

ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

ANÁLISIS Y PREDICCIÓN DE LA CAPACIDAD DE IRRADIACIÓN
SOLAR MEDIANTE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

Alex Bladimir Chancúsig Quinatoa

GUAYAQUIL - ECUADOR

Año: 2023

DEDICATORIA

A mis padres y mi hermano.

AGRADECIMIENTOS

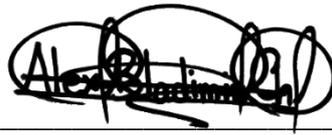
Mi más sincero agradecimiento a mis padres y hermano por su apoyo incondicional.

A la ing. Neira D. y a la Lic. Sanmartín E. por su paciencia, motivación y soporte brindado.

De igual manera a José Córdova Ph.D. por su constante retroalimentación y enseñanzas impartidas durante el desarrollo del presente proyecto.

DECLARACIÓN EXPRESA

"Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; Alex Bladimir Chancúsig Quinatoa y doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"



Alex Bladimir Chancúsig Quinatoa

COMITÉ EVALUADOR

PhD. José Córdova García
PROFESOR TUTOR

Allan Avendaño, MSc.
PROFESOR EVALUADOR

RESUMEN

La energía solar se ha convertido en la principal fuente de energía renovable como respuesta a la creciente demanda energética que se experimenta a nivel mundial y la necesidad de reducir la huella de carbono generada por el uso de combustibles fósiles, a través de la implementación de sistemas fotovoltaicos a pequeña y gran escala cuya producción dependerá de la cantidad de irradiación solar disponible en un área determinada considerando la influencia de factores externos como las estaciones del año, condiciones ambientales, ubicación geográfica, entre otros. En Ecuador no se dispone de una base de datos de irradiación solar que se encuentre actualizada y permita analizar eficientemente el potencial solar que puede ser aprovechado, en este proyecto se integraron modelos de Machine Learning para ajustar datos de irradiancia solar a partir de registros satelitales con respecto a mediciones en tierra de las diferentes estaciones meteorológicas distribuidas en la provincia de Pichincha, dando como resultado un conjunto de datos de irradiación solar que comprende todos los puntos coordinados dentro de la provincia con una resolución espacial de 4 Km. Se implementaron modelos LSTM estándar y Conv-LSTM en series de tiempo univariadas y secuencias de matrices 2D mediante el formato de imágenes respectivamente, con la finalidad de generar pronósticos de irradiancia solar a corto plazo, cuyo modelo resultante fue integrado en una aplicación web para el desarrollo de una herramienta de soporte para el cálculo y dimensionamiento de sistemas fotovoltaicos aplicable dentro de la provincia de Pichincha.

Palabras Clave: Energía solar, sistemas fotovoltaicos, irradiación solar, estación meteorológica, aplicación web.

ABSTRACT

Solar energy has become at the main source of renewable energy as the response to the growing energy demand experimented at worldwide, and the need of reduce the carbon footprint generated by the use of fossil fuels through the implementation of small and large-scale photovoltaic systems. The production depends on the amount of available solar irradiation in a given area considering the influence of external factors as the year seasons, environment conditions, geographic location, etc. Ecuador hasn't an updated solar irradiance database that allows an efficient analysis of the solar potential that can be used. In this project, Machine Learning models were integrated to adjust solar irradiance data from satellite records with regarding ground measurements of the different meteorological stations distributed in the province of Pichincha, resulting in a new solar irradiation dataset that includes all the coordinate points within the province with a spatial resolution of 4 km. Standard LSTM and Conv-LSTM models were implemented in univariate time series and 2D matrix sequences using the image format respectively. In order to generate short-term solar irradiance forecasts, the resulting model was integrated into a web application for the development of a support tool for the calculation and sizing of photovoltaic systems applicable within the province of Pichincha.

Keywords: *Solar energy, photovoltaic systems, solar irradiation, weather station, web application.*

ÍNDICE GENERAL

COMITÉ EVALUADOR.....	5
RESUMEN.....	I
<i>ABSTRACT</i>	II
ÍNDICE GENERAL.....	III
ABREVIATURAS	VII
SIMBOLOGÍA	IX
ÍNDICE DE FIGURAS.....	X
ÍNDICE DE TABLAS	XIV
CAPÍTULO 1	15
1. Introducción	15
1.1 Antecedentes.....	15
1.2 Descripción del problema	16
1.3 Justificación del problema.....	17
1.4 Solución propuesta	18
1.5 Objetivos.....	19
1.5.1 Objetivo General	19
1.5.2 Objetivos Específicos	19
1.6 Metodología	19
1.7 Resultados esperados	20
1.8 Dataset	20
1.8.1 Datos de estaciones meteorológicas.....	21
1.8.2 Datos satelitales	22
CAPÍTULO 2.....	24
2. MARCO TEÓRICO Y ESTADO DEL ARTE.....	24

2.1	Conceptos generales	24
2.1.1	Radiación solar.....	24
2.1.2	Irradiancia e irradiación	25
2.1.3	Tratamiento de datos meteorológicos	26
2.2	Fundamentos del problema	27
2.3	Soluciones basadas en datos	28
2.4	Análisis de variables para el desarrollo de modelos de predicción de irradiación 29	
2.5	Fuentes de datos	31
2.6	Pronóstico de irradiación solar a partir de datos satelitales	33
2.7	Light Gradient Boosting Machine (LGBM)	35
2.8	Línea base para predicción de series temporales – Holt Winters	37
2.9	Redes LSTM.....	38
2.10	Redes LSTM convolucionales	41
2.11	Optuna	43
2.12	Métricas para evaluación de modelos	45
2.13	Librerías y software a utilizar	47
CAPÍTULO 3.....		50
3.	DISEÑO E IMPLEMENTACIÓN.....	50
3.1	Exploración y validación de datos y fuentes	52
3.1.1	Validación de datos correspondientes a estaciones meteorológicas	52
3.1.2	Delimitación del área de estudio	53
3.1.3	Análisis exploratorio datos de estaciones meteorológicas	55
3.1.4	Descarga de datos satelitales	63
3.1.5	Análisis descriptivo de datos satelitales	65
3.2	Ajuste de datos locales y satelitales	69

3.3	Prototipos de métodos y modelos para elaboración de dataset artificial de irradiancia solar	71
3.3.1	Modelo Skforest	73
3.3.2	Modelo LGBM	76
3.4	Prototipos de métodos y modelos para pronóstico de series temporales	79
3.4.1	Método Holt	79
3.4.2	Método Holt-Winters.....	81
3.4.3	Redes LSTM one-step para series de tiempo univariadas	82
3.4.4	ConvLSTM	84
3.5	Infraestructura para el procesamiento y almacenamiento	87
3.6	Diseño base de plataformas y prototipos de visualización.....	87
3.6.1	Pantalla de inicio y selección de ubicación	87
3.6.2	Herramienta 1: Cálculo de ahorro eléctrico	89
3.6.3	Herramienta 2: Calcular sistemas fotovoltaicos	90
CAPÍTULO 4.....		93
4.	ANÁLISIS DE RESULTADOS.....	93
4.1	Recolección de datos y estrategias para validación del proyecto.....	93
4.1.1	Análisis de datos de estaciones meteorológicas	93
4.1.2	Tratamiento de valores atípicos en datos de estaciones meteorológicas	98
4.1.3	Análisis de datos satelitales	102
4.2	Puesta en marcha y funcionamiento.....	108
4.2.1	Modelos de regresión para construcción de dataset de irradiancia solar	108
4.2.2	Aplicación de modelo de regresión	118
4.2.3	Métodos estadísticos para pronóstico de irradiancia solar	123

4.2.4	Pronóstico de irradiancia solar por puntos	127
4.2.5	Pronóstico de irradiancia solar por áreas	130
4.2.6	Selección de modelo de pronóstico.....	136
4.3	Pruebas de funcionalidad	137
4.3.1	Funcionalidad de módulo “Análisis de irradiación solar”	138
4.3.2	Funcionalidad de módulo “Pronóstico de irradiación”.....	141
4.3.3	Funcionalidad de módulo “Calculadora solar”	144
4.4	Análisis costo/beneficio.....	149
4.4.1	Costos	149
4.4.2	Beneficios.....	150
CAPÍTULO 5.....		152
5.	Conclusiones Y Recomendaciones.....	152
	Conclusiones	152
	Recomendaciones	153
BIBLIOGRAFÍA.....		155
APÉNDICES		161

ABREVIATURAS

AC	Corriente Alterna
API	Interfaz de programación de aplicaciones
ARCERNNR	Agencia de Regulación y Control de Energía y Recursos Naturales no Renovables
AWS	Amazon Web Services
CART	Árboles de clasificación y regresión
ConvLSTM	Redes LSTM convolucionales
CSS	Hojas de estilo en cascada
CSV	Valores separados por comas
DC	Corriente Continua
DHI	Radiación difusa
DMQ	Distrito Metropolitano de Quito
DNI	Radiación directa
EEQ	Empresa Eléctrica Quito
EFB	Exclusive Feature Bundling
ESPOCH	Escuela Politécnica del Chimborazo
FONAG	Fondo para la Protección del Agua
GBM	Gradient boosting machine
GHI	Radiación global
GOSS	Grandient-based One-Side Sampling
HTML	Lenguaje de Marcas de Hipertexto
IDEAM	Instituto de Hidrología, Meteorología y Estudios Ambientales
IIGE	Instituto de investigación Geológico y Energético
INAMHI	Instituto Nacional de Meteorología e Hidrología
INER	Instituto Nacional de Eficiencia Energética y Energías Renovables
LGBM	Light Gradient Boosting Machine
LSTM	Long short-term memory
MAE	Error absoluto medio
MBE	Media del error del sesgo
MSE	Error cuadrático medio

NASA	National Aeronautics and Space Administration
NREL	Laboratorio Nacional de Energía Renovable
NSRDB	National Solar Radiation Database
PVGIS	Photovoltaic Geographical Information System
R2	Coeficiente de determinación
RMSE	Error de raíz cuadrada media
RNN	Redes neuronales recurrentes
SEDC	Sistema de Estandarización de Datos Hidroclimáticos Crudos
SRB	Surface Radiation Budget
UPS	Universidad Politécnica Salesiana
UTC	Tiempo Universal Coordinado

SIMBOLOGÍA

°	Grados
°C	Grados centígrados
A	Amperios
Ah	Amperios hora
h	Hora
hPa	Hectopascal
Km	Kilómetro
KW	Kilovatio
m/s	Metros por segundo
mm	Milímetro
MW	Megavatio
s	Segundo
V	Voltios
W	Vatios
W/m ²	Vatios por metro cuadrado
Wh/m ²	Vatios hora por metro cuadrado

ÍNDICE DE FIGURAS

Figura 1.1. Localización de estaciones del tipo climatológica. (Vaca & Ordóñez, 2020)	21
Figura 2.1. Componentes de radiación solar. (Cantos, 2016)	25
Figura 2.2. Representación de funcionamiento de LGBM (Barrios J., 2022)	36
Figura 2.3. Módulos repetitivos en red neuronal LSTM (Olah, 2015)	39
Figura 2.4. Funcionamiento de red LSTM. (Olah, 2015)	40
Figura 2.5. Estructura interna de red ConvLSTM (Shi et al., 2015)	42
Figura 2.6. Estructura codificadora y pronóstico de red ConvLSTM (Shi et al., 2015)	43
Figura 3.1. Flujo de trabajo para predicción de irradiación solar	50
Figura 3.2. Gráfica de caja del RMSE de estaciones meteorológicas agrupadas por propietario. (Ordoñez et al., 2019)	53
Figura 3.3. Comparativa de bases de intervalos de medición entre datos reales y estimados. (Ordoñez et al., 2019)	54
Figura 3.4. Ubicaciones de estaciones meteorológicas del DMQ	56
Figura 3.5. Flujo de proceso para el análisis exploratorio de irradiancia	58
Figura 3.6. Representación de disponibilidad de datos desde 2015 hasta 2022 en las estaciones del DMQ	59
Figura 3.7. Gráfica de distribución de datos en variables ambientales	62
Figura 3.8. Box Plot de radiación global y reflejada	62
Figura 3.9. Isolation Forest para detección de outliers	63
Figura 3.10. Distribución de puntos coordenados para descarga de variables ambientales	64
Figura 3.11. Estructura matriz 4D de archivos netCDF. (GeoSolutions, 2022)	66
Figura 3.12. Centro de análisis de archivo netCDF	67
Figura 3.13. Visualización de irradiancia solar de archivo netCDF	67
Figura 3.14. Representación de irradiancia a partir de archivo netCDF	68
Figura 3.15. Serie de irradiancia solar correspondiente a un punto	69

