáticas y estructuras vectorizadas.

- Pandas: permitió el procesamiento, limpieza y estructuración de los datos meteorológicos y de producción energética, facilitando tareas de análisis exploratorio y transformación de series temporales.
- Scikit-learn (sklearn): biblioteca clave para la implementación de algoritmos clásicos de aprendizaje automático como Random Forest y Support Vector Regression, así como para tareas de normalización, validación cruzada y métricas de evaluación.
- TensorFlow y Keras: utilizadas para la creación y entrenamiento de modelos basados en redes neuronales, particularmente redes neuronales recurrentes (RNN) y su variante Long Short-Term Memory (LSTM), que son especialmente adecuadas para el tratamiento de datos secuenciales como las series temporales energéticas.

La elección de estas herramientas tecnológicas permitirá desarrollar un sistema de predicción eficiente, flexible y con capacidad de generalización frente a nuevos datos climatológicos.

Capítulo III

Estado del arte

El aprendizaje automático ha emergido como una de las herramientas más poderosas para abordar problemas complejos en diversas disciplinas. En el contexto energético, estas técnicas son fundamentales para enfrentar los desafíos que plantea la transición hacia sistemas sostenibles, como la integración de energía solar fotovoltaica en redes eléctricas. Este capítulo analiza el avance de las técnicas más relevantes, como las Máquinas de Soporte Vectorial (SVM), los Bosques aleatorios (RF) y las Redes Neuronales LSTM, contextualizando sus aplicaciones específicas y destacando las áreas de mejora identificadas en investigaciones recientes.

3.1 Predicciones para dar la bienvenida a la era solar.

La investigación de (Visser, 2023) desarrolla, prueba y evalúa múltiples modelos de predicción solar bajo varias configuraciones, y combinando técnicas estadísticas tradicionales, modelos físicos, algoritmos de aprendizaje automático y métodos probabilísticos. El trabajo final tiene como objetivo principal respaldar la integración masiva de energía solar en los sistemas eléctricos, teniendo en cuenta desafíos técnicos y económicos de la creciente penetración de fuentes renovables, como en el caso de la fotovoltaica.

En relación con las técnicas, este trabajo cubre técnicas desde modelos simples de regresión lineal hasta complicados métodos como redes neuronales y redes LSTM, así como técnicas Bosques aleatorios. Una de las características más innovadoras es la introducción de variables expertas, provenientes de modelos físicos de irradiancia y comportamiento de los sistemas fotovoltaicos, que se incorporan al

preprocesamiento y se muestran altamente efectivas para mejorar la precisión de los modelos, a menudo de manera óptima y en muchos casos más allá de la mejora que se obtendría sólo haciendo más complejo el algoritmo.

Los resultados obtenidos en la tesis de Visser arrojan luz sobre la actuación de los diferentes modelos. Siendo los modelos basados en árboles de decisión en general superiores a las redes neuronales profundas y otros modelos más complicados, especialmente cuando se consideran horizontes de predicción de un día para otro e intradía. Esto cuestiona la creencia común de que la complejidad algorítmica por sí sola garantiza mejores resultados.

Siendo que en el mismo documento se resalta la importancia del nivel de agregación espacial como una variable crítica, refiriéndose a la agrupación de múltiples sistemas fotovoltaicos distribuidos geográficamente, generando un efecto de suavizado que reduce la volatilidad de la producción y mejora la precisión, aunque no siempre conlleva un aumento proporcional en los beneficios económicos. Por último, Visser destaca la relevancia de los datos abiertos y en tiempo real, tanto meteorológicos como de producción, como un factor que podría multiplicar el potencial de mejora en los modelos.

3.2 Predicción de la energía solar a través del aprendizaje automático.

El artículo de (Subramanian, Karthik, Krishna, Kumar, & Prasath, 2016) busca aportar al campo del aprendizaje automático en el contexto de producción fotovoltaica. Aquí explica la importancia de este tipo de investigaciones en este contexto, ya que, debido al auge de la tecnología en paneles solares y su cada vez más amplio uso, estos necesitan afrontar su mayor obstáculo, la gran volubilidad de las condiciones climáticas, las cuales afectan directamente a la producción de energía, esto conllevando graves problemas de gestión energética y costos económicos elevados.

Para esto surgen los algoritmos de aprendizaje automático implementados en este ámbito; estos buscan realizar predicciones exactas para permitir a las plantas de energía fotovoltaica gestionar de mejor manera sus recursos y optimizar su producción y almacenamiento de energía.

Por lo tanto (Subramanian, Karthik, Krishna, Kumar, & Prasath, 2016) implementaron algoritmos de aprendizaje automático, más específicamente utilizaron Maquinas de soporte vectorial (SVM), Bosques aleatorios (RF) y “Gradient Boosting”, hasta este punto el articulo no saldría de lo convencional en cuanto a implementacion de modelos inteligentes, sin embargo buscaron innovar combinando los anteriores algoritmos y aplicando un modelo de regresión linear como “meta-learner”, esta última adición fue usada para que atravez de las predicciones de los anteriores algoritmos y un conjunto de datos de validación el modelo tenga la capacidad de aprender de las predicciones de los demás algoritmos y ayude a ajustar las predicciones de estos. La implementación de este conjunto de algoritmos terminó generando como resultado una precisión basada en la métrica AUC (Área bajo la curva) del 99% de precisión.

A pesar de contar con resultados tan prometedores, en el mismo documento se destaca el mayor problema respecto a su propuesta, siendo este el gran costo computacional que implica implementar dichos algoritmos, siendo este un gran obstáculo para pequeñas plantas de producción fotovoltaica y una gran inversión de capital para otras con mayor capacidad.

3.3 Modelos de Aprendizaje Profundo para la Predicción de Energía Fotovoltaica

La revisión de (Yu, y otros, 2024) recopila el uso de múltiples algoritmos de aprendizaje profundo, segmentándolos en 4 grandes grupos, perceptrón multicapa (MLP), redes neuronales recurrentes (RNN), redes neuronales convolucionales (CNN) y red neuronal grafica (GNN) detallando su funcionamiento, pasando por sus beneficios al implementarlos en la predicción de produccion fotovoltaica, las limitaciones

existentes en este mismo ámbito y múltiples algoritmos que buscan dar solución a dichos problemas.

Comenzando con los algoritmos MLP, su arquitectura consiste en una capa de entrada, una capa de salida y una o más capas ocultas entre las dos capas anteriores. Gracias a las múltiples capas que contiene el modelo, este es capaz de manejar complejos problemas de patrones de clasificación.

Para su entrenamiento utiliza tecnicas de "*back-propagation*" para así al encontrar diferencias entre los resultados del modelo y los resultados deseados, este pueda ajustar los pesos correspondientes.

Debido a la necesidad de mayor precisión al momento de la implementacion en la predicción de produccion fotovoltaica han surgido múltiples opciones para dar solución, o mejorar aspectos de este tipo de algoritmos, como lo es "*robust-MLP*", que remplaza la funcion de perdida convencional por la funcion de perdida "*pseudo-Huber*" que combina la funcion de perdida cuadrática y la funcion de perdida absoluta, esto ajustándose a datos del día por venir que se encuentran comúnmente en la produccion fotovoltaica, dando como resultado mejoras en puntuaciones en métricas como RMSE y MAE.

Tras lo anterior, el documento continuó con las RNN, estas son redes que contienen un ciclo, este consiste en múltiples neuronas interconectadas, las cuales reciben y envian información unas a otras, repitiendo este proceso hasta llegar a un resultado. Gracias a esto son capaces de procesar secuencias de datos y capturar dependencias temporales con conexiones recurrentes. Se resalta la capacidad de este algoritmo de utilizar información previa en predicciones actuales, sin embargo, no es posible que este modelo procese secuencias de datos largos, debido a que no es capaz de mantener información por mucho tiempo.

Como solución a estos desafíos surgen las redes neuronales LSTM, estas cuentan con una arquitectura capaz de mantener memoria a largo y corto plazo, gracias a su sistema de compuertas incluidas en cada neurona de la red, consistiendo en puesta

de entrada, puerta de olvido y pueta de salida, permitiendo así administrar el tipo de información que se mantiene en la memoria a largo plazo y así ser utilizada en próximas tareas, este tipo de modelos son mayormente utilizadas en el procesamiento de lenguaje natural, sin embargo debido a su capacidad de procesamiento en series de tiempo, estan siendo utilizadas en diversos aspectos, siendo en la predicción de produccion fotovoltaica la que se resalta en esta revisión, dando resultados prometedores.

El articulo prosigue con los algoritmos CNN, estos son principalmente usados para el procesamiento de imágenes debido a su estructura que organiza las neuronas en 3 dimensiones, ancho, largo y profundidad, cada neurona en la misma capa se conecta con un bloque de resultados de la anterior capa, siendo estas características la razon de su efectividad en el procesamiento de imágenes, las capas que componen estos algoritmos consisten en una capa de entrada, una capa convolucional, una capa de agrupación, una capa completamente conectada y finalmente la capa de salida, recalcando el uso de la capa convolucional encargada de extraer características de las imágenes, y la capa de agrupación, cuya funcion es prevenir el sobreajuste del modelo y reducir la dimensionalidad de los datos.

A pesar de que este tipo de algoritmos tienen un excelente desempeño en el procesamiento de imágenes pueden ser utilizadas en la predicción de series de tiempo y aprender de forma automática representaciones de series de tiempo con múltiples escalas, encontrando en estas un buen desempeño en apartados como la explosión o desvanecimiento de gradiente y la retención de memoria.

Modelos como el TCN (redes convolucionales temporales) permiten al algoritmo el uso de anterior información para tareas actuales utilizando "*back-propagation*", este modelo es capaz de manejar secuencias largas y mantener un entrenamiento veloz obteniendo resultados con precisión durante sus tareas.

La última sección dedicada a modelos se centra en algoritmos del tipo GNN, este tipo de modelos se enfocan en el procesamiento de datos estructurados en graficas, en

tiempos recientes estos han sido aplicados en predicciones de series de tiempo construyendo graficas con dichas series. Estos algoritmos pueden capturar dependencias espacio temporales considerando la conexión topológica con la información de los nodos. Así mejorando su precisión tomando en cuenta la relación entre tiempo y variables.

Para construir las gráficas con las series de tiempo cada apartado de tiempo es representado como un nodo de la gráfica y los bordes entre los nodos se consideran la relación entre estos, el peso de estos bordes representa la relación o influencia entre cada dato en la gráfica, como resultado se entrega una gráfica codificada que contiene las dependencias entre datos y así poder alimentar al modelo para su procesamiento. Este enfoque permite mostrar de forma explícita las relaciones y dependencias entre los datos, manejar series de datos masivas, procesar de manera natural múltiples series de tiempo relacionadas y capturar datos de tiempo a largo plazo.

No obstante, también presenta sus propios retos, como lo son la propia creación de las gráficas, siendo este modelo sumamente sensible a las relaciones y correlaciones de datos expresadas en estas, a su vez este modelo cuenta con una complejidad computacional alta haciéndose más intrincada entra mayor sea la longitud de la serie.