Figura 3.16. Diagrama de bloques para regresión de irradiancia solar mediante Machine Learning.	70
Figura 3.17. Metodología de ajuste de datos y pronóstico.....	71
Figura 3.18. Segmentación de dataset para entrenamiento de modelos de regresión.	72
Figura 3.19. Reajuste Skforest con origen fijo. (Amat, 2021).....	73
Figura 3.20. Metodología para modelo Random Forest mediante Skforecast	74
Figura 3.21. Descripción de funcionamiento de la herramienta grid_search_forecaster	76
Figura 3.22. Metodología para modelo LGBM	77
Figura 3.23. Metodología para implementación de Optuna	78
Figura 3.24. Componentes de radiación solar de la estación M5021 en 2021.	79
Figura 3.25. Metodología para pronóstico de radiación solar método Holt.	81
Figura 3.26. Metodología para pronóstico de radiación solar método Holt-Winters... ..	82
Figura 3.27. Metodología para entrenamiento de red LSTM	84
Figura 3.28. Preparación de registros de irradiancia solar.....	85
Figura 3.29. Preprocesamiento de secuencia de imágenes.	86
Figura 3.30. Diseño propuesto de pantalla de inicio	88
Figura 3.31. Pantalla de definición de ubicación previa al inicio de cada programa. .	88
Figura 3.32. Pantalla para calcular el ahorro energético.....	89
Figura 3.33. Pantalla para el cálculo de sistemas fotovoltaicos.....	91
Figura 4.1. Revisión de registros nulos y válidos correspondientes a estaciones meteorológicas del DMQ.	94
Figura 4.2. Box Plot para irradiancia solar en estaciones meteorológicas del DMQ..	96
Figura 4.3. Análisis de diagrama de cajas para irradiancia solar por horas.....	97
Figura 4.4. Análisis de irradiación solar por estación.....	99
Figura 4.5. Detección de outliers en datos de radiación de la estación “El Carmen”	100
Figura 4.6. Registro de outliers captados por estación.	101
Figura 4.7. Suavizado outliers en serie de tiempo de irradiancia solar.	102
Figura 4.8. Gráfico de violín para variables ambientales de tipo satelital.	104
Figura 4.9. Histograma para variables ambientales de tipo satelital.	105
Figura 4.10. Análisis de irradiancia GHI satelital por horas.	106

Figura 4.11. Análisis de irradiancia GHI satelital por meses.....	107
Figura 4.12. Ejemplo de matrices de irradiancia horizontal global.....	108
Figura 4.13. Resultados de regresión de irradiación solar en estación Itulcachi.	110
Figura 4.14. Resultados modelo LGBM sin considerar irradiación nocturna.	112
Figura 4.15. Importancia de variables para regresión de irradiancia solar sin considerar ciclos nocturnos.....	113
Figura 4.16. Importancia de variables para regresión de irradiancia solar.....	114
Figura 4.17. Importancia de hiperparámetros para regresión de irradiancia solar. ..	114
Figura 4.18. Ajuste de hiperparámetros por iteración.	115
Figura 4.19. Evaluación de regresión de irradiancia en estaciones meteorológicas según métrica MAE.....	117
Figura 4.20. Evaluación de regresión de irradiancia en estaciones meteorológicas según métrica R2.....	117
Figura 4.21. Resultados de evaluación de modelos de regresión de radiación solar.	119
Figura 4.22. Presión atmosférica por ubicación en la provincia de Pichincha.	120
Figura 4.23. Evaluación de regresión de modelos ML para ubicaciones de prueba.121	
Figura 4.24. Visualización de serie de tiempo de radiación solar.	122
Figura 4.25. Irradiación solar anual en la provincia de Pichincha durante 2021.	122
Figura 4.26. Descomposición estacional de serie de tiempo de irradiancia en el punto (-0.31, -78.26).	124
Figura 4.27. Pronóstico de irradiancia solar mediante método Holt.....	124
Figura 4.28. Pronóstico de irradiancia solar mediante método Holt-Winters.	125
Figura 4.29. Estructura modelo LSTM para irradiancia solar.....	129
Figura 4.30. Comparativa de pronóstico y valores reales de irradiancia solar.....	130
Figura 4.31. Comportamiento de irradiancia solar en secuencia Nro. 220.	131
Figura 4.32. Estructura de modelo ConvLSTM para pronóstico de irradiación solar.	133
Figura 4.33. Gráfica de pérdida de entrenamiento y validación de modelo ConvLSTM.	133
Figura 4.34. Pronóstico de irradiancia solar mediante Conv-LSTM.....	136
Figura 4.35. Componentes de aplicación de pronóstico de irradiación solar.	137

Figura 4.36. Componentes de módulo de “Análisis de irradiación solar”	139
Figura 4.37. Distribución de elementos visuales en herramienta de “Análisis de irradiación solar”	141
Figura 4.38. Componentes del panel de pronóstico de irradiación solar.	142
Figura 4.39. Distribución de elementos visuales en herramienta de “Pronóstico de irradiación”	144
Figura 4.40. Componentes de panel de “Calculadora solar”	145
Figura 4.41. Panel de detalle de consumos AC y DC.	146
Figura 4.42. Componentes de módulo “Calculadora solar”	149

ÍNDICE DE TABLAS

Tabla 1.1. Descripción de variables disponibles por estación meteorológica del INAMHI.	22
Tabla 1.2. Descripción de variables disponibles de la base de datos satelital NREL.	23
Tabla 2.1. Análisis de variables ambientales para el desarrollo de modelos de predicción de irradiación solar.	30
Tabla 3.1. Marca, modelo, clase de sensores tipo ISO9060 disponibles en estaciones meteorológicas de Ecuador. (Ordoñez et al., 2019).....	52
Tabla 3.2. Datos identificativos y ubicación de estaciones meteorológicas ubicadas en las provincias de Pichincha y Napo. (FONAG, 2022)	56
Tabla 3.3. Análisis de variables ambientales disponibles en la estación M5021.	60
Tabla 3.4. Valores perdidos en estación M5021	61
Tabla 3.5. Atributos de fichero netCDF	65
Tabla 3.6. Características técnicas de equipos y servidor	87
Tabla 4.1. Resumen descriptivo de datos disponibles por cada estación meteorológica.	94
Tabla 4.2. Resumen descriptivo de datos satelitales en el punto (-0.19, -78.5).....	103
Tabla 4.3. Resultados de entrenamiento de modelos SkForest y LGBM.....	109
Tabla 4.4. Resultados de entrenamiento LGBM sin irradiancia nocturna.	111
Tabla 4.5. Evaluación de modelos de irradiancia solar por estación.	115
Tabla 4.6. Evaluación de pronóstico mediante método Holt-Winters.....	126
Tabla 4.7. Evaluación de métodos estadísticos para pronóstico de irradiancia.	126
Tabla 4.8. Resultados test Dickey-Fuller para serie de tiempo de irradiancia solar.	127
Tabla 4.9. Configuración de modelo LSTM.....	128
Tabla 4.10. Evaluación de modelo LSTM (one-step) para pronóstico de irradiancia.	130
Tabla 4.11. Configuración de modelo ConvLSTM.	132
Tabla 4.12. Evaluación de pronósticos generado por modelo ConvLSTM.	135
Tabla 4.13. Comparación de pronóstico entre modelos ConvLSTM.....	135
Tabla 4.14. Costos de implementación.....	150

CAPÍTULO 1

1. INTRODUCCIÓN

1.1 Antecedentes

Durante las últimas décadas, el aprovechamiento de la energía solar ha venido evolucionando progresivamente dando como resultado una alternativa energética confiable, limpia y económicamente rentable. El estudio de diversos materiales que pueden captar la luz y transformarla en electricidad, en conjunto con el desarrollo de técnicas de manufactura han facilitado la producción en masa de paneles solares y sistemas auxiliares que permiten integrar fácilmente diferentes tipos de energía para el uso cotidiano ha hecho que la energía solar reciba una amplia acogida a nivel mundial, sin embargo, el comportamiento de la irradiación solar no es constante en el tiempo y dependiendo de la ubicación geográfica la cantidad de luz solar es variable.

Para facilitar el estudio de la factibilidad de proyectos solares, se han desarrollado múltiples herramientas enfocadas en el análisis del componente solar, tal es el caso de Project Sunroof perteneciente a Google y creado en 2015, hace el uso de información satelital y variables ambientales para analizar la cantidad de luz solar que puede captarse en un techo, esta iniciativa permite al usuario dimensionar rápidamente un sistema fotovoltaico y mostrar el ahorro que supondría el cambio a la energía solar hasta 20 años, lastimosamente la herramienta se encuentra disponible solamente para Estados Unidos (Plena energía, 2022).

Otra alternativa desarrollada por la Comisión Europea es la herramienta Photovoltaic Geographical Information System (PVGIS) que permite analizar la irradiación solar en gran parte del planeta utilizando como referencia imágenes satelitales, de esta manera es una herramienta enfocada al estudio de la luz solar para el dimensionamiento de sistemas fotovoltaicos, no obstante, su base de datos no se encuentra actualizada y se encuentra optimizada para el uso en Europa (López de Benito, 2019).

Dentro de Latinoamérica, Chile desarrolló el “Explorador de Energía Solar” en el cual hace uso de información satelital y aplicando diferentes modelos matemáticos ha logrado desarrollar una herramienta capaz de analizar la irradiación solar e integrar estos datos para la planificación de sistemas solares, la aplicación se encuentra segmentada para diferentes tipos de usuarios variando la cantidad de herramientas y complejidad de las mismas para obtener un estudio en detalle de la luz solar y su aplicación en sistemas fotovoltaicos, se encuentra disponible únicamente para el territorio chileno (Molina, 2017).

1.2 Descripción del problema

El sol constituye una fuente de energía prácticamente ilimitada debido a las reacciones de fusión que tienen lugar en su núcleo, liberando grandes cantidades de energía de las cuales, una pequeña fracción es recibida en la atmósfera del planeta y dependiendo de condiciones como: las estaciones, la ubicación geográfica, condiciones ambientales, hora, etc. influyen en la capacidad utilizable en la superficie terrestre (Tous, 2010).

El uso constante de combustibles de origen fósil como el carbón, el petróleo y el gas han aportado al incremento del nivel de contaminación y al ser energías no renovables con yacimientos limitados, sumado al incremento de la demanda energética se requiere optar por nuevas alternativas que sean amigables con el medio ambiente, mantengan un flujo continuo y no representen el riesgo de agotarse a largo plazo tales como: biomasa, energía hidráulica, eólica y solar (Moro, 2010).

El avance tecnológico en las últimas décadas ha aportado al incremento de la eficiencia energética de los paneles solares, sumado a los bajos costos de producción, ha hecho de la energía fotovoltaica una apuesta clave de muchos países para producir electricidad de forma económica, de acuerdo al informe Global Market Outlook for Solar Power 2021 – 2025, el 39% de la energía producida en el 2020 corresponde al componente solar, actualmente, los líderes de generación de energía fotovoltaica son: China (33%), Estados Unidos (12%), Japón (9%) y Alemania (7%) quienes aprovechan su extensión, capacidad de irradiación solar y

fomentan la conciencia ambiental mediante el uso de energías renovables (Gil, 2021).

De acuerdo al Atlas del sector eléctrico ecuatoriano elaborado por la Agencia de Regulación y Control de Energía y Recursos Naturales No Renovables (2020), la capacidad de generación eléctrica del país en el año 2020 fue 8712.29 MW de los cuales, la energía solar aportó con 27.63 MW que representa apenas el 0.32% de la energía total, donde la mayor generación de energía fotovoltaica se encuentra concentrada en las provincias de El Oro, Loja e Imbabura (pp. 19-30).

Pese a registrar niveles elevados de captación de irradiación solar en toda la superficie país, son pocos los proyectos impulsados para generar electricidad a partir de la luz solar, una de las limitantes es la ausencia de registros de irradiación que permitan dimensionar adecuadamente un sistema fotovoltaico e identificar la rentabilidad a corto plazo de la instalación en una ubicación determinada mediante la aplicación de predicciones en base a la información histórica, en la actualidad se disponen de atlas solares desactualizados que proporcionan una visión general del potencial solar sobre el territorio ecuatoriano.

1.3 Justificación del problema

En la actualidad existe una amplia fuente de información satelital que proporciona registros de diferentes variables ambientales, además de la cantidad de irradiación en cualquier ubicación en el mundo que han sido de utilidad para desarrollar exploradores solares que son herramientas dedicadas al análisis histórico de luz solar en zonas determinadas, generalmente abarcan extensiones de países en concreto, de tal forma que permiten establecer un acercamiento tanto técnico utilizando métricas de diseño como de propósito general y libre acceso para cualquier persona, facilitando una visión general del potencial solar y el ahorro que supondría la puesta en marcha de una instalación solar.