El trabajo de (Yu, y otros, 2024) muestra ser una revisión bastante completa que se adentra en profundidad no solo a los arquetipos de modelos anteriormente mostrados si no que ahonda en sus avances , los retos y mejoras de estos mismos.

Como conclusión este articulo propone que cada modelo presentado al contar con diversas limitaciones y beneficios los hacen perfectamente utilizables en diversos ámbitos incluyendo el de predicción de produccion fotovoltaica, siendo el contexto en el que busca ser implementados el mayor punto de inflexión al momento de seleccionar el modelo ideal para su aplicación.

3.4 Un Modelo Híbrido de Aprendizaje Automático para la Predicción de Energía Solar

(R. Kumar, P. Sundari, Parakash, & Sathya, 2023)centraron su artículo en la implementación y evaluación de un modelo de aprendizaje profundo híbrido, buscando con esto una opción entre los algoritmos de aprendizaje automático con mayor precisión y menor costo computacional, para así ayudar al sector de producción fotovoltaica.

Comienzan el documento dando un repaso respecto al contexto actual de las energías renovables y cómo estas han incrementado su penetración en el mercado, y hacen énfasis en la producción de energía a través de celdas fotovoltaicas. Remarcan la importancia de la implementación de modelos inteligentes para la predicción de energía, ya que este tipo de producción cuenta con múltiples variables con grandes fluctuaciones, que impiden la correcta administración de la energía. Para esto, la implementación de modelos capaces de tener predicciones precisas facilitaría este aspecto, otorgando mayor penetración en el mercado y facilitando el uso de tecnologías fotovoltaicas.

Para esto, el artículo propone la implementación de un modelo inteligente híbrido, combinando modelos de “support vector regression” (SVR), “random forest regression” (RFR) y “artificial neural network” (ANN). El modelo solicita una predicción a cada algoritmo y así sumar sus predicciones junto con pesos respectivos a la importancia que tiene cada modelo para el resultado final. Estos pesos fueron ajustados durante la experimentación buscando la configuración óptima para el experimento.

Este modelo fue comparado con modelos individuales como SVR y árboles de decisión utilizando MAE, RMSE y MSE como métricas de evaluación, llegando a tener resultados positivos en su precisión respecto a los modelos individuales con los que se le comparó. A pesar de esto, el modelo mostró la necesidad de una gran cantidad de información para su entrenamiento, al igual que un tiempo de entrenamiento tardado.

Capítulo IV

Implementación

En este capítulo se describen en detalle el proceso de implementación del sistema propuesto para la predicción de la producción energética de una planta fotovoltaica. Se presenta la arquitectura general del sistema, así como los modelos de aprendizaje automático seleccionados.

4.1 Configuración del experimento.

La implementación de tres modelos inteligentes: Bosques aleatorios (RF), Maquinas de Soporte Vectorial (SVM) y "*Long Short-Term Memory*" (LSTM). Estos modelos se entrenarán con datos climatológicos y de producción energética, previamente procesados para garantizar su calidad. El tratamiento de los datos incluirá la eliminación de ruido, la estandarización a un formato común y la normalización utilizando técnicas como "*Min-Max normalization*", elegida por su simplicidad y eficiencia en la escala de datos numéricos, como se detalla en (Krishna & Kumar, 2015) y (Géron, 2022).

Por lo tanto, el proyecto se dividirá en la fase de "*data mining*" y la fase de implementación y evaluación, descritas a continuación.

Tratamiento y recopilación de datos

- Recopilación de datos climatológicos y de producción energética, identificando las variables clave que afectan la producción fotovoltaica.

- Aplicación de técnicas de preprocesamiento, incluyendo la eliminación de ruido y la normalización, para garantizar la calidad de los datos y su integración en los modelos de aprendizaje automático.
- Uso de técnicas como "*Min-Max Normalization*" para escalar los datos de manera eficiente, lo que facilita su procesamiento y análisis posterior.

Evaluación y comparación de modelos inteligentes

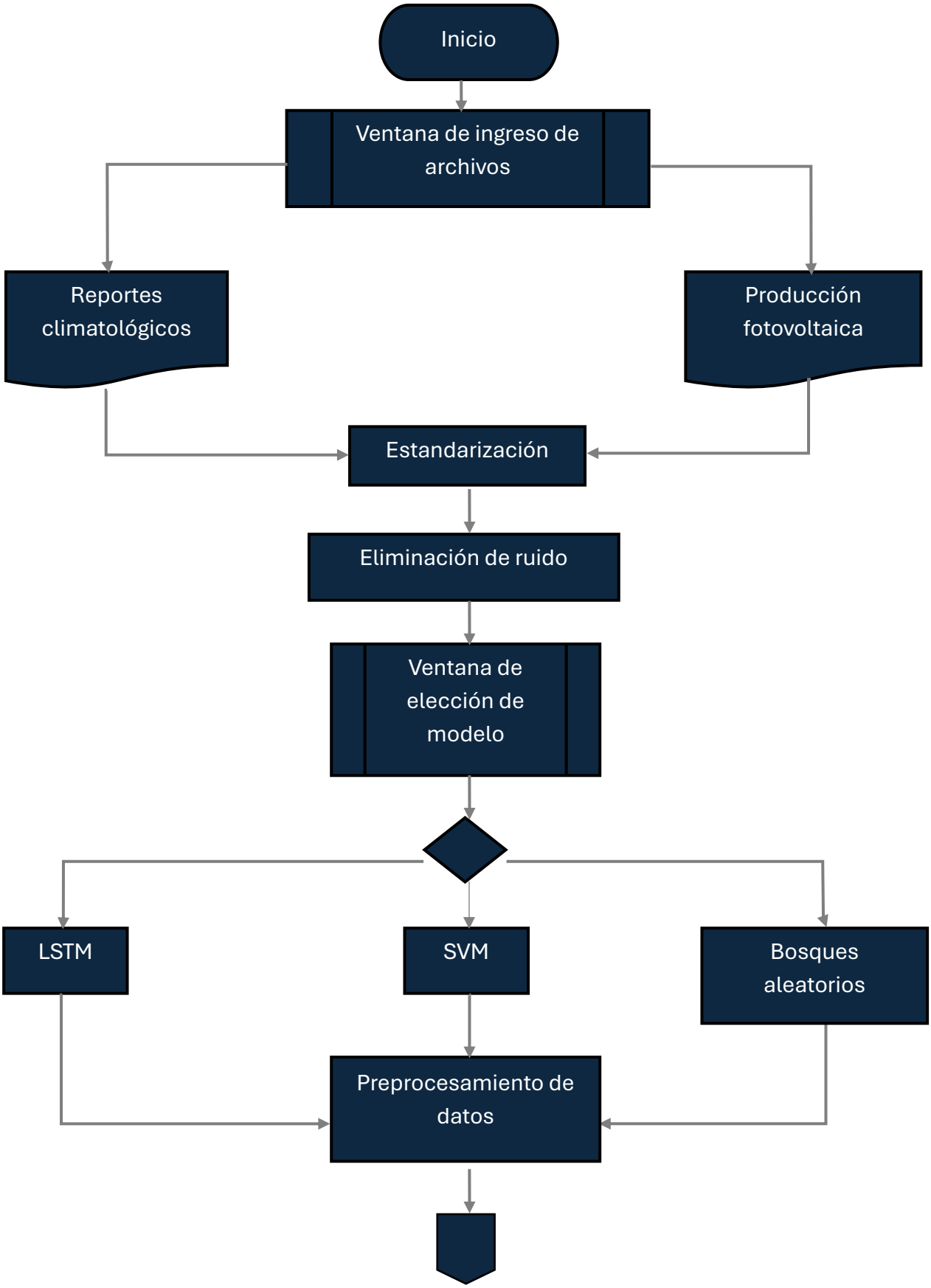
- Implementación de los modelos RF, SVM y LSTM, entrenándolos con los datos preprocesados.
- Evaluación de su desempeño utilizando métricas como el Error Cuadrático Medio (MSE, por sus siglas en inglés), que mide las discrepancias entre las predicciones y los valores reales, otorgando mayor peso a los errores significativos.
- Análisis comparativo para identificar el modelo más eficiente y adecuado para solucionar el problema de predicción de producción energética.

En última instancia, este análisis permitirá determinar el modelo más preciso y confiable, lo que contribuirá al desarrollo de sistemas inteligentes más eficientes en el sector energético, impulsando la integración de la energía fotovoltaica como una fuente confiable y sostenible de generación eléctrica.

4.2 Arquitectura del sistema

Una vez definidos los requerimientos técnicos y funcionales del sistema procedemos a definir cómo será la arquitectura del sistema a crear, se tomarán en cuenta los componentes principales de este como lo son la interfaz de usuario, el preprocesamiento de los datos, los modelos predictivos y la interacción con el usuario.

Para esto se realizó el siguiente diagrama de flujo (Figura 11) que demuestra paso a paso el funcionamiento, procesos, entradas y salidas de datos del sistema a crear.



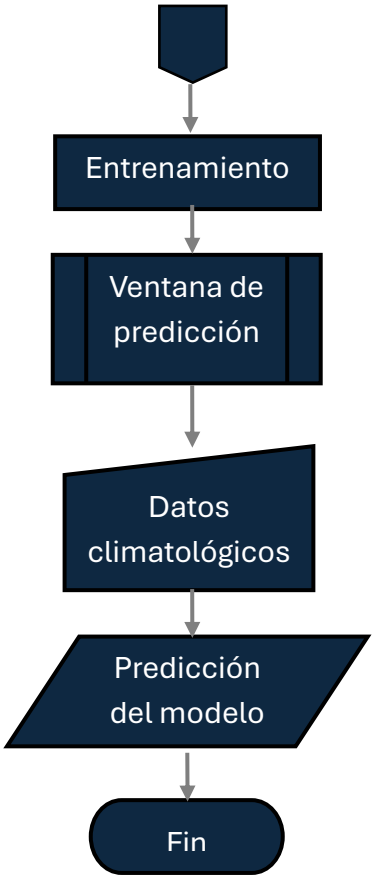


Figura 11 Diagrama de flujo del sistema de predicción fotovoltaica

Con la Figura 11 se ve el cómo los datos proporcionados por el usuario y sus interacciones con el sistema van fluyendo a lo largo del sistema proporcionando los elementos y variables necesarios para nuestro objetivo final, la predicción de producción fotovoltaica. Comenzando con ingresar los archivos que contiene reportes climáticos y de producción fotovoltaica para procesar los archivos y estandarizarlos con el mismo formato y una frecuencia de entrada de datos de 1 hora por entrada, y posteriormente la eliminación de ruido para mejorar la calidad de nuestro conjunto de datos, una vez procesados los archivos se implementa un modelo inteligente entre LSTM, SVM y Bosques aleatorios. Una vez seleccionado el modelo los datos que se procesaron antes son normalizados para posteriormente alimentar y entrenar el modelo, una vez entrenado se ingresan datos climatológicos para así realizar una predicción utilizando el modelo seleccionado y entrenado.