La importancia de este tipo de exploradores radica en la necesidad de identificar los periodos con menor aporte solar en un lugar donde se requiera una instalación fotovoltaica y de esta forma, calcular la cantidad de paneles solares necesaria para

asegurar un correcto funcionamiento del proyecto. Teniendo en cuenta la naturaleza irregular de la luz solar que es recibida en la superficie terrestre es importante realizar un seguimiento y generar predicciones que permitan identificar la rentabilidad de la instalación.

En Ecuador no existe ninguna herramienta similar que permita realizar un estudio a profundidad de la cantidad de irradiación solar y generar estimaciones futuras confiables para gestionar eficientemente una instalación solar a pequeña escala o en plantas de mayor producción, es indispensable aprovechar las fuentes de información para detectar zonas de alto potencial para la generación de energía utilizando información satelital, respaldada con las mediciones de la red de estaciones meteorológicas disponibles en diferentes puntos del país de esta manera incentivar al uso de energías limpias y aportar al cambio de la matriz energética.

1.4 Solución propuesta

Para solucionar la problemática planteada se pretende elaborar una herramienta que permita integrar la precisión procedente de fuentes de información locales de estaciones meteorológicas, con la estabilidad de registros de irradiación solar disponibles en fuentes satelitales mediante la aplicación de técnicas de aprendizaje automático, de tal manera que se dispongan de una base de datos actualizados, continuos y precisos dentro del área de interés.

De esta manera, se puede emplear un conjunto de datos estándar para llevar a cabo una predicción de la cantidad de luz solar que se podría aprovechar en un intervalo de tiempo a futuro mediante el empleo de aprendizaje profundo, los resultados de predicción se mostrarán al usuario de forma dinámica permitiendo tomar un punto de partida en el diseño de sistemas fotovoltaicos que puedan satisfacer a la demanda energética conociendo el comportamiento de la irradiación solar.

1.5 Objetivos

1.5.1 Objetivo General

Evaluar el potencial de irradiación solar captado en una superficie terrestre mediante técnicas de aprendizaje automático para la elaboración de un explorador solar interactivo como herramienta preliminar al diseño de sistemas fotovoltaicos.

1.5.2 Objetivos Específicos

1. Extraer información para la construcción de un conjunto de datos necesario para el análisis de la irradiación solar utilizando repositorios de libre acceso.
2. Identificar técnicas de aprendizaje automático afines al pronóstico de irradiación solar para la selección del mejor modelo predictor en función de métricas de evaluación.
3. Desarrollar una interfaz interactiva para la visualización de la predicción de irradiación solar en una ubicación determinada mediante la implementación de un modelo basado en aprendizaje automático.

1.6 Metodología

En el desarrollo del proyecto se analizarán diversas fuentes de información para integrar un conjunto de datos contemplando la capacidad de irradiación solar y variables externas para la extracción de información con la finalidad de generar un modelo de predicción y posterior visualización de resultados de acuerdo a una ubicación geográfica.

La primera etapa del proyecto permite comprender el problema de investigación y la búsqueda de información a ser extraída, posteriormente acondicionada mediante la aplicación de técnicas de limpieza de datos, estadística y visualizaciones para facilitar su análisis y generar un nuevo conjunto de datos teniendo en consideración potenciales variables de interés.

La segunda etapa consiste en el análisis exploratorio de la información para identificar el comportamiento del nivel de irradiancia solar, su posible interacción con variables externas y el enriquecimiento del conjunto de datos incorporando variables adicionales como producto de una revisión literaria que puedan ser

adquiridas y generen valor dentro del estudio propuesto, tomando en cuenta ubicaciones geográficas de acuerdo a la disponibilidad y calidad de la información.

La tercera etapa del proyecto consiste en la creación de un modelo de pronóstico utilizando diversas técnicas de aprendizaje automático disponibles a manera de librerías siguiendo el proceso de configuración, entrenamiento, calibración iterativa, validación, finalmente evaluar los resultados y seleccionar el mejor modelo.

Finalmente, el desarrollo de una herramienta interactiva que permita integrar los resultados de pronóstico en una ubicación determinada mediante visualizaciones que faciliten la identificación de parámetros de irradiación elementales para el diseño de sistemas fotovoltaicos.

1.7 Resultados esperados

- Conjunto de datos generado a través de la compilación de diferentes fuentes de información para el análisis de irradiancia solar.
- Modelo de predicción de irradiación como producto del empleo de un conjunto de algoritmos de aprendizaje automático.
- Interfaz didáctica para la visualización de los resultados del modelo implementado en una ubicación determinada.

1.8 Dataset

La principal fuente de información son los registros satelitales proporcionados por el Laboratorio Nacional de Energía Renovable (NREL) localizado en Estados Unidos y el compendio de datos ambientales del proyecto Copernicus de la Unión Europea que proporciona información mundial de irradiancia solar y sus componentes disponibles en forma de mapas, archivos CSV, NetCDF y GeoJson poseen datos desde 2007 hasta 2021, las bases de datos son de libre acceso.

En Ecuador no existe ninguna herramienta que permita obtener datos actualizados de radiación solar, sin embargo, se dispone información de estaciones meteorológicas pertenecientes al Instituto Nacional de Meteorología e Hidrología (INAMHI), estaciones meteorológicas del Distrito Metropolitano de Quito (DMQ) y

el Instituto de investigación Geológico y Energético (IIGE) que se encuentran instaladas en puntos estratégicos en todo el territorio ecuatoriano, para este caso de estudio se tomará en cuenta las estaciones del tipo Climatológica que se encuentran en funcionamiento en distintos puntos alrededor del país y se ajusten a las condiciones de disponibilidad y calidad de datos para definir el área de estudio, la mayoría de ellas proporciona de forma pública sus registros desde el 2018 a la actualidad, en la Figura 1.1. se muestra sus ubicaciones respectivas.

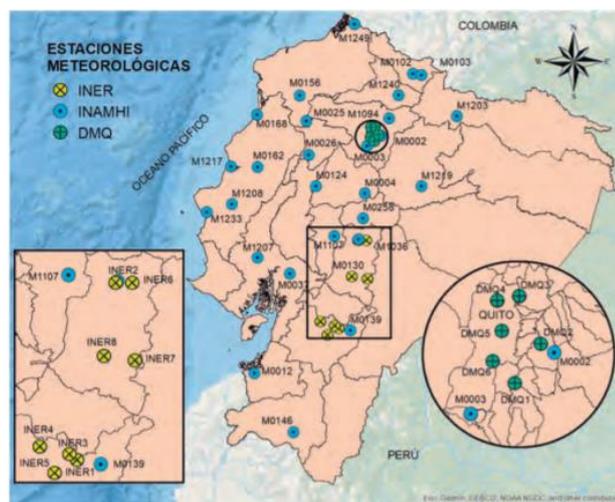


Figura 1.1. Localización de estaciones del tipo climatológica. (Vaca & Ordóñez, 2020)

1.8.1 Datos de estaciones meteorológicas

Para el estudio de variables meteorológicas obtenidas desde estaciones terrestres se utilizará la base de datos del Instituto Nacional de Meteorología e Hidrología (INAMHI), cuyo acceso se realiza a través del portal web del Sistema de Estandarización de Datos Hidroclimáticos Crudos (SEDC) donde se puede acceder a los registros de cada estación en formato CSV, permitiendo organizar el dataset como se muestra en la Tabla 1.1.

Tabla 1.1. Descripción de variables disponibles por estación meteorológica del INAMHI.

Variable	Tipo	Descripción	Cantidad de registros	Resolución
Fecha	Temporal	Registros horarios de variables ambientales medidas en la estación meteorológica.	36990	Horario (UTC -5)
Humedad	Numérica	Humedad relativa promedio medida en el ambiente expresada en porcentaje.	36990	Horario
Presión	Numérica	Presión ambiental promedio registrada en hPa.	36990	Horario
R. Global	Numérica	Cantidad de radiación solar global promedio registrada en W/m ² .	36990	Horario
R. Reflejada	Numérica	Cantidad de radiación solar reflejada promedio registrado en W/m ² .	36990	Horario
Temperatura	Numérica	Temperatura ambiental promedio registrada en °C.	36988	Horario
Dirección V.	Numérica	Dirección del viento instantánea registrada en grados (°).	31832	Horario
Velocidad V.	Numérica	Velocidad del viento instantánea registrada en m/s.	31832	Horario

Los datos satelitales detallados en la Tabla 1.1 se tomarán en cuenta como información referencial, debido a que cada equipo físicamente puede verse afectado por diversas condiciones ambientales, mantenimiento, vida útil de sensores que implican el cambio en la cantidad y calidad de registros.

Otra característica a tener en consideración es el formato de fecha y hora que se encuentran calibradas todas las estaciones, siendo considerado el Tiempo Universal Coordinado (UTC) para Ecuador UTC -5 y de esta forma, la hora se corresponda con la cantidad de luz medida.

1.8.2 Datos satelitales

Para complementar el estudio de variables meteorológicas se utilizará las bases de datos de la red satelital Copernicus que ha sido implementada por la Unión Europea en conjunto con los registros satelitales del laboratorio NREL, teniendo en cuenta la amplia variedad de mediciones, intervalos de tiempo entre medidas y la resolución espacial. Los datos se pueden obtener en formato netCDF (.nc) o en CSV, específicamente para el análisis de irradiancia solar la información se encuentra disponible en el repositorio

NREL desde el 2015 hasta 2022 teniendo en cuenta que el formato de tiempo por defecto es UTC 0, las características del dataset se describen en la Tabla 1.2.

Tabla 1.2. Descripción de variables disponibles de la base de datos satelital NREL.

Variable	Tipo	Descripción	Cantidad de registros	Resolución temporal	Resolución espacial
Fecha	Temporal	Registros horarios de variables ambientales captadas por satélites.	17520	30 minutos (UTC 0)	No aplica
Latitud	Espacial	Componente de ubicación geográfica (grados).	27	No aplica	0.036° x 0.036° (Aprox. 4 x 4 Km)
Longitud	Espacial	Componente de ubicación geográfica (grados).	40	No aplica	0.036° x 0.036° (Aprox. 4 x 4 Km)
DHI	Numérica	Irradiancia directa horizontal (W/m ²)	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
DNI	Numérica	Irradiancia directa normal (W/m ²)	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
GHI	Numérica	Irradiancia horizontal global (W/m ²)	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
Temperatura	Numérica	Temperatura ambiental promedio registrada en °C.	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
Velocidad del viento	Numérica	Velocidad del viento instantánea registrada en m/s.	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
Dirección del viento	Numérica	Dirección del viento instantánea registrada en grados (°).	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
Humedad relativa	Numérica	Humedad relativa promedio medida en porcentaje (%)	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
Presión	Numérica	Presión ambiental promedio registrada en hPa.	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)
Precipitación	Numérica	Altura alcanzada por el nivel de lluvia registrada por pluviómetros (mm)	17520	30 minutos	0.036° x 0.036° (Aprox. 4 x 4 Km)

CAPÍTULO 2

2. MARCO TEÓRICO Y ESTADO DEL ARTE

2.1 Conceptos generales

2.1.1 Radiación solar

La radiación solar es la energía que se genera en el sol como resultado de las reacciones de fusión que se llevan a cabo en la estrella, que se transmiten y alcanzan la atmósfera terrestre en forma de radiación electromagnética, la misma que se ve afectada por diversos factores geográficos, atmosféricos y temporales, que van a incidir en la cantidad de energía disponible en la superficie terrestre para su aprovechamiento usualmente a través de sistemas térmicos y fotovoltaicos (Corcobado & Rubio, 2010).

Durante el diseño de sistemas fotovoltaicos es importante considerar los componentes de la radiación solar y su interacción con diversas condiciones ambientales, la representación gráfica de cada componente se muestra en la Figura 2.1 siendo principalmente:

- **Radiación directa (DNI):** Es la energía que se recibe directamente desde el sol hacia el panel.
- **Radiación difusa (DHI):** Es la radiación dispersa después de atravesar la atmósfera, sin la presencia de nubes la radiación difusa ocupa el 15% de la radiación global, a diferencia de días nublados puede superar el 50% de la radiación total disponible (Cantos, 2016).
- **Radiación reflejada o de albedo:** Es la radiación remanente producida por la reflexión de la luz solar sobre el entorno e incide en un punto de interés.
- **Radiación global (GHI):** Es la radiación total disponible y la sumatoria de la radiación directa, difusa y reflejada que es de interés para el diseño de sistemas solares.

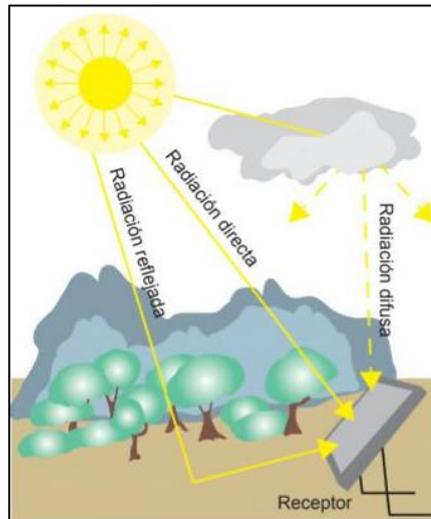


Figura 2.1. Componentes de radiación solar. (Cantos, 2016)

2.1.2 Irradiancia e irradiación

Para poder cuantificar cantidad de energía procedente de la radiación solar disponible en un área determinada existen dos tipos de unidades que se harán uso durante el desarrollo del presente proyecto, siendo importante identificarlas adecuadamente y asociarlas correctamente en función a su origen y aplicación.

- Irradiancia

Se define como la potencia de la luz solar disponible por unidad de área, se mide en W/m^2 , esta unidad es importante debido a que el estándar de irradiancia solar para el uso de paneles solares y evaluar sus parámetros es $1000 W/m^2$, considerando implícitamente los efectos de reducción de la radiación por efectos de la atmósfera. Esta unidad es la que se encuentra disponible los registros de luz solar a partir de estaciones meteorológicas y registros satelitales (Cantos, 2016).

- Irradiación

Se define como la suma consecutiva de irradiancias que se captan de forma regular en un periodo de tiempo dado, de esta manera se puede cuantificar la cantidad de energía solar que incide en una superficie en un tiempo determinado, su unidad de medida se expresa en $W \cdot h/m^2$ y sus múltiplos.

Esta unidad permite designar a la irradiación global en términos de suma total y su respectivo promedio con respecto a un periodo de tiempo que puede ser mensual (m) o anual (a) y es la unidad que se emplea para el dimensionamiento de sistemas fotovoltaicos.

2.1.3 Tratamiento de datos meteorológicos

Como se ha mencionado previamente, en el presente proyecto se dispondrán de diversas fuentes de datos, haciendo énfasis a la información captada desde estaciones en tierra requieren un tratamiento de los datos debido a que los sensores empleados se encuentran a la intemperie y pueden acarrear inconsistencias en sus registros debido a fallos en los instrumentos de medida.

De acuerdo a Cantos (2016), propone una serie de consideraciones a tener en cuenta para el correcto manejo de datos de irradiancia, siendo:

- No se deben considerar las mediciones de radiación durante ciclos nocturnos debido a la ausencia de luz se espera una irradiancia igual a cero, son datos que no aportan valor.
- En el caso de disponer de registros de las componentes de radiación solar, las componentes no pueden ser mayores a la radiación global, esto indicaría una clara descalibración de los sensores empleados.
- Los piranómetros tienden a presentar ruido en sus mediciones en los cambios diurnos y nocturnos, debido a la poca cantidad de luz en el entorno los registros no son fiables.
- Se recomienda validar los datos mediante la comparación de registros con estaciones aledañas siempre y cuando las condiciones ambientales sean similares.
- Realizar un tratamiento estadístico de los datos para simplificar la cantidad de datos medidos, mantener un registro de valores mínimos, máximos, promedios y medianas permiten analizar eficazmente los rangos de irradiación solar y la tendencia que se ha presentado en un intervalo de tiempo definido.

2.2 Fundamentos del problema

En la actualidad, el desarrollo e investigación asociados a fuentes de energía renovables ha tomado relevancia debido a la disminución de las reservas de combustibles de origen fósil y el incremento de la contaminación ambiental debido al uso de los mismos; con la finalidad de aprovechar los recursos renovables para la producción de energía limpia, económicamente rentable y eficiente ha dado paso al avance tecnológico y el desarrollo de métodos para incrementar su viabilidad ante la creciente demanda energética global, siendo los mayores exponentes la energía solar, eólica y biomasa.

La energía solar es un recurso renovable prácticamente ilimitado, cuya energía puede ser aprovechado de diversas maneras siendo la más estudiada la generación eléctrica, sin embargo, este recurso puede verse afectado por las condiciones climáticas, ubicación geográfica, la influencia de las estaciones durante el año que impiden un flujo constante que pueda ser aprovechado.

Debido a su naturaleza estocástica, se ha vuelto una necesidad el poder anticipar la cantidad de luz solar para la producción de electricidad en diversos intervalos de tiempo, a corto plazo permite realizar un seguimiento en tiempo real de la cantidad de energía que puede ser producida con respecto al consumo y poder gestionar eficientemente el suministro eléctrico de una planta solar (Ordoñez Palacios et al., 2020).

Para lidiar con la alta variabilidad del recurso solar es necesario determinar con eficacia un pronóstico con un margen suficiente para la toma de decisiones se han propuesto soluciones basadas en técnicas estadísticas, machine learning y deep learning integrando una amplia gama de fuentes de información que puedan aportar a la predicción de irradiación de tal forma que se pueda reducir el margen de error en el pronóstico en un horizonte temporal dado, aprovechando el potencial de la inteligencia artificial para este tipo de aplicaciones (Sinc, 2020).

Finalmente, para impulsar el uso de energía solar a nivel local y proporcionar un nuevo panorama asociado a los beneficios del aprovechamiento del recurso solar es necesario el desarrollo de una herramienta que pueda integrar la información y utilizarla para generar pronósticos fiables a través del empleo de técnicas de inteligencia artificial, de tal manera que los usuarios puedan conocer la realidad energética a la cual tienen acceso y pueda usarse como punto de referencia para la planificación de sistemas fotovoltaicos.

2.3 Soluciones basadas en datos

El pronóstico de irradiación solar es una problemática de interés general debido a la necesidad de conocer con anticipación la cantidad de luz en un área determinada con la finalidad de administrar estratégicamente la cantidad de energía producida con respecto a la demanda, a su vez, existe una amplia cantidad de técnicas y fuentes de información que se han considerado en diversos estudios con la finalidad de generar un modelo que presente los mejores resultados de predicción.

De acuerdo a Ordoñez et al. (2020) se plantea el pronóstico de irradiación solar a partir de datos registrados localmente, a través de estaciones meteorológicas distribuidas en diferentes ciudades de Colombia y su integración con la base de datos World Weather Online, permitiendo identificar los mejores predictores basados en variables ambientales siendo la velocidad, dirección del viento, temperatura, humedad, cantidad de precipitación y posteriormente, la evaluación del desempeño de diversas técnicas de aprendizaje automático (árboles de regresión, regresión lineal múltiple, regresión de soporte vectorial, Gradient Boosting, etc.) para generar pronósticos.

La importancia de determinar eficientemente la cantidad de luz solar para aprovecharlo para la producción de energía ha permitido desarrollo de una amplia gama de estudios enfocados en la predicción solar, tal es el caso de Pasion et al., (2020) donde realiza una comparativa de los algoritmos Stacked ensemble build y Gradient boosting machine (GBM), asimismo, se determina la importancia de variables ambientales como temperatura ambiental, la humedad relativa y el techo

de nubes como las variables predictoras más importantes en el desarrollo de su investigación, otro factor a tener en cuenta es la calidad de datos de las estaciones en tierra siendo éstas las que delimitarán el área de estudio.

Tomando otra perspectiva para llevar a cabo el pronóstico de irradiación solar Taniguchi et al., (2001) expone el método de pronóstico a partir del análisis de vectores de deriva de cúmulos de nubes obtenidos desde imágenes satelitales con una resolución de 1800 x 1800 píxeles (1.25 x 1.25 Km²), utilizando modelos lineales y características del tipo de nube, permite estimar su posición en intervalos de 1 a 3 horas en el futuro, con esta información calcular eficientemente la cantidad de irradiación solar que recibe la superficie analizada.

El uso de las bases de datos procedentes de estaciones tiende a presentar un problema en común, el almacenamiento de los datos no se mantiene constante durante el tiempo y puede verse afectado por daños en los sensores, ejecución de mantenimientos, cortes de energía u otros factores que impiden un registro constante, no obstante, los datos satelitales son otra fuente de información habitual que no presenta los problemas de continuidad temporal pero sus mediciones no son precisas con respecto a los valores reales medidos en tierra (Ordoñez et al., 2019).

Para consolidar las ventajas existentes entre diversas fuentes de datos Narváez et al. (2021) proponen un método de integración basado en técnicas de aprendizaje automático que permite ajustar la información satelital con respecto a las mediciones en tierra, de esta forma se construye un conjunto de datos artificial aprovechando las fortalezas de cada método de captura de datos.

2.4 Análisis de variables para el desarrollo de modelos de predicción de irradiación

Teniendo en cuenta diversas metodologías propuestas en investigaciones afines a la predicción de irradiación solar, como resultado se han podido determinar variables que juegan un rol indispensable en la mejora del rendimiento de los

modelos de aprendizaje automático y a su vez, tienden a ser datos que se encuentran disponibles en la mayoría de estaciones meteorológicas.

Existe una amplia variedad de estudios enfocados en el pronóstico de irradiación solar donde se han explorado diversos enfoques y variables predictoras, la Tabla 2.1 muestra las variables principales que se han considerado en diversas investigaciones, siendo mayoritariamente: la temperatura ambiental, humedad relativa, la dirección y velocidad del viento e irradiación global horizontal las que se han tomado en múltiples estudios y deben ser considerados prioritariamente para la ejecución del presente proyecto.

Tabla 2.1. Análisis de variables ambientales para el desarrollo de modelos de predicción de irradiación solar.

Variable	(Narvárez & Giraldo, 2020)	(Ordoñez & León, 2020)	(Pasion et al., 2020)	(Kayri et al., 2017)	(Viscondi & Alves-Souza, 2021)	(Urraca et al., 2016)
Temperatura	X	X	X	X	X	X
Irradiancia Global Horizontal (GHI)	X	X		X	X	X
Velocidad del viento		X	X	X	X	X
Dirección del viento	X	X	X	X		
Humedad relativa		X			X	X
Precipitación		X			X	
Presión atmosférica			X		X	
Zenith solar	X			X		
Irradiancia Directa Horizontal (DHI)	X					
Irradiancia Directa Normal (DNI)	X					
Latitud			X			
Longitud			X			
Hora del día			X			

Cabe destacar la presencia de ciertas variables ambientales que se han usado frecuentemente en la mayoría de estudios tomados como referencia, siendo la temperatura ambiental, la irradiancia global horizontal, humedad relativa, dirección y velocidad del viento como variables predictoras efectivas basadas en la revisión

bibliográfica que serán adoptadas como punto de partida durante el desarrollo del presente proyecto.

Teniendo en cuenta la disponibilidad de información a partir del repositorio de registros satelitales, las variables que se emplearán para el pronóstico de irradiación solar son:

- Irradiancia Global Horizontal (GHI)
- Irradiancia directa horizontal (DHI)
- Irradiancia Directa Normal (DNI)
- Temperatura ambiental
- Presión atmosférica
- Humedad relativa
- Dirección del viento
- Dirección del viento
- Nivel de precipitación
- Hora del día

2.5 Fuentes de datos

Una vez identificadas las variables de mayor relevancia para el desarrollo de pronóstico de irradiación solar de acuerdo a la Tabla 2.1, es importante identificar las fuentes de información adecuadas para la construcción de un dataset que contemple diversas variables ambientales en un mismo intervalo de tiempo, siendo las principales los datos disponibles de forma local a través de estaciones de tipo meteorológico y datos satelitales de acceso libre.

De acuerdo a Narvaez et al., (2021) en su investigación realiza el análisis de irradiación solar en el departamento de Nariño, Colombia mediante el uso de registros ambientales medidos en tierra a partir de la red de estaciones meteorológicas comprendidas por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) y para los datos satelitales se utilizó el repositorio

público National Solar Radiation Database (NSRDB) de acuerdo a Sengupta et al., (2018) este repositorio se caracteriza por mantener una amplia gama de mediciones ambientales, con una resolución aproximada de 4 km² y con registros con intervalos de 30 minutos desde 1998 hasta la actualidad.