4.3 Tratamiento de datos

Una vez definida la arquitectura del sistema y el flujo de los datos, es fundamental establecer el proceso que deben seguir las principales funciones del sistema. Estas funciones incluyen el procesamiento de datos para su estandarización y la normalización de los datos. Estos pasos son esenciales para garantizar la claridad en el manejo de la información y para comprender cómo el sistema procesará los datos para generar predicciones precisas y confiables.

4.4.1 Definición y clasificación de datos

Como punto inicial en el tratamiento de datos es menester definir las variables que se manejaran durante la implementacion de los modelos y su respectiva clasificación para el posterior analisis. Estas son variables climatológicas que fueron seleccionadas para el entrenamiento de los modelos, se obtuvieron de bases de datos de la NASA enfocadas en el almacenamiento de información ambiental a largo plazo, contando con múltiples variables climatológica asociadas a la fecha y hora en la que se registraron.

Para lograr un entrenamiento eficiente en modelos inteligentes se seleccionaron las variables con mayor influencia en la produccion fotovoltaica encontradas en estas bases de datos con base a la ponderación encontrada en (Visser, 2023), las cuales fueron:

- Temperatura del aire a 2 metros de altura sobre la superficie, expresada en grados Celsius (°C)[T2M].

- Humedad relativa del aire a 2 metros de altura, expresada en porcentaje (%)[RH2M].

- Velocidad del viento a 10 metros de altura, expresada en metros por segundo (m/s)[WS10M].

- Radiación fotosintéticamente activa total en superficie bajo condiciones de cielo despejado, en W/m^2[CLRSKY].

- Precipitación total acumulada en superficie, expresada en milímetros (mm)[Lluvia].

Con las variables de entrenamiento seleccionadas se realizó una clasificación por tipo de día correspondiente a los comportamientos del conjunto de estas variables a lo largo de 24 horas, esto para permitir un mejor analisis comparativo entre los modelos al solicitar predicciones basadas en conjuntos de datos especificos y así evaluar el desempeño en condiciones específicas. Los conjuntos de datos se catalogaron en día normal, día soleado, día nublado, día lluvioso y día parcialmente nublado, estas clases se encuentran definidas en Tabla 2.

Tabla 2 Definición de variables por tipo de día

Tipo de día	**T2M**	**RH2M**	**WS10M(m/s)**	**CLRSKY**	**Lluvia**
Normal	20°-35°	40%-70%	1-5	600-1000	0
Soleado	25°-38°	30%-60%	1-4	700-1100	0
Nublado	18°-28°	60%-90%	2-6	100-400	0-0.2
Lluvioso	16°-26°	>80%	2-8	0-200	>0.5
Parcialmente nublado	20°-32°	50%-80%	1-5	300-800	0-0.3

Por último, se definió la variable objetivo, esta es la variable que se busca predecir a través de los modelos inteligentes, la cual es el volumen de energia eléctrica producida por la planta fotovoltaica, en WH. Estos datos se obtuvieron de la base de datos de una planta de producción fotovoltaica de la empresa “COCOA”, ésta contiene información sobre la producción diaria repartida en entradas por cada hora (24 horas).

El conjunto de información total utilizada en este proyecto corresponde a las fechas de entre el 30/01/2022 al 30/03/2022.

4.4.1 Procesamiento de los datos para su estandarización

En esta etapa, el sistema se encargará de recopilar y modificar los datos proporcionados por el usuario con el objetivo de garantizar su calidad y consistencia. A continuación, se describen los pasos clave que se llevarán a cabo:

1. **Unificación de la columna temporal:**
   Se creará una columna única que registre la fecha y hora exacta en la que se realizó cada medición. Esta columna incluirá el mes, día, año y hora, lo que permitirá una correcta sincronización y ordenamiento temporal de los datos. Este paso es crucial para asegurar que los datos estén alineados cronológicamente y puedan ser procesados de manera secuencial.

2. **Detección y corrección de datos ruidosos o inconsistentes:**
   Se identificarán y corregirán valores atípicos, datos nulos o inconsistentes que puedan afectar la calidad del conjunto de datos. Por ejemplo, se detectarán valores fuera de rangos válidos, como meses mayores a 12 o días negativos. Estos datos serán corregidos asignándoles un valor de cero o, en su defecto, interpolándolos a partir

de valores adyacentes. Este proceso garantiza la integridad de los datos y evita que errores en la medición afecten negativamente las predicciones.

3. **Ajuste de la frecuencia de los datos:**
   Para mejorar la precisión de las predicciones, se asegurará que los datos tengan una frecuencia constante de 1 hora. En caso de que existan intervalos de tiempo con datos faltantes, se calculará la media de los valores disponibles para generar una aproximación de los datos faltantes. Este paso es fundamental para mantener una secuencia temporal uniforme, lo que facilita el análisis y la modelización de los datos.

4. **Estandarización del formato de los datos:**
   Finalmente, se asegurará que todos los datos estén en un formato consistente y estandarizado. Esto incluye la normalización de unidades de medida, la alineación de columnas y la eliminación de redundancias. Este proceso facilita la manipulación de los datos y asegura que puedan ser utilizados de manera eficiente por los modelos de predicción.

Estos pasos no solo mejoran la calidad de los datos, sino que también preparan el conjunto de datos para su posterior preprocesamiento y modelización. La estandarización es un paso crítico para garantizar que los modelos LSTM, Bosques aleatorios y SVM puedan procesar la información de manera efectiva y generar predicciones precisas sobre la producción de energía en la planta fotovoltaica.

4.4.2 Normalización

Una vez que los datos han sido estandarizados y preprocesados, el siguiente paso crítico es la normalización. Este proceso es fundamental para garantizar que los datos

estén en una escala adecuada, lo que facilita el entrenamiento eficiente de los modelos de inteligencia artificial y mejora su capacidad de generalización.

En el sistema implementado, la normalización Min-Max se aplica tanto a las características de entrada como a la variable objetivo. El proceso se realiza en las siguientes etapas:

1. **Variables objetivo:**
   - Las características de entrada se obtienen de variables como la radiación solar (ALLSKY_KT), la temperatura (T2M), la humedad (RH2M), entre otras.
   - La variable objetivo corresponde a la producción de energía (WH).
2. **Cálculo de parámetros de normalización:**
   - Se calculan los valores mínimos para cada característica y la variable objetivo utilizando únicamente el conjunto de entrenamiento. Esto evita la fuga de información (data leakage), que podría ocurrir si se incluyeran datos de prueba en el cálculo.
3. **Aplicación de la normalización:**
   - Se utiliza la clase MinMaxScaler de la biblioteca Scikit-learn para aplicar la transformación Min-Max a los datos.
4. **Almacenamiento de los escaladores:**
   - Esto permite aplicar la misma transformación a nuevos datos durante la fase de inferencia, garantizando consistencia en el preprocesamiento.

Una vez explicado el cómo se abordará cada aspecto teórico de la funcionalidad del sistema se procederá a su implementación.

4.5 Aplicación de modelos

En este apartado se detalla el proceso de implementación del sistema predictor de producción de energía fotovoltaica. Se describen los modelos de aprendizaje automático utilizados, su configuración y los criterios empleados para su entrenamiento y validación.

El sistema ha sido diseñado para procesar datos meteorológicos y de producción de energía, con el objetivo de generar predicciones precisas sobre la energía que producirá la planta fotovoltaica en intervalos de 1 hora. Para ello, se han empleado diversos algoritmos de inteligencia artificial, incluyendo redes neuronales LSTM, Bosques aleatorios y SVM.

A continuación, se presentan los fragmentos de código correspondientes a cada modelo, acompañados de una explicación detallada de su funcionamiento, estructura y parámetros utilizados. También se discuten las estrategias implementadas para el preprocesamiento de datos, la optimización de hiperparámetros y la validación de los resultados.

Está sección proporciona una visión clara del proceso de desarrollo, facilitando la comprensión de los métodos empleados y sirviendo como base para futuras mejoras o adaptaciones del sistema.

4.5.1 Implementación de Bosques Aleatorios

Para comenzar en la implementación del modelo de Bosques aleatorios se deben preparar los datos climáticos(dfNasaResampled) y de producción(dfCocoaResampled) para su uso en el modelo. Se eliminan columnas no numéricas, como la fecha, y se conservan únicamente los valores numéricos relevantes para el análisis como se muestra en Figura 12.

```
# ========== 1. PREPROCESAMIENTO ==========
# Excluir fecha y seleccionar solo numéricos
X = dfNasaResampled.drop(columns=["Fecha"]).select_dtypes(include=['number'])
y = dfCocoaResampled["WH"]

# Verificación de rangos
print("\nEstadísticas descriptivas de los datos:")
print(X.describe())
print("\nRango de WH:", y.min(), "-", y.max())

# Normalización de y si los valores son muy grandes
y_scaling_factor = 1
if y.max() > 1000:  # Si WH tiene valores muy grandes
    y_scaling_factor = 1000
    y = y / y_scaling_factor
    print(f"\nAplicado factor de escala a WH: 1/{y_scaling_factor}")

# Limpieza robusta
X = X.replace([np.inf, -np.inf], np.nan).fillna(X.median())
y = y[X.index]
```

Figura 12 implementación para Bosques aleatorios (Preparación de datos).

Se realiza un análisis estadístico descriptivo sobre los datos de producción (y) y se verifica su rango. En caso de que los valores sean muy grandes (por ejemplo, mayores a 1000), se aplica un factor de escala para facilitar el aprendizaje del modelo.

Posteriormente, se aplica una limpieza robusta para reemplazar valores infinitos y ausentes (NaN) por la mediana de cada columna. Finalmente, se asegura que los índices de las variables predictoras coincidan con los de la variable objetivo.

Antes del entrenamiento, se estandarizan los datos mediante "*StandardScaler*", asegurando que todas las características tengan media cero y desviación estándar uno. Esto mejora la eficiencia del algoritmo de aprendizaje y evita sesgos relacionados con escalas diferentes. Se utiliza un modelo de tipo "*RandomForestRegressor*", y se

realiza una búsqueda aleatoria (RandomizedSearchCV) sobre un espacio definido de hiperparámetros para encontrar la mejor configuración, visualizado en Figura 13. Los parámetros a ajustar incluyen:

```python
# Escalado simplificado
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# ========== 3. OPTIMIZACIÓN ==========
param_grid = {
    'n_estimators': [50, 100, 200, 300],
    'max_depth': [None, 5, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 0.5, 0.7],
    'bootstrap': [True]
}

# Configuración de búsqueda
rf = RandomForestRegressor(random_state=42, n_jobs=-1)
search = RandomizedSearchCV(
    rf, param_grid, n_iter=50, cv=5,
    scoring='neg_mean_squared_error',
    verbose=2, random_state=42, n_jobs=-1
)

print("\nIniciando búsqueda de hiperparámetros...")
start_time = time()
search.fit(X_train_scaled, y_train)
print(f"\nBúsqueda completada en {(time()-start_time)/60:.1f} minutos")
```

Figura 13 implementación para Bosques aleatorios (Estructura del modelo).