Según Kayri et al., (2017) propone la extracción de datos a partir de la generación eléctrica de plantas fotovoltaicas y considerando a la irradiación dentro del conjunto de variables predictoras, siendo de vital importancia pronosticar la producción energética para satisfacer la demanda. Otro caso particular tratado por Pasion et al., (2020) establece su fuente de datos a partir de un sistema previamente desarrollado para la medición de variables ambientales empleando sistemas embebidos facilitando el control de la calidad y resolución de las mediciones de acuerdo a la necesidad del estudio.

Finalmente, Urraca et al., (2016) y Ordoñez Palacios et al., (2020) desarrollan su investigación utilizando datos de estaciones meteorológicas locales debido a su disponibilidad y libre acceso, amplio historial de mediciones y factibilidad para obtener información actualizada, la baja tolerancia de los sensores ambientales lo que permite garantizar la fiabilidad de los datos almacenados siendo la principal alternativa a ser considerada para construir una base de datos.

En Ecuador como se expone en el apartado *1.8.1. Datos de estaciones meteorológicas*, se dispone de 53 estaciones pertenecientes al INAMHI, DMQ e INER que se encuentran distribuidas en diferentes puntos del territorio y se encuentran en capacidad de obtener registros de irradiación y otras variables ambientales (radiación solar global, radiación solar reflejada, presión atmosférica, temperatura ambiental, humedad relativa, etc.).

El intervalo de tiempo de registros que se puede acceder es 30 y 60 minutos, siendo una fuente importante de información a tenerse en cuenta para la construcción de la base de datos, sin embargo, es importante tener en consideración factores que pueden afectar a la calidad de los datos tales como: rutinas de mantenimiento que impidan la continuidad de los registros, calibración de sensores, cortes de energía,

fallos en los dispositivos de medición y condiciones ambientales que alteren el funcionamiento adecuado de la estación.

Teniendo en cuenta estos aspectos, es importante realizar una validación de la calidad y veracidad de los datos que se disponen en las estaciones permitiendo a primera instancia, delimitar el área de estudio, la resolución temporal y establecer los periodos para el ajuste con respecto a los datos satelitales.

Finalmente, para la obtención de datos satelitales las principales fuentes de información se han considerado previamente son los repositorios libres de NREL y Copernicus debido a la gran cantidad de variables ambientales disponibles para su análisis, amplio historial de registros, resolución temporal compatible con las estaciones meteorológicas, disponibilidad de registros que abarcan toda la superficie del país y facilidad de descarga de datos.

Para el desarrollo del presente proyecto se tendrá en consideración la menor resolución espacial disponible para proporcionar una perspectiva en detalle de la realidad con respecto a la cantidad de irradiación solar disponible en el país.

2.6 Pronóstico de irradiación solar a partir de datos satelitales

La necesidad de conocer de forma precisa la cantidad de irradiación solar en una ubicación determinada, ha tratado de ser cubierta a través del uso de datos locales siendo una alternativa, la aproximación con respecto a datos de estaciones meteorológicas cercanas, sin embargo, esta práctica es válida siempre y cuando el nivel del terreno sea llano en un radio inferior a 10 km (Zarzalejo et al., 2006).

Otro método que presenta mejores resultados es la interpolación de medidas de irradiación entre estaciones cercanas, que puede ser implementada si el área de interés dispone de una red de estaciones concentradas en una misma zona con una densidad comprendida entre 20 a 50 km de separación entre sí, la estimación a partir de diferentes puntos de referencia y considerando las condiciones del terreno permiten aproximar satisfactoriamente la cantidad de irradiación captada en un punto específico (Zarzalejo et al., 2006).

El uso de datos registrados localmente a partir de sensores especializados presenta una confiabilidad elevada si el dispositivo de medición se encuentre calibrado y se realice frecuentemente tareas de mantenimiento preventivo del equipo e implica grandes costos operativos, caso contrario, los registros pueden contener valores fuera de rango, desfasados en el tiempo, encontrarse incompletos, etc. imposibilitando el uso de esta información.

Para subsanar la escasez de información, el uso de imágenes satelitales para estimar la radiación solar se ha convertido en una herramienta efectiva debido a su capacidad de abarcar amplias extensiones de terreno y la disponibilidad de datos en diversos periodos de tiempo de datos pasados hasta la actualidad, a partir de esta iniciativa se han derivado diversas formas para estimar la irradiación solar, siendo las más representativas los métodos estadísticos y físicos (Zarco et al., 1996).

Dentro de los modelos estadísticos también denominados modelos empíricos, su objetivo es construir funciones de regresión que permitan relacionar la cantidad de irradiación solar medida de forma local con los canales visuales tomadas por el sensor del satélite en el mismo punto e instante en el tiempo, este tipo de metodología ha sido implementado en los datasets de HelioClim que abarcan territorios como Europa, África y el Mediterráneo (Sengupta et al., 2018).

En los modelos físicos se pueden categorizar de acuerdo al método empleado para calcular la radiación solar:

- Modelo Single-step, calcula la irradiancia horizontal global (GHI) a partir de observaciones satelitales en conjunto con teorías de transferencia de radiación.
- Modelo Two-step, abarcan todos los efectos físicos que inciden en la transmisión de radiación solar desde la atmósfera hasta al suelo considerando múltiples variables atmosféricas como el efecto del aerosol y la nubosidad en el área de interés y al igual que los modelos Single-step se

integra las teorías de transferencia de radiación para obtener mediciones de alta confiabilidad, la aplicación más representativa es el dataset Surface Radiation Budget (SRB) desarrollado por la NASA (Sengupta et al., 2018).

Otra variante a considerar, son los modelos semi-empíricos que emplean el cálculo de la radiación reflejada por las nubes y que retornan al satélite, esta radiación remanente es empleada como un “índice de claridad” que indica la proporción de la radiación total que llega a la superficie de interés, este método es ampliamente utilizado en datos que actualmente se utilizan de forma comercial por SolarGIS y Solar-AnyWhere (Cebecauer et al., 2010).

2.7 Light Gradient Boosting Machine (LGBM)

A primera instancia los árboles de decisión de aumento de gradiente (GBDT) se han caracterizado por ser algoritmos de Machine Learning que han tomado relevancia en un sinnúmero de aplicaciones en el campo de clasificación y regresión, sin embargo, el incremento progresivo de la cantidad de información disponible para ser analizada ha generado la necesidad de incorporar métodos que sean precisos y eficientes que puedan procesar una gran cantidad de instancias optimizando el costo computacional.

Para dar solución a esta necesidad, se plantean técnicas como Gradient-based One-Side Sampling (GOSS) que se encarga de analizar el gradiente correspondiente a cada instancia, considerando que gradientes elevadas favorecen al incremento de la ganancia de información permitiendo eliminar variables menos representativas sin alterar el rendimiento del modelo (Ke et al., 2017).

Otra técnica considerada es Exclusive Feature Bundling (EFB) cuya finalidad es identificar eficientemente variables exclusivas y posteriormente, agruparlas a través de un algoritmo que reduce el problema del empaquetamiento a un problema de grafos utilizando las variables como vértices y nodos, cuya resolución se realiza mediante relaciones de aproximaciones constantes (Ke et al., 2017).

La integración de las técnicas GOSS y EFB en los algoritmos basados en árboles de decisión dan como origen a Light GBM, que se caracteriza por su alto rendimiento debido a su particularidad de división de hojas del árbol de decisión a partir de los mejores ajustes, a diferencia de otros algoritmos que tienden a enfocarse en los niveles de profundidad del árbol ocasionando que el modelo se vuelva complejo, el tiempo de entrenamiento sea considerablemente más alto y la precisión del modelo sea menor (Barrios, 2022).

En la página oficial de Light GBM se presenta la representación gráfica del funcionamiento del algoritmo como se muestra en la Figura 2.2, donde se encuentra representado el método de división de hojas permitiendo reducir considerablemente las pérdidas por nivel y, en consecuencia, incrementar la precisión del modelo de salida.

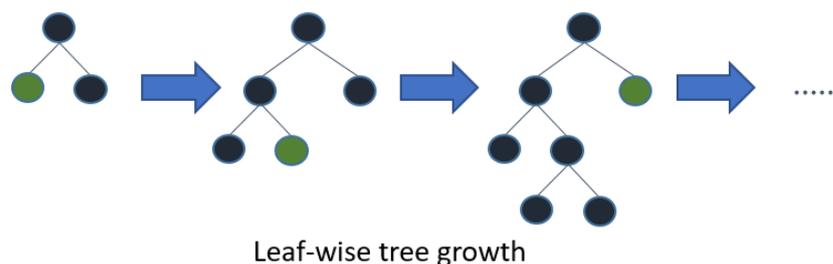


Figura 2.2. Representación de funcionamiento de LGBM (Barrios J., 2022)

Como se había mencionado previamente, el incremento constante de la cantidad de datos imposibilita que métodos convencionales de ciencia de datos puedan satisfacer los requerimientos del mercado actual, siendo indispensable optimizar los tiempos de entrenamiento. Es por ello que LGBM ha hecho énfasis en su velocidad de entrenamiento para manejar cantidades masivas de datos y la optimización en el uso de recursos computacionales, de ahí se ha adoptado el prefijo 'Light'.

Dentro de las ventajas más importantes de LGBM se destacan las siguientes:

- Incorporación de algoritmos basados en histogramas que permiten agrupar variables para acelerar el entrenamiento de modelos y mejorar su eficiencia.

- Se reemplazan valores continuos por contenedores discretos, permitiendo optimizar el uso de la memoria.
- Se mejora la precisión del modelo al trabajar sobre árboles de decisión con menor profundidad, pero con una mayor distribución de sus hojas, para conjuntos de datos pequeños existe el riesgo de sobreajuste.
- Tiempo de entrenamiento bajo para conjuntos de datos de gran tamaño.
- Admite el aprendizaje en paralelo.

2.8 Línea base para predicción de series temporales – Holt Winters

Para el desarrollo del proyecto se utilizará Holt para consolidar una línea base al tratarse de una solución basada en estadística simple y su bajo coste computacional, de tal forma que permita establecer un punto de referencia con respecto a otros modelos de pronóstico más complicados (Rivera, 2020).

En esencia, los métodos de suavizado exponencial se emplean cuando se disponen de datos que no presentan una tendencia constante, además para realizar pronósticos el método asigna una mayor ponderación a observaciones recientes mientras que datos antiguos son suavizados exponencialmente, siendo el pronóstico a futuro el resultado de la suma ponderada del registro más reciente y los valores pasados (Hanke & Wichern, 2006).

En el caso del estudio del comportamiento de la luz solar no se presentan tendencias lineales generales que puedan ser aprovechadas para realizar pronósticos, por lo tanto, una alternativa es estudiar tendencias lineales locales cuyo nivel y pendiente van evolucionando a la par con la serie de tiempo y a partir de dichas tendencias generar nuevas predicciones.

El método de suavización exponencial doble se caracteriza por presentar una alta flexibilidad en los coeficientes que permiten controlar el nivel y la pendiente, las ecuaciones a ser empleadas son:

1) Serie suavizada exponencialmente correspondiente al nivel actual.

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1}) \quad (2.1)$$

2) Estimado de la tendencia.

$$T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1} \quad (2.2)$$

3) Pronóstico para p periodos en el futuro.

$$\hat{Y}_{t+p} = L_t + pT_t \quad (2.3)$$

Donde:

- L_t : Nuevo valor de suavizado.
- α : Constante de suavizado de nivel, puede tomar valores comprendidos entre 0 a 1.
- Y_t : Valor real de la serie en el periodo t .
- β : Constante de suavizado para el estimado de tendencia, puede tomar valores comprendidos entre 0 a 1.
- T_t : Estimado de la tendencia.
- p : Periodos a pronosticar en el futuro.
- \hat{Y}_{t+p} : Pronóstico de la serie para un periodo p en el futuro.

2.9 Redes LSTM

Inicialmente se ha empleado a las redes neuronales recurrentes (RNN) como la principal alternativa para manejar secuencias de datos, cuya respuesta depende de sus entradas y su retroalimentación con el estado actual del sistema que permite mantener la información a modo de memoria, sin embargo, ante la necesidad de analizar dependencias a largo plazo la efectividad de este tipo de red disminuye considerablemente debido a su incapacidad para conectar la información capturada a largo plazo (Olah, 2015).

Para solventar este problema Hochreiter & Schmidhuber (1997) proponen la red neuronal "Long short-term memory" (LSTM) como una red recurrente especializada en "recordar" información durante largos periodos de tiempo para generar un

pronóstico, al incorporar un sistema de puertas que regulan las funciones para agregar o eliminar información a través de la aplicación de funciones de activación que se encuentran integrados en cuatro capas de red neuronal que interactúan en conjunto para completar este proceso, el módulo de la red se representa como se muestra en la Figura 2.3.

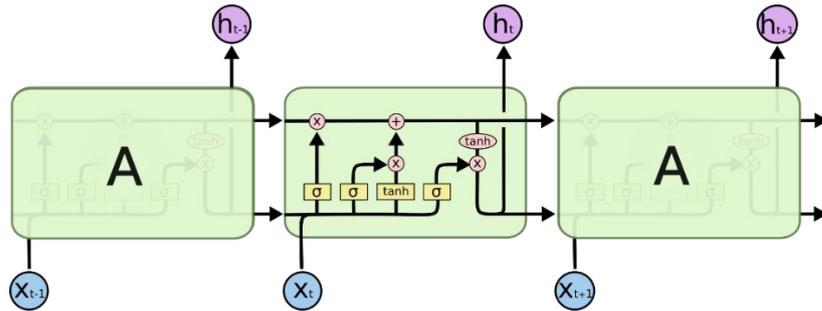


Figura 2.3. Módulos repetitivos en red neuronal LSTM (Olah, 2015).

De acuerdo al esquema propuesto se puede estudiar el comportamiento de la red por etapas:

- 1) La primera capa determina la información que será eliminada en función de las entradas h_{t-1} y x_t que pasan a través de la capa sigmoide que decide si mantener o desechar los datos, y posteriormente inyectada al estado previo de la celda.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2.4)$$

- 2) La siguiente etapa se encarga de decidir la nueva información que será almacenada en el estado actual de la celda, consiste en dos etapas a ser combinadas para actualizar el estado.

- a. La información nueva x_t accede a través de la “capa de puerta de entrada” compuesta por una capa sigmoide que selecciona los potenciales datos a actualizar.
- b. En conjunto se dispone de una capa de tipo tangencial hiperbólica que determinar valores que pueden agregarse al estado de la celda.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2.5)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

- 3) Una vez que se tiene la información a eliminar y su respectiva actualización, debe ser incorporada a través de una serie de operaciones vectoriales para obtener el nuevo estado de la celda.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (2.6)$$

- 4) Finalmente, la salida genera un nuevo estado oculto que se determina a partir del escalamiento del estado actualizado mediante una capa tangente hiperbólica y para seleccionar las partes del estado a generar se aplica una capa sigmooidal.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (2.7)$$

$$h_t = o_t * \tanh(C_t)$$

La representación gráfica de la red LSTM y la nomenclatura empleada en las ecuaciones (2.4 a 2.7) se muestra en la Figura 2.4.

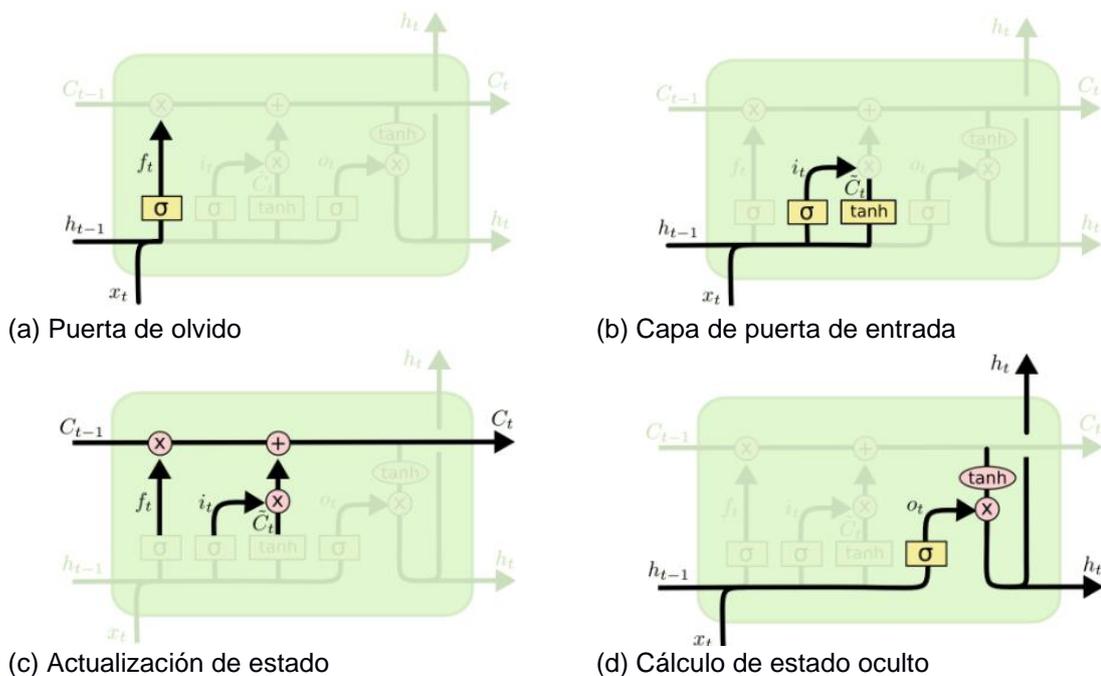


Figura 2.4. Funcionamiento de red LSTM. (Olah, 2015).

2.10 Redes LSTM convolucionales

Las redes de tipo LSTM al mantener una memoria a largo plazo es una herramienta potente para el análisis de series temporales en una dimensión, no obstante, para analizar cambios secuenciales que integran la componente espacial las alternativas se reducen, una alternativa viable definida por Liu et al. (2020) comprende la extracción de características a partir de imágenes ordenadas secuencialmente mediante la aplicación de redes convolucionales apiladas y generar una predicción un paso en el futuro, sin embargo, en múltiples aplicaciones es importante tener múltiples puntos de referencia para la toma de decisiones.

Siguiendo la línea de investigación Graves (2014) propone una configuración basada en redes neuronales Fully connected en combinación con una red LSTM para poder realizar pronósticos en secuencias en dos dimensiones, este método a pesar de abarcar eficientemente la dimensión temporal no puede codificar información espacial y su interacción con puntos cercanos.

Tomando como premisa la posibilidad de combinar dos tipos de redes neuronales Shi et al. (2015) propone la composición de redes convolucionales para aprovechar su capacidad de extracción de características especialmente cuando los datos de entrada son conjuntos de imágenes/videos y las redes LSTM para analizar la secuencia en el tiempo, de esta manera se pretende pronosticar uno o varios fotogramas en el futuro.

De acuerdo al diseño propuesto, se requiere emplear un tensor de 3 dimensiones conformado principalmente por las entradas (x_1, \dots, x_t) , salidas de celda (C_1, \dots, C_t) , capas ocultas (h_1, \dots, h_t) y puertas i_t, f_t, o_t pertenecientes a la red LSTM estándar como una dimensión, mientras que las dos dimensiones restantes corresponden a la componente espacial, distribuidos a manera de una cuadrícula (Shi et al., 2015).

La red denominada como “ConvLSTM” permite realizar un pronóstico de un estado futuro correspondiente a una de las celdas contenidas en la cuadrícula tomando en cuenta la entrada actual y estados pasados de los puntos cercanos a la celda de

referencia, esto se puede efectuar mediante un operador convolucional en las transiciones *state-to-state* e *input-to-state*, la representación gráfica propuesta se muestra en la Figura 2.5.

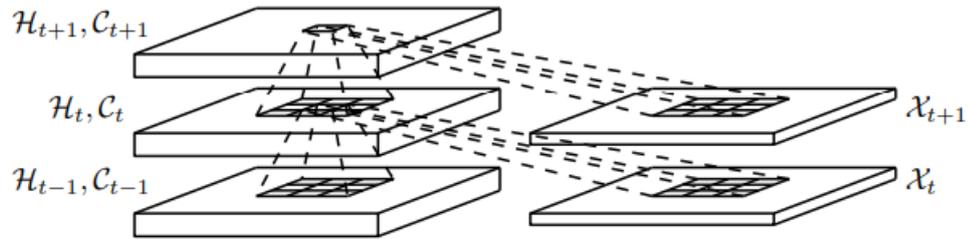


Figura 2.5. Estructura interna de red ConvLSTM (Shi et al., 2015).

Las ecuaciones de la red ConvLSTM son:

$$\begin{aligned}
 i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\
 f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\
 C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\
 o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ C_{t-1} + b_o) \\
 h_f &= o_t \circ \tanh(C_t)
 \end{aligned} \tag{2.8}$$

De acuerdo a las pruebas realizadas, se recomienda el empleo de un Kernel transicional grande cuando se requieren detectar cambios que se desarrollan rápido, y los Kernel pequeños son apropiados para detectar cambios lentos dentro de las escenas.

La estructura de la red ConvLSTM se caracteriza por incorporar la combinación de una red codificadora y otra especializada en pronóstico, donde los estados iniciales y salidas de celda de la red de pronóstico es copiada del último estado de la red de codificación, la representación gráfica se muestra en la Figura 2.6.

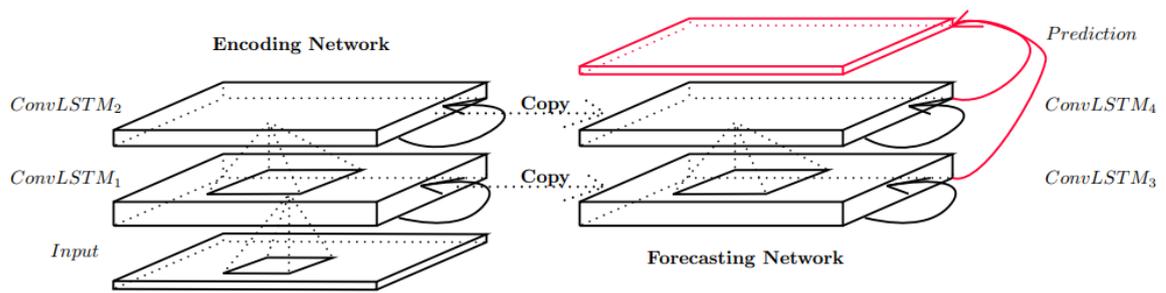


Figura 2.6. Estructura codificadora y pronóstico de red ConvLSTM (Shi et al., 2015).

Como se puede notar en la gráfica, la estructura para codificación y pronóstico se encuentran construidas al apilar varias capas de la red ConvLSTM, asimismo, hay que tener en consideración que el pronóstico generado debe tener las mismas dimensiones que los elementos de entrada, para ello los estados se concatenan en la red de pronóstico y finalmente se aplica una capa convolucional de 1x1 para obtener el resultado (Shi et al., 2015).

2.11 Optuna

Optuna es un marco de software diseñado específicamente para automatizar el proceso de búsqueda de hiperparámetros basándose en una amplia variedad de métodos que incorpora el paquete informático y a su vez, puede evaluar constantemente el rendimiento durante el entrenamiento del modelo y cancelar la iteración si el desempeño del mismo no es favorable, facilitando la gestión de los recursos computacionales y disminuyendo el tiempo de espera durante el entrenamiento (Karani, 2022).

A día de hoy, con la incorporación de modelos de Deep Learning cada vez más complejos que poseen una amplia variedad de parámetros configurables, resulta una tarea complicada poder abarcar con cada uno de ellos teniendo en cuenta que métodos convencionales como Grid Search y las muestras aleatorias muchas veces no logran generar resultados satisfactorios y requieren un uso constante de recursos de hardware e inversión de tiempo para completarse.

Los métodos de muestreo disponibles en Optuna son:

- Grid search: Se define un subconjunto de búsqueda en cuadrícula donde se analizará cada combinación y se seleccionará el conjunto que presentó mejor rendimiento.
- Muestras aleatorias: Se define un espacio de búsqueda donde se podrá seleccionar aleatoriamente el valor del parámetro y abarcar una mayor cantidad de candidatos que no se encuentran restringidos por una cuadrícula.
- Método bayesiano: La búsqueda de hiperparámetros se realiza mediante la construcción de un modelo probabilístico a través de los resultados de una métrica preestablecida, hasta formar una distribución que permita seleccionar la mejor configuración para la siguiente iteración con la intención de mejorar en rendimiento del siguiente modelo (Lim, 2022).
- Algoritmos evolutivos: Se establece un espacio de búsqueda que será evaluado en función de una muestra inicial donde cada hiperparámetro será clasificado de acuerdo a su aptitud y los de rendimiento bajo serán descartados, la búsqueda se repite combinando los conjuntos restantes hasta que el rendimiento no presente ninguna mejora (Lim, 2022).