- Número de árboles (n_estimators).
- Profundidad máxima del árbol (max_depth).
- Mínimo de muestras para dividir un nodo (min_samples_split).
- Mínimo de muestras en una hoja (min_samples_leaf).
- Selección de características (max_features).
- Método de muestreo con reemplazo (Bootstrap).

La búsqueda se configura para realizar validación cruzada con cinco particiones (cv=5) y evaluar el error cuadrático medio como métrica de desempeño.

Se mide el tiempo de búsqueda y se ejecuta el ajuste sobre los datos de entrenamiento escalado

Finalmente, el modelo entrenado se guarda en un archivo llamado "*Model_RF.pkl*" utilizando la función "*dump*" de la biblioteca "*joblib*". Esto permite guardar el modelo para su uso futuro sin necesidad de volver a entrenarlo cada vez. Se imprime un mensaje indicando que el modelo ha sido guardado correctamente.

4.5.2 Implementación de SVM

El proceso de implementación para Support Vector Machines (SVM) en su variante de regresión (LinearSVR) inició con la integración de los conjuntos de datos climáticos (dfNasaResampled) y de producción (dfCocoaResampled) como se ve en Figura 14. A diferencia del enfoque utilizado en Bosques aleatorios, aquí se realizó una combinación explícita mediante "*merge*" en la columna "Fecha", conservando solo los registros coincidentes (inner join).

```
# 1. Combinar datos
dfCombined = pd.merge(dfNasaResampled, dfCocoaResampled, on="Fecha", how="inner")
features = ["T2M", "RH2M", "WS10M", "CLRSKY", "Lluvia"]
y = dfCombined["WH"].values
X = dfCombined[features].values

# 2. Escalado con StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
jb.dump(scaler, 'scaler_SVM.pkl')

# 3. Dividir datos
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# 4. Configuración óptima para regresión
best_mse = float('inf')
best_params = None
mse_history = []
```

Figura 14 implementación para SVM (Preparación de datos).

Se aplicó estandarización mediante "StandardScaler()" para normalizar las características utilizando transformación de los datos a escala unitaria (media = 0, desviación estándar = 1).

Se implementó una estrategia de búsqueda aleatoria manual para el modelo LinearSVR mostrado en Figura 15, evaluando combinaciones de hiperparámetros:

- Parámetro de regularización “C “(trade-off entre margen y error).
- Margen de tolerancia en la función de pérdida (Épsilon).
- Tipo de función de pérdida (Loss).
- Máximo de iteraciones para garantizar convergencia (max_iter).

```
param_ranges = {
    'C': [0.1, 1, 10, 100],
    'epsilon': [0.01, 0.1, 0.5],
    'loss': ['epsilon_insensitive', 'squared_epsilon_insensitive'],
    'max_iter': [10000]  # Fijo para garantizar convergencia
}

print("=== Entrenamiento LinearSVR (Regresión) ===")
start_time = time.time()

# 5. Búsqueda de parámetros
for i in range(20):  # Iteraciones reducidas pero efectivas
    params = {
        'C': random.choice(param_ranges['C']),
        'epsilon': random.choice(param_ranges['epsilon']),
        'loss': random.choice(param_ranges['loss']),
        'max_iter': param_ranges['max_iter'][0]
    }

    model = LinearSVR(**params, random_state=42)

    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    current_mse = mean_squared_error(y_test, y_pred)
    mse_history.append(current_mse)
```

Figura 15 implementación para SVM (estructura del modelo).

Finalmente, el modelo entrenado se guarda en un archivo llamado "*Model_SVM.pkl*" utilizando la función “*dump*” de la biblioteca “*joblib*”. Esto permite almacenar el modelo para su uso posterior sin necesidad de reentrenarlo. Se imprime un mensaje confirmando que el modelo ha sido guardado correctamente como "*Model_SVM.pkl*".

4.5.3 Implementación de LSTM

El proceso para el modelo LSTM difiere significativamente en el tratamiento de los datos debido a su naturaleza secuencial. Se combinaron los datos climáticos (df_nasa) con los datos de producción (df_cocoa['WH']) en un solo "*DataFrame*".

Se eliminaron explícitamente columnas de tipo "*datetime*" para conservar solo datos numéricos vistos en la (Figura 16).

Una vez procesados los datos se aplicó MinMaxScaler por columna (rango [0,1])

Cada variable se normalizó independientemente, almacenando sus escaladores para uso futuro.

```
# === Preparar datos ===
df_total = df_nasa.copy()
df_total['WH'] = df_cocoa['WH'].reset_index(drop=True)
df_total = df_total.select_dtypes(exclude=['datetime'])
data = df_total.values.astype(np.float32)

# Normalización por columnas (como MATLAB)
scalers = []
data_norm = np.zeros_like(data)
for i in range(data.shape[1]):
    scaler = MinMaxScaler()
    data_norm[:, i:i+1] = scaler.fit_transform(data[:, i:i+1])
    scalers.append(scaler)

# Guardar escaladores
os.makedirs('model', exist_ok=True)
for i, scaler in enumerate(scalers):
    joblib.dump(scaler, f'LSTM_scaler_{i}.save')

# Separar datos (80% entrenamiento, 20% validación)
train_size = int(len(data_norm) * 0.8)
train_data = data_norm[:train_size]
val_data = data_norm[train_size - 24:]
```

Figura 16 implementación para LSTM (Preparación de datos).

Los escaladores se guardaron individualmente para permitir la inversión de transformaciones.

Se implementó una función personalizada "create_sequences" que transforma los datos en formato adecuado para LSTM:

- Longitud de secuencia (lookback): 24 pasos temporales (horas).
- Paso de predicción (horizonte): 1 paso adelante (next-step prediction).

- Estructura 3D requerida por LSTM: [muestras, pasos temporales, características].

Se diseñó una red neuronal profunda con las siguientes características (Figura 17):

- 6 capas con 5 conexiones.
- 24 pasos, 5 características (Input).
- Capa recurrente con memoria con 100 unidades (LSTM).
- Regularización para evitar “overfitting” (Dropout).
- Capa oculta tradicional con 50 unidades y activación ReLU (Dense).
- Predicción de WH (Dense [Output]).

```python
def create_sequences(data, seq_len=24):
    x, y = [], []
    for i in range(len(data) - seq_len):
        x.append(data[i:i+seq_len, :-1])  # entradas
        y.append(data[i+seq_len, -1])     # salida
    return np.array(x), np.array(y)

X_train, y_train = create_sequences(train_data)
X_val, y_val = create_sequences(val_data)

# === Definir modelo LSTM profundo ===
model = Sequential([
    tf.keras.layers.Input(shape=(24, 5)),
    LSTM(100, return_sequences=True),
    Dropout(0.1),
    LSTM(100, return_sequences=False),
    Dense(50, activation='relu'),
    Dense(1)
])

lr_schedule = tf.keras.optimizers.schedules.PiecewiseConstantDecay([100, 200], [0.01, 0.005, 0.001])
optimizer = Adam(learning_rate=lr_schedule)
model.compile(optimizer=optimizer, loss='mse')

# Entrenamiento
model.fit(X_train, y_train, epochs=350, batch_size=16, validation_data=(X_val, y_val), verbose=0)

# Guardar modelo
model.save('modelo_LSTM.keras')
```

Figura 17 implementación para LSTM (Estructura del modelo).

Con esta configuración de entrenamiento:

- Optimizador: Adam con planificación adaptable.
- Función de pérdida: Error Cuadrático Medio (MSE).
- Épocas: 350 (entrenamiento prolongado para convergencia).

- Batch size: 16 (balance entre velocidad y estabilidad).
- Validación en cada época con conjunto de validación.

Finalmente, el modelo final se guardó en formato Keras (modelo_LSTM.keras), para ser utilizado posteriormente.

Tras la implementación de los modelos, se llevó a cabo un proceso de optimización con el objetivo de mejorar su precisión y capacidad predictiva. Esta etapa consistió en la realización de múltiples pruebas y ajustes sistemáticos para identificar la combinación de hiperparámetros que ofreciera los mejores resultados en la reproducción de los valores reales. En la Tabla 3 se presentan los hiperparámetros óptimos encontrados para cada uno de los modelos evaluados.

Tabla 3 hiperparámetros usados en los modelos

SVM	**Bosques aleatorios**	**LSTM**
C=1, *epsilon=0.01,* *loss='squared_epsilon_insensitive',* *max_iter=10000,* *fit_intercept=True,* *tol=0.0001,* *random_state=42*	n_estimators=200, max_depth=15, criterion='squared_error', max_features=1.0, bootstrap=True, random_state=42, min_samples_split=2, min_samples_leaf=1	Layers 2 con 100 units cada una, Activación 'tanh', dropout=0.1, capa densa de 50 units con 'relu', salida de 1 unit con 'linear', optimizador “Adam”

4.6 Comparación de resultados

Para este análisis se seleccionará un conjunto de datos tanto climatológicos, así como de producción fotovoltaica y se tomará una parte de estos como población muestra. Con esta población muestra seleccionada se alimentarán a los tres modelos inteligentes (“*LSTM*”, “*SVM*” y “*Random Forest*”) y posterior a esto se realizarán una

serie de predicciones utilizando conjuntos de datos que conformen la sucesión de un día entero (24 horas) basados en el conjunto de datos con el que no fueron alimentados los modelos inteligentes, para así poder realizar un análisis de resultados acorde ya que los modelos inteligentes estarán alimentados con la misma información.

Una vez realizada las pruebas correspondientes los resultados obtenidos con los modelos son las siguientes:

Día lluvioso

Durante el día lluvioso, se observaron niveles significativamente bajos de producción fotovoltaica reales (WH), con un comportamiento plano y valores mínimos durante la mayor parte del día (Figura 18). Esto concuerda con lo esperado en condiciones meteorológicas adversas, donde la radiación solar se ve notablemente afectada por la nubosidad y las precipitaciones.

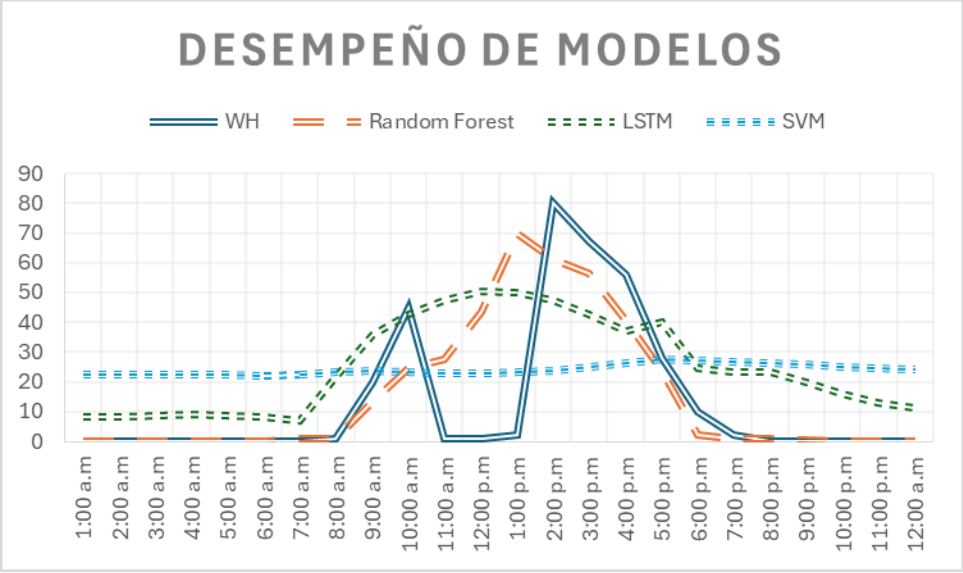


Figura 18 desempeño de modelos (Día lluvioso).

Al analizar el desempeño de los modelos inteligentes, se pueden destacar varios puntos importantes:

- El modelo de bosques aleatorios logró seguir de manera más cercana la tendencia del valor real (WH) durante la mayoría de las horas del día, especialmente en la franja horaria de 10:00 a.m. a 4:00 p.m. donde, aunque hubo sobrestimación, esta fue moderada.
- El modelo LSTM también replicó adecuadamente la forma de la curva, aunque con un leve sesgo hacia la sobreestimación durante las horas de mayor producción teórica (alrededor del mediodía).
- El modelo SVM presentó un comportamiento relativamente constante, con una línea de predicción casi plana a lo largo del día. Esto sugiere una baja capacidad del modelo para ajustarse a los cambios dinámicos en las condiciones climáticas propias de un día lluvioso.

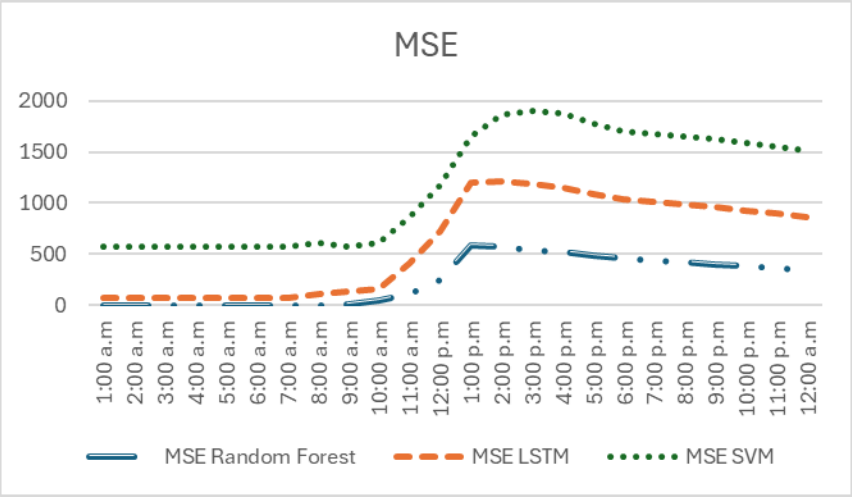


Figura 19 MSE de modelos (Día lluvioso).

En cuanto al análisis del error cuadrático medio (MSE)(Figura 19):

- El modelo de bosques aleatorios presentó el menor error general, destacando su robustez ante las condiciones meteorológicas adversas. El MSE se mantuvo por debajo de los 600 durante todo el día.
- El modelo LSTM tuvo un desempeño intermedio, con valores de MSE que se incrementaron notablemente a partir de las 11:00 a.m., alcanzando picos por

encima de 1200, lo que sugiere que su predicción fue más afectada por los cambios abruptos del clima.

- El modelo SVM mostró el mayor error, con valores de MSE superiores a 1800 durante gran parte de las horas pico, reflejando su poca adaptación a los patrones de un día con condiciones climáticas atípicas.

Los resultados indican que, en condiciones de lluvia, bosques aleatorios es el modelo con el desempeño general optimo, tanto por su mayor capacidad de seguimiento de los datos reales como por su menor error, LSTM sigue siendo una alternativa razonable, mientras que SVM se muestra poco efectivo bajo este tipo de condiciones.

Día normal

Para el analisis de resultados en este tipo de día se tomarán en cuenta los valores definidos en la Tabla 2.

Aquí se observa el patrón típico de producción fotovoltaica en forma de campana, con un pico máximo al mediodía (12:00 p.m.–2:00 p.m.) y sin perturbaciones climáticas significativas(Figura 20).

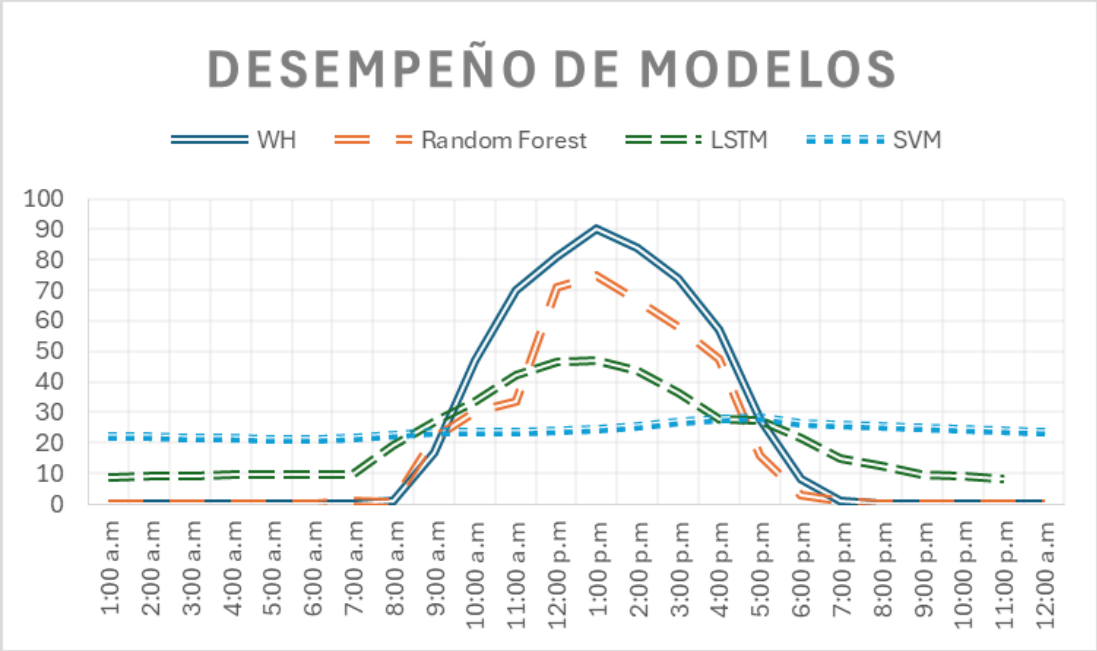


Figura 20 desempeño de modelos (Día normal).

En cuanto al comportamiento de los modelos inteligentes podemos visualizar lo siguiente:

- **Bosques aleatorios** demostró un excelente ajuste a los valores reales durante todo el ciclo diurno, particularmente en las horas críticas de transición (8:00-10:00 a.m. y 4:00-6:00 p.m.). Sin embargo, mostró una subestimación durante el pico de producción solar.
- **LSTM** capturó de buena forma la dinámica de crecimiento y decrecimiento de la producción, en los puntos críticos de crecimiento y decrecimiento (8:00-10:00 a.m. y 4:00-6:00 p.m.) sin embargo durante estos picos de producción subestimó está de manera significativa.

- **SVM** presentó un comportamiento más rígido, con subestimación consistente durante las horas de mayor producción (10:00 a.m.-3:00 p.m.)

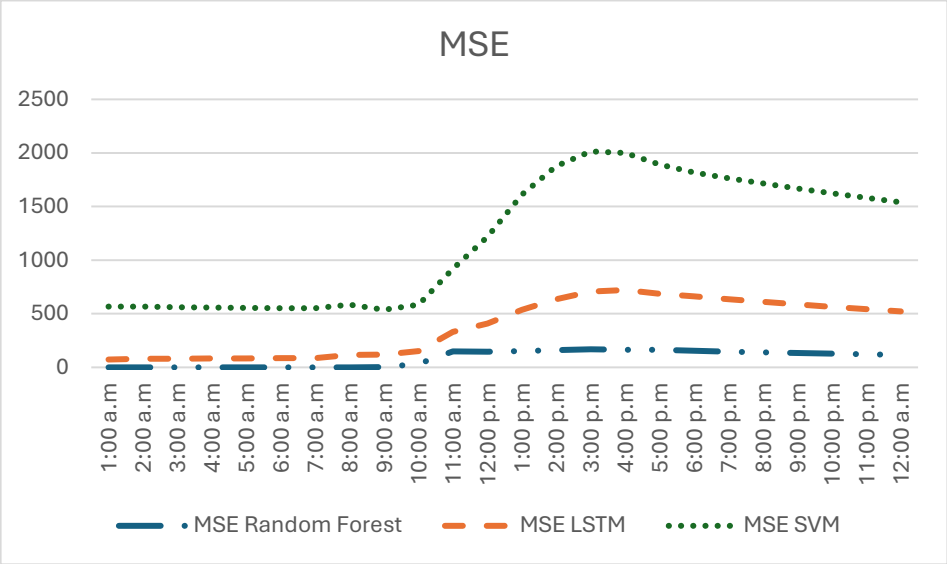


Figura 21 MSE de modelos (Día normal).

En cuanto al comportamiento mostrado en la gráfica de MSE observamos los siguientes resultados(Figura 21):

- Bosques aleatorios fue consistentemente mejor en MSE global, con un error más estable.
- LSTM tuvo los malos picos de error, pero destacó en ventanas específicas (10:00-11:00 a.m.).
- SVM mostró un MSE elevado durante todo el ciclo diurno, confirmando su limitación para modelar no-linealidades.

Con este conjunto de datos podemos ver que durante entornos comunes sin variables atenuantes a la producción de energía Bosques aleatorios presenta una predicción más fiel a los valores de producción, en cuanto a LSTM presentó un rendimiento relativamente bueno, aunque sumamente mejorable, y concluyendo con SVM con un mal desempeño siendo incapaz de realizar predicciones que imiten el comportamiento de campana de los datos.

Día nublado

Durante el día nublado, la producción fotovoltaica mostró un patrón atípico con fluctuaciones pronunciadas. A diferencia de los días normales, no se observó una curva de producción en forma de campana, sino variaciones abruptas causadas por cambios en la densidad nubosa. La producción se mantuvo en niveles bajos entre las 9:00 y 15:00 horas, con caídas repentinas alrededor de las 10:30 y 13:45 horas (Figura 22).