Teniendo en cuenta estas consideraciones, de acuerdo a Akiba et al. (2019) han planteado características de Optuna que lo hacen una herramienta versátil para la optimización de hiperparámetros:

- El usuario puede construir de forma dinámica el espacio de búsqueda posibles en cada hiperparámetro, denominada como programación “definida por ejecución”.
- Algoritmos enfocados para realizar exploración de hiperparámetros y detección de modelos de bajo rendimiento totalmente personalizables.
- Adaptable, fácil de configurar e implementar independientemente de la complejidad de los cálculos requeridos.
- Incorpora representaciones gráficas que permiten observar la interacción entre hiperparámetros durante cada iteración y la convergencia hacia sus valores óptimos.

2.12 Métricas para evaluación de modelos

Para poder cuantificar el rendimiento de ajuste de los datos pronosticados mediante modelos de Machine Learning y Deep Learning con respecto a los datos reales, existe una amplia variedad de métricas que se pueden emplear para validar los resultados y dependiendo del tipo de aplicación en concreto se pueden seleccionar las mejores métricas a ser empleadas.

El presente proyecto contiene dos etapas, siendo la etapa de regresión de irradiancia solar a partir de datos satelitales mediante el empleo de modelos de Machine Learning, la métrica que mejor se adapta para este tipo de aplicación es el Error cuadrático medio (MSE) que se define como la diferencia al cuadrado de los valores reales y predichos al cuadrado, matemáticamente se expresa como:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2.9)$$

Donde:

- N es la cantidad de puntos.
- y es el valor real.
- \hat{y} es el valor predicho (estimación).

Una de las ventajas del MSE es su sensibilidad ante errores grandes ocasionados por valores atípicos y su disponibilidad a ser diferenciable para calcular valores mínimos y máximos, el objetivo de esta métrica es buscar un modelo que presente la menor puntuación que idealmente indica mejor precisión en las predicciones (Heras, 2018).

Otra métrica relevante para regresión es el Error absoluto medio (MAE) que es menos sensible a valores atípicos, se define como el promedio de la diferencia absoluta entre los valores reales con respecto a los valores predichos, su expresión matemática se define como:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (2.10)$$

Posteriormente, en el proyecto se propone una etapa de pronóstico de irradiancia a partir del análisis de series de tiempo mediante modelos de Deep Learning, para evaluar el ajuste de los modelos durante su etapa de validación se empleará la métrica MSE a ser minimizada en cada iteración, al tratarse de una serie de valores numéricos que requieren ser comparados entre pronósticos y valores reales la métrica seleccionada se ajusta a los requerimientos de la aplicación de pronóstico.

De las métricas principales seleccionadas previamente, pueden ser empleadas eficientemente para el entrenamiento de los modelos de inteligencia artificial, sin embargo, son percepciones subjetivas a nivel operativo que requieren ser expresadas de forma general para poder interpretar si los pronósticos generados son similares a los datos reales, para ello el coeficiente de determinación R^2 es otra métrica de evaluación cuya característica principal son sus puntuaciones libres de escala que se encuentra contenido entre $-\infty$ y 1 (López, 2017).

El coeficiente de determinación permite cuantificar la bondad de ajuste de un modelo, con respecto a una línea base que matemáticamente se expresa como:

$$R^2 = 1 - \frac{MSE (modelo)}{MSE (línea base)} \quad (2.11)$$

El error cuadrático medio de la línea base se calcula como:

$$MSE (línea base) = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 \quad (2.12)$$

Donde:

- N es la cantidad de puntos.
- y es el valor observado.
- \bar{y} es el promedio del conjunto de datos.

Durante el desarrollo del presente proyecto es necesario evaluar el pronóstico de los diferentes métodos y modelos teniendo en consideración las métricas propuestas y aplicadas reiteradas veces, para agilizar el proceso se propone el

desarrollo de una función que pueda ser reciclada para evaluar los resultados de la predicción utilizando principalmente la librería *sklearn.metrics*.

2.13 Librerías y software a utilizar

Para el desarrollo del presente proyecto se utilizará Python 3 debido a la amplia disponibilidad de herramientas para desarrollar las tareas de lectura de ficheros, limpieza, preprocesamiento de datos, representación de datos espaciales y librerías especializadas en Machine Learning y Deep Learning enfocados al pronóstico en series de tiempo.

Como se ha expuesto en el apartado 1.8. Dataset, se plantea la descarga de datos procedentes de estaciones en tierra a partir del repositorio del Sistema de Estandarización de Datos Hidroclimáticos Crudos cuyos ficheros se encuentran disponibles en formato CSV y Excel (xlsx), para la lectura de los archivos se empleará la librería Pandas que incluye la función de lectura *read_csv* y *read_excel*, respectivamente.

Para la descarga de datos satelitales se utilizará el repositorio NSRDB, al requerirse información de una gran cantidad de puntos geográficos se empleará la librería *nrel_dev_api* que permite acceder fácilmente a la API del repositorio para descarga de múltiples variables ambientales correspondientes a una ubicación geográfica determinada y posteriormente, estos registros serán almacenados en un archivo NetCDF especializado para almacenar datos ambientales y geográficos mediante la librería *netCDF4*.

La lectura de datos espaciales se hará uso de la librería *xarray* que permite el procesamiento de archivos NetCDF (.nc) que contienen el registro de variables ambientales en diferentes ubicaciones geográficas considerando la variable temporal, asimismo se utilizará la librería *rioxarray* que permite agregar un mayor manejo de datos espaciales para su proyección e interacción con otros formatos comunes (geotiff), finalmente, la herramienta *cartopy* que facilitará la integración de datos con su respectiva ubicación mediante la creación de mapas facilitando la interpretación de la información satelital.

Se ha establecido el requerimiento de ajuste de datos satélites con respecto a datos locales mediante el empleo de modelos de regresión basados en Machine Learning, para ello se utilizará la librería *skforecast* que permite habilitar las herramientas de la librería Scikit-Learn para preparación de datos, entrenamiento y validación de modelos enfocados específicamente en series de tiempo, el modelo referencial que se empleará es *RandomForestRegressor*.

A la par, se realizará la comparativa del ajuste de datos a través del modelo de regresión de Light Gradient Boosting Machine desarrollado por Microsoft y que ha sido liberado al público, debido a su gran aplicabilidad para trabajar con grandes cantidades de datos en series de tiempo y su facilidad de implementación, se hará uso de la librería *lightgbm*.

La tarea de ajuste de hiperparámetros es una tarea de vital importancia para mejorar el rendimiento del modelo e implica la inversión de tiempo y recursos computacionales para probar múltiples configuraciones de cada modelo, para automatizar el proceso de búsqueda se empleará la herramienta *grid_search_forecaster* disponible en la librería Skforest y la librería *Optuna* que permite definir rangos de parámetros que permitirán evaluar cada modelo y retornar la configuración con el mejor rendimiento.

Para integrar el componente de Deep Learning, se tiene como referencia la librería “TensorFlow” específicamente el módulo “Keras” al ser una herramienta desarrollada específicamente para facilitar el pronóstico de series de tiempo utilizando diversas arquitecturas de redes neuronales recurrentes, otra referencia a ser considerada son las redes LSTM que se especializan en el estudio de series temporales al mantener características durante largos periodos de tiempo que pueden ser empleados en sus pronósticos.

Posteriormente, todos los resultados obtenidos a partir de la implementación de métodos y modelos de pronóstico deben ser presentados al usuario a través de una

interfaz que permita interactuar con los resultados obtenidos, visualizar de forma sintetizada aspectos relevantes del conjunto de datos para ello se empleará la metodología de construcción de una aplicación web bajo el marco de código abierto *Dash* que permite crear visualizaciones e integrarlas fácilmente en un entorno web.

Esta librería disponible en Python se encuentra conformada por diferentes módulos como *Flask* que permite inicializar un servidor web de forma local, *React.js* permite elaborar la interfaz mostrada al usuario a través de una página web y finalmente *Plotly.js* que contiene una biblioteca con una amplia variedad de gráficas disponibles para representar los resultados obtenidos y permitiendo al usuario interactuar con los gráficos.

Finalmente, la librería *Bootstrap* la cual es un framework CSS empleado para estilizar los componentes de una página de HTML y se dispone de elementos prediseñados que se encuentran optimizados para ser visualizados desde ordenador o un dispositivo móvil, esta librería puede ser integrada en conjunto con *Dash* siendo una alternativa viable para el desarrollo del prototipo de visualización de pronóstico de irradiación solar.

CAPÍTULO 3

3. DISEÑO E IMPLEMENTACIÓN

Para el desarrollo del presente proyecto una vez definidos la fuente de datos y modelos que se utilizarán para realizar el pronóstico de irradiación solar, en la Figura 3.1 presenta el flujo de trabajo que se seguirá para generar predicciones partiendo de datos ambientales captados desde estaciones en tierra y registros satelitales.

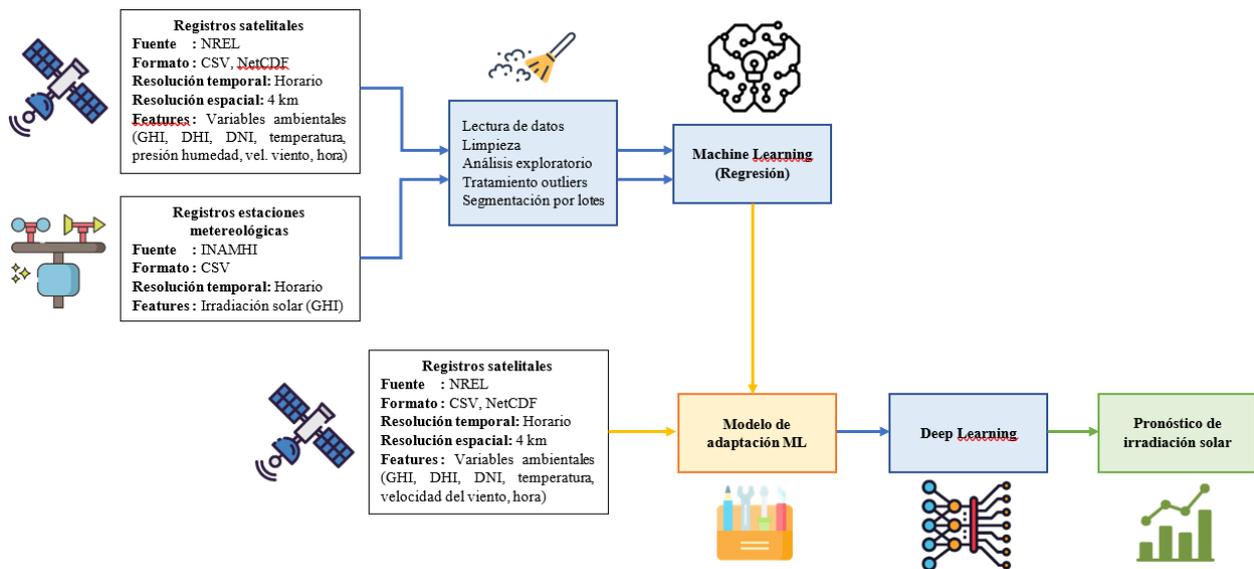


Figura 3.1. Flujo de trabajo para predicción de irradiación solar.

En el esquema propuesto, se establece la fuente de datos y variables que se utilizarán para cada tipo de registro, es decir, para los datos satelitales serán obtenidos desde el repositorio NREL donde se extraerán múltiples variables ambientales (GHI, DHI, DNI, temperatura ambiental, presión, humedad relativa, velocidad del viento) en diversas ubicaciones siguiendo una resolución espacial de aproximadamente 4 km en intervalos horarios, de esta forma los datos pueden ser archivados en ficheros individuales por cada ubicación en formato CSV o a su vez, integrar los datos ambientales de acuerdo a su ubicación y temporalidad en un único fichero NetCDF.

Otra fuente de datos será las mediciones horarias de irradiancia solar mediante estaciones terrestres pertenecientes al INAMHI, que se encuentran disponibles en

formato CSV por cada estación asociada a una ubicación en diversos puntos del país.

Teniendo en cuenta la naturaleza de los datos ambientales es necesario realizar un preprocesamiento y limpieza de los mismos para asegurar su calidad, especialmente en los registros procedentes de estaciones meteorológicas, debido a que cada variable es registrada por diferentes sensores que integran a la estación y a su vez, las condiciones externas, los intervalos de mantenimiento, vida útil de los sensores, calidad del sistema de alimentación y otros factores pueden afectar a los datos que se encuentran disponibles.

De acuerdo a Narvaez et al., (2021) propone el uso de datos satelitales y de estaciones terrestres para crear un conjunto de datos artificial que permita representar la cantidad de irradiación solar en un punto geográfico aprovechando la continuidad temporal durante el registro de datos satelitales y la precisión de los datos procedentes de las estaciones meteorológicas.