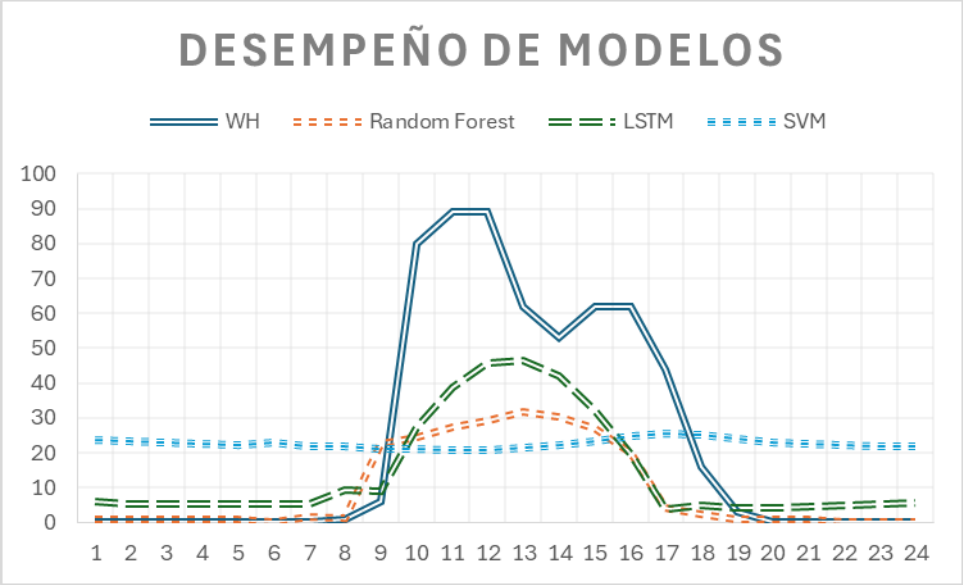


Figura 22 desempeño de modelos (Día nublado).

En la gráfica anterior pudimos visualizar los siguientes comportamientos:

- El modelo LSTM demostró su fortaleza principal, logro detectar las caídas abruptas de producción con mayor precisión que los otros modelos, especialmente en los momentos críticos de cambios rápidos en la nubosidad. Sin embargo, mostró una tendencia a subestimar la producción durante los periodos prolongados de alta generación.
- En el modelo de Bosques aleatorios se presentó un comportamiento más estable pero menos sensible a los cambios rápidos. Mantuvo un error constante

durante todo el periodo, pero falló en capturar las fluctuaciones abruptas, suavizando demasiado la curva de predicción. Su principal ventaja fue la consistencia en su desempeño.

- El modelo SVM mostró el peor desempeño, generando predicciones casi planas que no reflejaron las variaciones reales. Alcanzó desviaciones durante las horas críticas, demostrando ser inadecuado para estas condiciones climáticas variables.

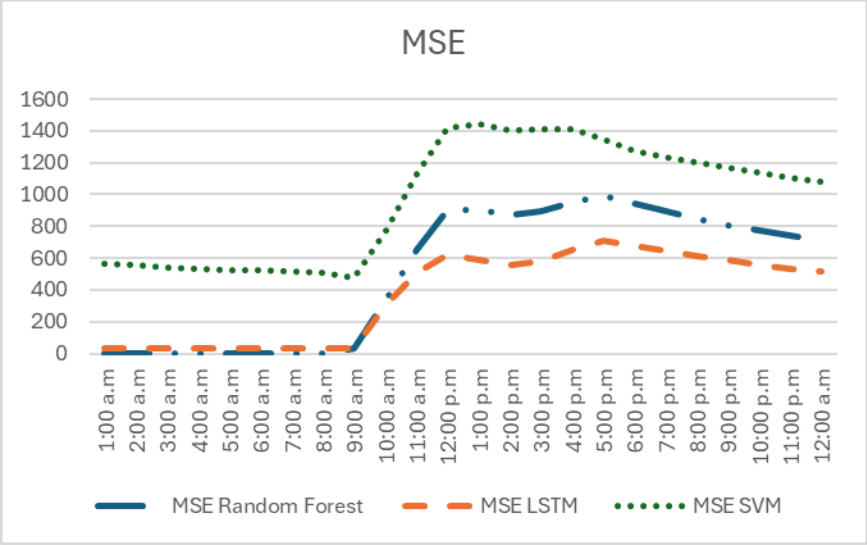


Figura 23 MSE de modelos (Día nublado).

En cuanto al desempeño en MSE se visualizaron los siguientes resultados (Figura 23):

- LSTM presentó un MSE promedio bueno, con una notable variabilidad horaria. Alcanzó su mejor desempeño durante cambios abruptos de nubosidad, demostrando su capacidad para capturar variaciones rápidas. Sin embargo, mostró picos de error en periodos de baja producción sostenida, donde tendió a sobrestimar los valores reales.

- Bosques aleatorios mantuvo un MSE más estable, con fluctuaciones limitadas. Este modelo mostró consistencia en su desempeño, pero con una incapacidad

para reaccionar a cambios bruscos, manteniendo un error sistemático todo el día.

- SVM registró los peores resultados, con MSE promedio malo y picos muy altos. Su rigidez estructural lo hizo insensible a las variaciones climáticas, generando predicciones casi planas que no reflejaban la realidad operativa.

En condiciones de nubosidad variable, el modelo LSTM demostró mayor capacidad para seguir las fluctuaciones abruptas en la producción, aunque con cierta tendencia a sobrestimar durante periodos prolongados de baja generación. Bosques aleatorios mantuvo un comportamiento estable pero menos preciso ante cambios rápidos, subestimando consistentemente la producción. Por su parte, SVM nuevamente mostró un desempeño deficiente, generando predicciones rígidas que no lograron capturar la dinámica variable típica de un día nublado.

Día parcialmente nublado

La producción fotovoltaica mostró un patrón intermedio con fluctuaciones moderadas, presentando un ascenso matutino con pequeñas oscilaciones (6:00-9:00), variabilidad marcada al mediodía (10:00-14:00) con picos durante despejes, y descenso vespertino más estable (15:00-18:00). A diferencia de días completamente nublados, las transiciones fueron progresivas y sin caídas bruscas, mostrando mayor

producción base, pero conservando una estructura menos definida que en días soleados (Figura 24).

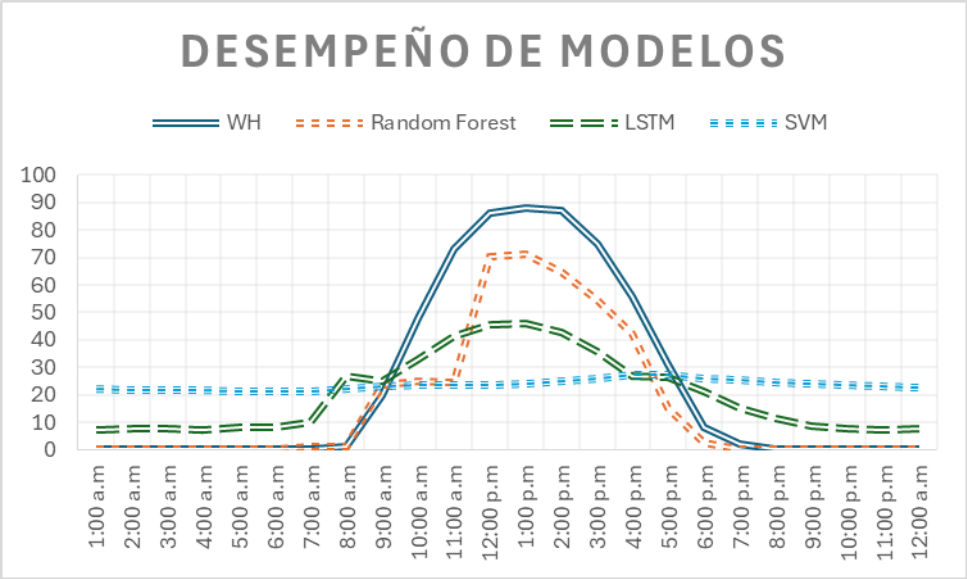


Figura 24 desempeño de modelos (Día parcialmente nublado).

Los modelos mostraron capacidades diferenciadas para capturar el patrón intermitente de producción fotovoltaica durante este día:

- El modelo de Bosques aleatorios mostró un comportamiento estable, siguiendo adecuadamente la tendencia general, pero suavizando las variaciones rápidas. Subestimó los picos de producción durante los despejes, aunque mantuvo un error bajo en periodos estables. Su fortaleza fue la consistencia, pero falló en capturar fluctuaciones abruptas.
- El modelo LSTM mostró capacidad limitada para replicar el comportamiento real, subestimando consistentemente la producción durante los periodos clave. Aunque identificó correctamente la ocurrencia de los picos de producción (11:00-14:00 horas), falló en predecir su magnitud exacta, mostrando valores por debajo de los datos reales. Durante las fases de transición presentó un retraso en ajustar sus predicciones, y en los periodos estables (9:00-10:30 y 15:00-17:00) mantuvo una sobreestimación constante.

Estos resultados indican que, si bien reconoció los patrones generales, no logró la precisión esperada en la cuantificación de los valores de producción

- SVM presentó el peor desempeño, generando predicciones casi planas alrededor. No logró adaptarse a la variabilidad real, mostrando un error constante. Su rigidez lo hizo ineficaz para este escenario dinámico.

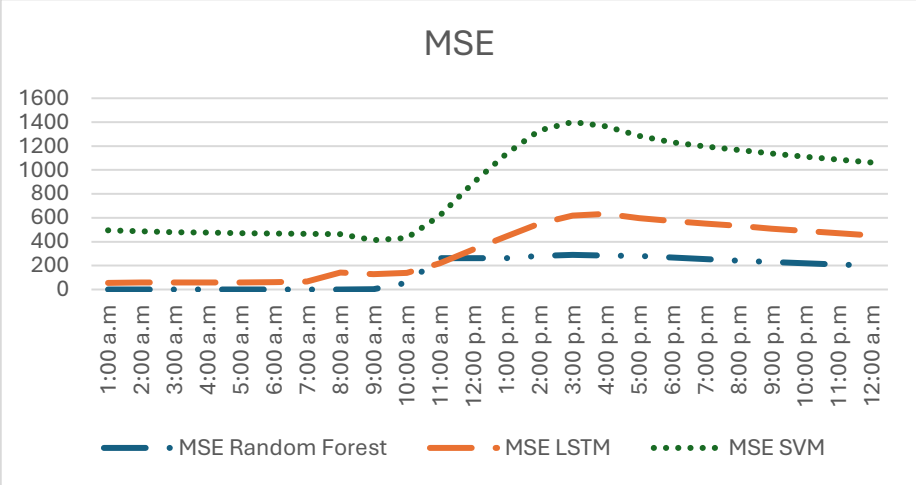


Figura 25 MSE de modelos (Día parcialmente nublado).

En cuanto al desempeño en las gráficas de error MSE se obtuvieron los siguientes resultados (Figura 25):

- Bosques aleatorios registró el menor MSE global, demostrando consistencia en sus predicciones. Su error se mantuvo estable a lo largo del día, con ligeros incrementos durante los periodos de mayor variabilidad (mediodía), pero siempre dentro de un rango controlado. Esto confirma su robustez como modelo predictivo.
- LSTM mostró un MSE intermedio, con valores más altos que Bosques aleatorios pero inferiores a SVM. Presentó picos de error durante las transiciones climáticas (cambios bruscos de nubosidad), evidenciando cierta dificultad para ajustarse rápidamente a condiciones variables, aunque superó claramente a SVM en precisión general.

- SVM obtuvo el peor desempeño con el MSE más elevado. Sus errores fueron consistentemente altos durante todo el periodo, especialmente en horas de mayor producción (10:00-14:00), confirmando su inadecuación para modelar la relación no lineal entre variables climáticas y producción energética.

En condiciones de nubosidad variable, los modelos mostraron comportamientos diferenciados. El modelo de Bosques aleatorios demostró ser el más robusto, manteniendo una alta consistencia en sus predicciones y el menor MSE global, aunque con cierta dificultad para capturar las fluctuaciones más abruptas. LSTM, si bien identificó correctamente los patrones temporales y los periodos de transición, presentó limitaciones en la precisión cuantitativa, con subestimaciones en los picos clave y cierto retraso en ajustarse a los cambios. Por su parte, SVM confirmó su total inadecuación para este escenario, con errores sistemáticos elevados y una incapacidad manifiesta para adaptarse a la variabilidad característica de este tipo de día.

Día soleado

La producción mostró la clara curva en forma de campana característica de los días despejados, alcanzando su pico máximo entre las 12:00 y 14:00 horas. Se observó un ascenso progresivo desde las 6:00 horas y un descenso simétrico hacia el atardecer, con variaciones mínimas durante las horas centrales (Figura 26).

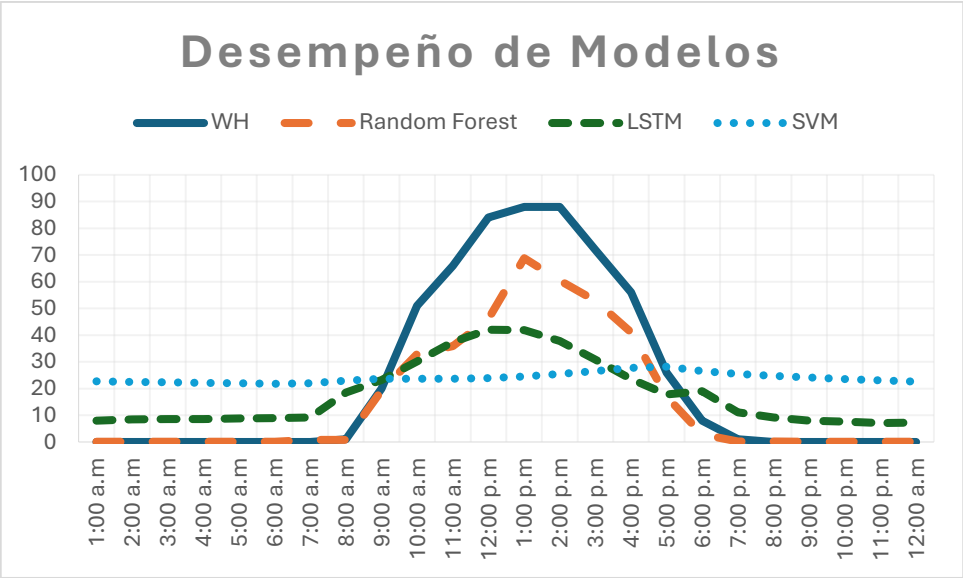


Figura 26 Desempeño de modelos (Día soleado).

- El modelo de Bosques aleatorios demostró un buen desempeño en condiciones de cielo despejado. Su arquitectura basada en árboles capturó la relación no lineal entre las variables climáticas y la producción energética. Durante las horas centrales del día (11:00 a 15:00), donde la producción alcanzó su máximo. Solo en los extremos del día mostró una ligera subestimación, debido probablemente a la menor cantidad de ejemplos similares en los datos de entrenamiento para estas condiciones.
- La red neuronal recurrente mostró un comportamiento característico en este escenario. Si bien logró mantener el pico de producción, introdujo pequeñas oscilaciones artificiales en la fase estable de generación. Estas variaciones, inexistentes en los datos reales, revelan cierta tendencia al sobreajuste. Durante las transiciones matutinas y vespertinas, el modelo sobrestimó consistentemente la producción, posiblemente por su dificultad para ajustarse rápidamente a los cambios en el ángulo de incidencia solar.
- El modelo de Support Vector Machines mostró graves limitaciones para este tipo de condiciones. Su predicción se mantuvo prácticamente plana, completamente ajena a la curva de campana característica de los días soleados. El error alcanzó su máximo precisamente durante las horas de mayor

producción (12:00 a 14:00), cuando más se alejó de los valores reales. Este desempeño confirma que los modelos lineales tradicionales son incapaces de capturar las relaciones complejas presentes en sistemas fotovoltaicos, incluso en condiciones climáticas estables.

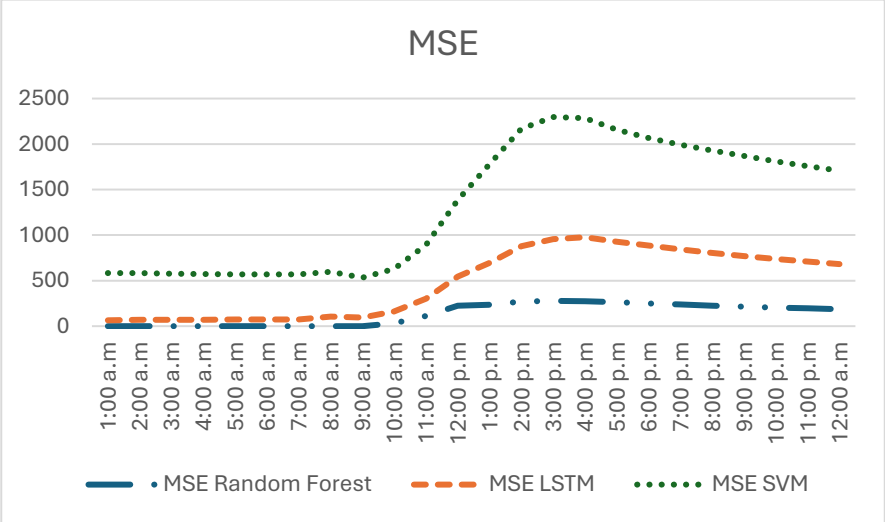


Figura 27 MSE de modelos (Día soleado).

Tras la visualización de la gráfica de MSE se rescataron las siguientes dinámicas en los modelos (Figura 27):

- El modelo de Bosques aleatorios demostró el mejor desempeño en MSE durante el día soleado, manteniendo valores consistentemente bajos. Alcanzó su mínimo error precisamente en las horas de máxima producción (12:00-14:00), confirmando su excelente ajuste a condiciones estables. Solo mostró ligeros incrementos durante las transiciones matutinas y vespertinas, manteniendo siempre la ventaja sobre otros modelos. Su curva de error fue la más plana y estable de los tres.
- LSTM presentó un MSE intermedio con notable variabilidad horaria. Si bien logró su mejor precisión durante el pico solar, mostró picos de error significativos en horas marginales. La gráfica revela frecuentes oscilaciones en su error, evidenciando su sensibilidad a cambios en las condiciones de irradiación, incluso en días despejados.

- SVM registró el peor desempeño con MSE consistentemente alto. La curva muestra valores críticamente elevados durante todo el ciclo solar, alcanzando su máximo coincidiendo con las horas de máxima producción. La gráfica confirma su total incapacidad para adaptarse al patrón característico de los días soleados.

En condiciones de cielo despejado, el modelo de Bosques aleatorios demostró ser el más confiable, manteniendo alta precisión durante todo el ciclo solar y minimizando errores incluso en las horas de máxima producción. LSTM mostró capacidad para seguir la tendencia general, pero con fluctuaciones artificiales y mayor sensibilidad a transiciones, revelando limitaciones en su consistencia. Por su parte, SVM confirmó su total inadecuación para este escenario, con errores sistemáticos que lo invalidan como opción predictiva.

A partir del análisis gráfico y numérico realizado para cada uno de los días representativos, se puede observar un comportamiento diferenciado de los modelos de predicción frente a las condiciones climáticas. Mientras algunos modelos mantienen un desempeño consistente en distintos escenarios, otros muestran limitaciones importantes en situaciones específicas, como días con alta variabilidad o radiación extrema. Para resumir estos hallazgos y facilitar la comparación general, se presenta la **¡Error! No se encuentra el origen de la referencia.** que sintetiza el desempeño de cada modelo según el tipo de día analizado utilizando el valor obtenido en MSE para cada modelo, entre menor sea el valor obtenido menor diferencia entre los valores predichos y los reales se habrá encontrado.

Tabla 4 Comparativa de resultados del desempeño de los modelos por tipo de día.

Modelo	Bosques aleatorios	LSTM	SVM
Lluvioso.	*MSE = 349.32*	*MSE = 512.94*	*MSE = 655.03*
Normal.	*MSE = 116.32*	*MSE = 107.31*	*MSE = 1017.36*
Nublado.	*MSE = 707.88*	*MSE = 515.31*	*MSE = 1079.32*
Parcialmente nublado	*MSE = 200.21*	*MSE = 453.14*	*MSE = 1062.35*
Soleado	*MSE = 187.55*	*MSE = 493.25*	*MSE = 1026.84*

4.7 Análisis

El análisis del desempeño de los modelos inteligentes utilizados para la predicción de producción fotovoltaica revela comportamientos diferenciados según las condiciones meteorológicas y la estructura interna de cada algoritmo. El modelo de Bosques aleatorios demostró una fuerte consistencia en sus predicciones, especialmente en días normales y soleados, donde la producción presenta patrones predecibles. En estos casos, logró errores bajos y una buena correspondencia con la curva real de producción. Sin embargo, en días nublados o parcialmente nublados, donde la radiación presenta una mayor variabilidad, Bosques aleatorios tendió a suavizar las fluctuaciones, mostrando una menor sensibilidad a cambios abruptos en las condiciones. Esto resultó en una subestimación de los picos de producción,

principalmente cuando se produjeron despejes rápidos. Su error cuadrático medio (MSE) fue estable y plano a lo largo del día, lo cual indica que el modelo capturó con éxito patrones promedio, pero no logró adaptarse completamente a las variaciones de alta frecuencia que caracterizan días con intermitencias en la radiación solar.

El modelo LSTM, por su parte, mostró una respuesta más dinámica frente a escenarios de alta variabilidad climática. En días nublados o parcialmente nublados, fue capaz de seguir con mayor precisión las caídas o aumentos rápidos en la producción, lo que evidencia su fortaleza para modelar secuencias temporales complejas. Sin embargo, esta misma sensibilidad generó ciertos inconvenientes en días soleados, donde introdujo oscilaciones no presentes en los datos reales, indicando una posible sobrestimación de las pequeñas variaciones o ruido en los datos de entrada. Durante periodos prolongados sin cambios significativos, LSTM tendió a sobreestimar la producción, lo que revela una debilidad para adaptarse a condiciones estacionarias. Su comportamiento en el MSE fue más irregular, con picos importantes en momentos de transición, aunque en general se mantuvo por debajo del error mostrado por el modelo SVM.

El modelo SVM fue el que presentó el rendimiento más débil en todos los escenarios considerados. Sus predicciones se caracterizaron por una escasa variabilidad, lo que se tradujo en curvas de producción planas o poco dinámicas, sin capacidad para seguir el patrón real de producción solar. Esta limitación fue especialmente notoria en días soleados, donde la forma de campana de la producción fue ignorada por el modelo, que mantuvo predicciones casi constantes. En días nublados, tampoco respondió adecuadamente a las caídas abruptas ni a los incrementos rápidos, lo que refleja una estructura poco flexible para adaptarse a relaciones no lineales entre las variables climáticas y la producción. Su MSE fue el más alto de los tres modelos, con picos que coincidieron con las horas de mayor variabilidad, donde su desfase respecto a la realidad fue más evidente.

Capítulo V

Conclusiones

En este capítulo final se hace el análisis y comparación de desempeños final de los resultados con lo cual, se espera identificar al modelo inteligente con el desempeño optimo entre "*LSTM*", "*SVM*" y "*Random Forest*". Para esto se harán diversas pruebas en cada modelo para observar sus predicciones ante casos meteorológicos diferentes y comparar sus resultados con los resultados esperados.

5.1 Conclusiones

Este trabajo de tesis ha evaluado exhaustivamente el desempeño de tres modelos de aprendizaje automático (Bosques aleatorios, LSTM y SVM) para predecir la producción de energía fotovoltaica bajo diversas condiciones climáticas, se programó dichos modelos y se optimizaron sus hiperparámetros para busca un buen desempeño posible, a la par que se recopilaron datos útiles para los entrenamientos y sus posteriores predicciones. A lo largo de múltiples escenarios meteorológicos diferenciados, se compararon las predicciones con valores reales de producción. Los resultados obtenidos proporcionan información valiosa sobre el desempeño de modelos inteligentes para su aplicación en la industria y así poder encontrar formas innovadoras y eficientes para mejorar la producción y eficacia de las plantas fotovoltaicas.

El análisis de los modelos predictivos bajo diversas condiciones climáticas reveló patrones determinantes para la selección de algoritmos en sistemas fotovoltaicos. Los resultados demuestran que el desempeño de cada modelo varía significativamente

según las condiciones ambientales, lo que exige un enfoque diferenciado para cada escenario operativo.

En condiciones climáticas extremas (días lluviosos y nublados), el modelo de Bosques aleatorios emergió como la solución más robusta. Su arquitectura basada en árboles demostró una capacidad excepcional para mantener predicciones estables con errores controlados, incluso cuando la producción fotovoltaica presentaba comportamientos atípicos. Particularmente en días lluviosos, donde otros modelos mostraban desviaciones significativas, Bosques aleatorios mantuvo un MSE consistentemente bajo, confirmando su superioridad en entornos adversos. Sin embargo, su principal limitación apareció en la incapacidad para capturar fluctuaciones abruptas durante cambios rápidos de nubosidad, donde su naturaleza conservadora lo llevó a suavizar en exceso las predicciones.

El modelo LSTM mostró un comportamiento paradójico. Si bien en teoría su arquitectura recurrente lo hace ideal para patrones temporales, en la práctica exhibió limitaciones notables. Durante días con variabilidad intermitente (parcialmente nublados), presentó retrasos en ajustar sus predicciones y errores sistemáticos en la cuantificación de picos de producción. Su tendencia al sobreajuste se hizo evidente en días soleados, donde introdujo oscilaciones artificiales inexistentes en los datos reales. No obstante, su capacidad para detectar cambios abruptos lo convierte en un complemento valioso para sistemas de alerta temprana.

SVM confirmó su total inadecuación para todos los escenarios analizados. Sus errores sistemáticos y su incapacidad para aprender patrones no lineales lo descartan como opción viable. Su rigidez conceptual lo hace insensible a las dinámicas fundamentales de los sistemas fotovoltaicos, particularmente en condiciones variables

Con este análisis se demostró que, si bien cada modelo tiene características distintivas, Bosques aleatorios ofreciendo el mejor equilibrio entre precisión y estabilidad en condiciones adversas (lluvia/nubosidad), manteniendo un MSE significativamente más bajo que LSTM y SVM, manejando eficientemente los

comportamientos atípicos de producción fotovoltaica, en días soleados, alcanzó una buena precisión replicando fielmente la curva de campana característica y mostrando superioridad computacional, siendo más eficiente que LSTM en recursos requeridos. Aunque LSTM demostró potencial para capturar cambios abruptos, su inconsistencia en estimaciones cuantitativas y tendencia al sobreajuste lo relegaron a un rol secundario. SVM, por su parte, resultó inviable en todos los escenarios.

La implementación de Bosques aleatorios promete mejorar la precisión en las predicciones, respecto a métodos convencionales, optimizando tanto la gestión energética como los programas de mantenimiento, convirtiéndose así en la solución más confiable y práctica para la operación diaria de plantas fotovoltaicas en entornos reales con climas variables.

5.2 Perspectivas del trabajo

Este estudio ofrece aplicaciones prácticas inmediatas para plantas fotovoltaicas, donde los modelos desarrollados pueden implementarse como sistemas predictivos para mejorar la gestión energética. Al proporcionar estimaciones precisas de producción, permitirán una mejor planificación del almacenamiento en baterías y una administración más eficiente de la distribución energética. Además, al comparar las predicciones con la producción real, podrán detectarse posibles fallos en los paneles o equipos cuando existan discrepancias significativas. Estas aplicaciones contribuirán a optimizar la operación de las plantas solares y reducir pérdidas por mala planeación o fallos no detectados.

El proyecto enfrentó limitaciones importantes que deben considerarse. Los altos costos computacionales representaron un desafío, particularmente en el entrenamiento de modelos complejos como LSTM, donde el tiempo y recursos requeridos aumentaban considerablemente con el volumen de datos. Otra limitación

fue la escasez de datos para condiciones climáticas extremas (días muy nublados, lluviosos o excepcionalmente soleados), lo que afectó la precisión de los modelos en estos escenarios. Un conjunto de datos más amplio y diverso podría mejorar significativamente los resultados, especialmente para eventos meteorológicos poco frecuentes.

5.3 Trabajos futuros

Como trabajo futuro, se plantea explorar otros modelos inteligentes para comparar su desempeño con los enfoques ya evaluados. También se propone mejorar las arquitecturas existentes mediante técnicas avanzadas de optimización y regularización. Es fundamental ampliar los conjuntos de datos, incorporando más registros de condiciones climáticas extremas y variables adicionales que puedan afectar la producción. Finalmente, la implementación en plantas reales permitirá validar y ajustar los modelos según necesidades operativas concretas, cerrando la brecha entre la investigación y su aplicación práctica en la industria fotovoltaica.

Referencias

Arna, C. (2021). *Redes neuronales recurrentes: Análisis de los modelos especializados en datos secuenciales [Libro]*. (J. M. Streb, Ed.) University of CEMA, CEMA Working Papers: Serie Documentos de Trabajo 797.

Berry, M. J., & S.Linoff, G. (2004). *Data Mining Techniques[Libro]*. Indiana: Wiley Publishing Inc.

Betancourt, G. A. (2005). *Las máquinas de soporte vectorial (SVMs)[Libro]*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning[Libro]*. Springer Science+Business Media.

Breiman, L. (2001). *Random forests[Libro]*. Machine Learning, Springer.

Bueno, B. M. (2017). *Predicción semanal de predios de la energía eléctrica utilizando bosque aleatorios[Tesis de Maestría]*. Madrid España: Universidad Politécnica de Madrid.

Chollet, F. (2021). *Deep Learning with Python[Libro]*. Manning Publications.

Estupiñan, J. J., Giral, D., & Santa, F. M. (2016). *Implementation of algorithms based on support vector machine (SVM) for electric systems: topic review[Tesis]*.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow[Libro]*. O'Reilly Media.

Gonzalez, L. (23 de 08 de 2019). *AprendeIA*. Obtenido de https://aprendeia.com/kernel-maquinas-vectores-de-soporte-clasificacion-regresion/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning[Libro]*. The MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction[Libro]*. Springer.

IEA PVPS. (2022). *Trends in photovoltaic applications 2022 [Informe]*. Paris: IEA PVPS.

Krishna, P. S., & Kumar, s. K. (2015). *Normalization: A Preprocessing Stage[Artículo]*. Veer Surendra Sai University of Technology.

León, E. C. (2016). *Introducción a las máquinas de vector soporte (SVM) en aprendizaje supervizádo[Libro]*. Universidad de Zaragoza.

León, M. G. (2018). *Análisis de Sensibilidad Mediante Random Forest[Tesis]*. Madrid: Universidad Politécnica de Madrid.

Organizacion de las Naciones Unidas [Informe]. (2021). *Naciones Unidas*. Obtenido de https://www.un.org/sustainabledevelopment/es/energy/

R, D. K., Prakash, K., P. Abirama, S., & Sathya, S. (2023). *A hybrid machine learning model for solar power forecasting*. New Prince Shri Bhavani College Of Engineering and Technology.

R. Kumar, D., P. Sundari, A., Parakash, K., & Sathya, S. (2023). *Ahybrid machine learning model for solar power forecasting [Artículo]*. Erode, India: New Prince Shri Bhavani College Of Engineering and Technology.

Rodríguez Montequín, M. T., Álvarez Cabal, J. V., & Mesa Fernández. (2003). Metodologías para la realización de proyectos de data mining[Tesis]. Universidad Pública de Navarra.

Runger, G. C., & Montgomery, D. C. (2014). *Applied Statistics and Probability for Engineers[Libro]*. John Wiley & Sons, Inc.

Smola, A. J., & Schölkopf, B. (2004). *A tutorial on support vector regression[Libro]*. Statistics and Computing, Springer.

Suárez, C. J., López, M. V., & Ramos, F. J. (2009). Automatic Generation of Learning Outcomes based on Long Short–Term Memory Artificial Neural Network[Artículo].

Subramanian, E., Karthik, M., Krishna, G., Kumar, V., & Prasath, D. (2016). *Solar power prediction using machine learning[Artículo].* Sri Shakthi Institute of Engineering and Technology.

Visser, L. (2023). Forecasting to Welcome the Solar Era[Tesis].

Yu, J., Li, X., Yang, L., Li, L., Huang, Z., Shen, K., . . . Yang, X. (2024). *Deep Learning Models for PV Power Forecasting: Review [Artículo].* Wuhan, China: Huazhong University of Science and Technology,.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). *Forecasting with artificial neural networks: The state of the art[Artículo].* International Journal of Forecasting, Elsevier.