Mediante técnicas de Machine Learning se realizará una regresión de irradiancia solar tomando como punto de partida las variables ambientales disponibles en el conjunto de datos satelitales con respecto a la radiación solar medida en tierra, de esta forma obtener un modelo de adaptación que pueda ser aplicado en su totalidad sobre los datos satelitales pudiendo obtener un nuevo conjunto de información mejorado que pueda ajustarse a los requerimientos del presente proyecto.

Finalmente, con los datos de irradiancia ajustados a los datos reales se hará uso de modelos de Deep Learning para realizar pronóstico de irradiación en varios puntos geográficos, de esta forma poder desarrollar una herramienta enfocada al dimensionamiento de sistemas fotovoltaicos.

3.1 Exploración y validación de datos y fuentes

3.1.1 Validación de datos correspondientes a estaciones meteorológicas

Como se había definido previamente existen a nivel nacional 53 estaciones meteorológicas de las cuales mantienen el registro de humedad relativa, presión atmosférica, radiación global, radiación reflejada, temperatura ambiental, dirección y velocidad del viento.

En el capítulo 1 se había dado a conocer las fuentes de datos que se utilizarán para la recolección de mediciones en tierra, siendo principalmente las estaciones del DMQ, INAMHI e INER debido a su capacidad de medición de irradiancia mediante piranómetros de tipo ISO9060 de primera y segunda clase asegurando la calidad de los datos, de acuerdo a la Tabla 3.1 se muestran los sensores usados por cada institución.

Tabla 3.1. Marca, modelo, clase de sensores tipo ISO9060 disponibles en estaciones meteorológicas de Ecuador. (Ordoñez et al., 2019)

Institución	Número de estaciones	Marca/Modelo de piranómetro	Clase
DMQ	6	Kipp&Zonen CM3	Segunda clase
INER	13	Kipp&Zonen CMP6	Primera clase
	31	Hukseflux SR11	Primera clase
INAMHI	2	Kipp&Zonen CM3	Segunda clase
	1	Kipp&Zonen CMP6	Primera clase

Como se puede notar en la Tabla 3.1, existen 45 estaciones de primera clase y 8 estaciones de segunda clase, siendo los sensores de primera clase los que presentan una mayor confiabilidad de los datos, sin embargo, cada institución mantiene un plan diferente de limpieza y calibración de los sensores y afecta directamente a la calidad de los datos.

Por parte de las estaciones del DMQ se mantiene vigente un protocolo de mantenimiento donde se realiza una calibración anual y la limpieza del sensor se realiza de forma mensual, a diferencia del INHAMI que realiza la calibración y

limpieza de los piranómetros semestralmente siendo un dato a tener en consideración en cuanto a la calidad de los registros. Finalmente, las estaciones del INER se encuentran bajo el custodio de la Escuela Politécnica del Chimborazo (ESPOCH) y Universidad Politécnica Salesiana (UPS) quienes realizan las tareas de mantenimiento (Ordoñez et al., 2019).

De acuerdo a Ordoñez et al., (2019) menciona la importancia de las labores de mantenimiento para obtener datos reales asociados a las diferentes variables ambientales, de esta manera en la Figura 3.2 se muestra la precisión de los datos captados por las diferentes estaciones de acuerdo a los diferentes propietarios, tomando como referencia la métrica del error cuadrático medio.

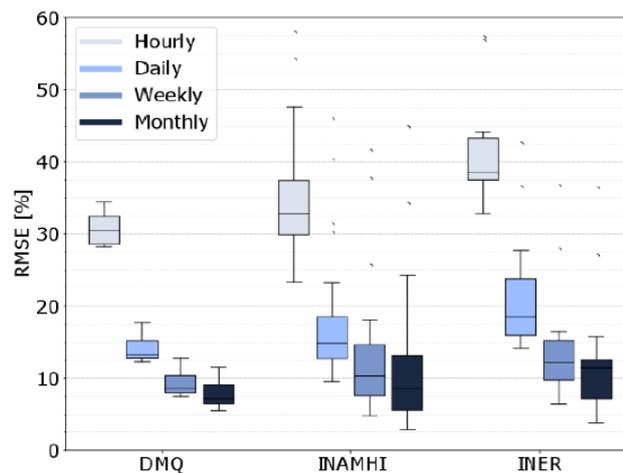


Figura 3.2. Gráfica de caja del RMSE de estaciones meteorológicas agrupadas por propietario. (Ordoñez et al., 2019)

Los datos agrupados por hora, día, semana y mes que presentan un margen de confiabilidad más estrecho corresponden a las estaciones del DMQ cuyo RMSE para datos agrupados por día, semana y mes varía desde el 5% al 15%, indicando ser una fuente de datos apropiada para la extracción de registros de irradiancia.

3.1.2 Delimitación del área de estudio

La mayoría de las estaciones meteorológicas usadas en la comparación, usan piranómetros de primera clase (ISO 9060). Adicionalmente, tanto el INAMHI como el IIGE implementaron estos piranómetros en 2014 y se ha mantenido un ciclo

contante de calibración de los sensores de forma anual. En cuanto al protocolo de mantenimiento adoptado, para asegurar la fiabilidad de las mediciones se realizan tareas de limpieza de los domos de manera mensual a diferencia de las estaciones del INAMHI donde las tareas de calibración y limpieza se realizan de forma semestral (Ordoñez et al., 2019).

La acumulación de suciedad en los piranómetros durante largos periodos de tiempo incide directamente en la precisión de las mediciones, especialmente a partir del año 2015 no se disponen de registros de mantenimiento preventivo de los equipos, lo que tiene como consecuencia que muchos de los datos almacenados se encuentren incompletos y no garantizan mediciones precisas.

Ordoñez et al., (2019) realizó una comparativa de la semejanza de los datos proporcionados por los datos NREL (satelitales) con respecto a datos tomados por estaciones meteorológicas cuyos datos fueron agrupados por horas, días, semanas y meses como se muestra en la Figura 3.3. utilizando métricas RMSE y MBE para medir el ajuste de los datos reales y estimados.

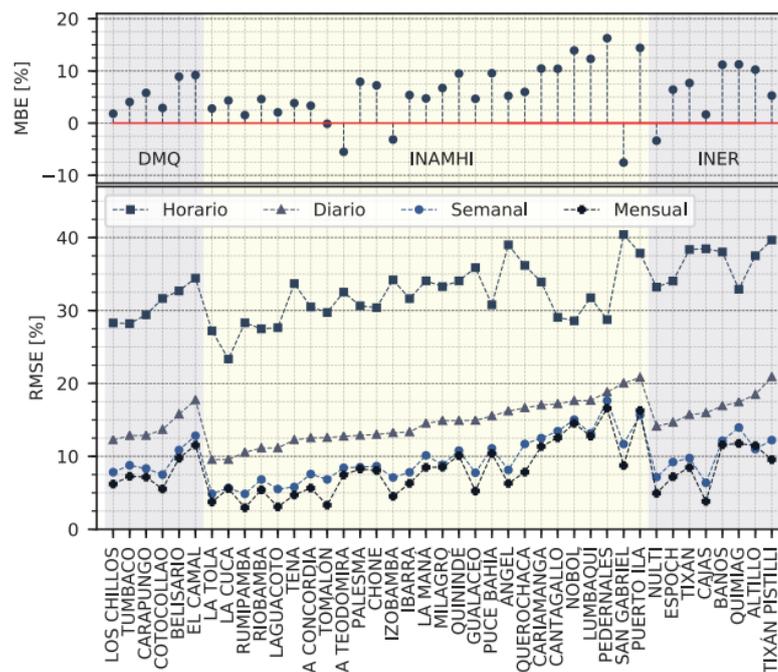


Figura 3.3. Comparativa de bases de intervalos de medición entre datos reales y estimados. (Ordoñez et al., 2019)

De acuerdo a los resultados mostrados en la gráfica, el RMSE de la base horaria alcanza valores entre el 30 al 50% indicando una baja confiabilidad en el uso de este modelo, sin embargo, la base de registros de forma diaria una gran mayoría de estaciones presenta un RSME inferior al 20% indicando una alta confiabilidad para el uso de los datos y manteniendo una resolución temporal aceptable.

Respecto al MBE, al ser una resta simple, este coincide en todas las bases. Un valor positivo de este indicador significa una sobre-estimación del modelo frente a las mediciones. En 36 de las 41 las estaciones este valor es positivo, lo que indica que el modelo sobre-estima el recurso solar en este mismo valor. En promedio, el sesgo de la comparación (error MBE) es del 8%.

3.1.3 Análisis exploratorio datos de estaciones meteorológicas

Existe una amplia cantidad de estaciones de tipo meteorológicas distribuidas en diversos puntos del país correspondientes a diversas instituciones encargadas del estudio de variables ambientales, sin embargo, no todas las estaciones proporcionan datos válidos que puedan representar la realidad de una variable ambiental en un punto determinado.

Para el desarrollo del presente proyecto se utilizarán los datos de 26 estaciones ubicadas en la provincia de Pichincha correspondientes al DMQ, en la Figura 3.4 se muestra la ubicación geográfica de cada estación y se puede verificar que existe una amplia variedad de puntos que permitirán analizar la luz solar en varios puntos de la provincia de Pichincha.

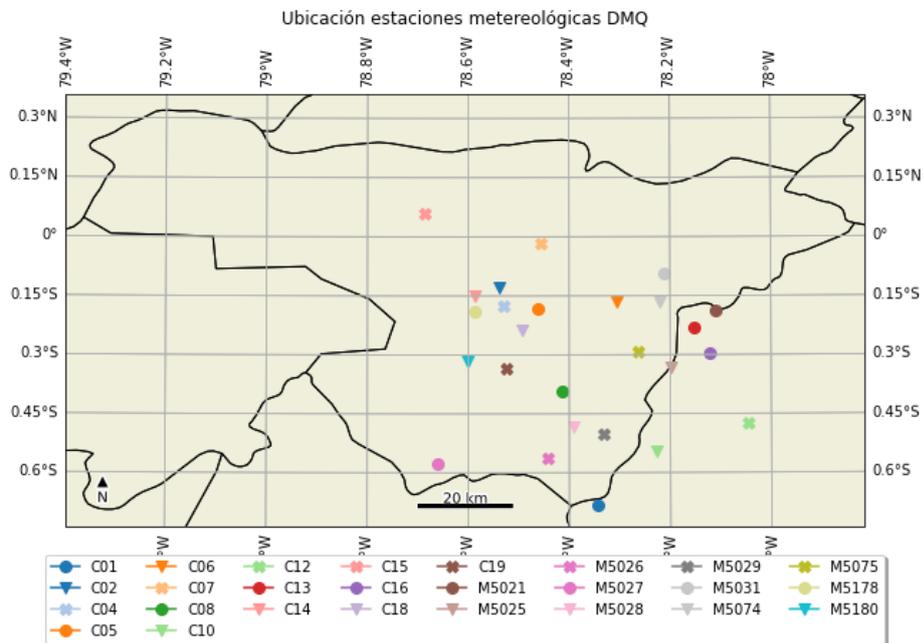


Figura 3.4. Ubicaciones de estaciones meteorológicas del DMQ.

En la Tabla 3.2 se muestra información y la ubicación geográfica correspondiente a las 26 estaciones de tipo meteorológica que se encuentran distribuidas en las provincias de Pichincha y Napo que comprenden la red del DMQ, cabe recalcar que los datos geográficos se encuentran en formato UTM-WGS84.

Tabla 3.2. Datos identificativos y ubicación de estaciones meteorológicas ubicadas en las provincias de Pichincha y Napo. (FONAG, 2022)

Nro.	Código	Nombre	Cantón	Parroquia	Este	Norte	Altura (m.s.n.m.)	Inicio operación
1	C01	Maucatambo	Archidona	Archidona	9924699	795679	3840	jul-2003
2	C02	Rumihurco	Quito	Condado	9985645	774102	3563	may-2000
3	C04	Rumipamba	Quito	Rumipamba	9980593	774790	3327	oct-2000
4	C05	Bellavista	Quito	Jipijapa	9979904	782541	2960	may-2000
5	C06	Yaruquí	Quito	Yaruquí	9981813	799837	2679	oct-2000
6	C07	San Antonio	Quito	Belisario Quevedo	9997961	783242	2455	dic-2017
7	C08	IASA	Quito	Sangolquí	9956633	787805	2728	dic-2001
8	C10	La Mica Presa	Archidona	Cotundo	9939530	808537	3922	jul-2008
9	C11	Pita Campamento	Mejía	Machachi	9945105	785140	3360	nov-2002
10	C13	Salve Faccha Campamento	Chaco	Oyacachi	9974472	816985	3889	oct-2015

11	C15	Nanegalito	Quito	Nanegalito	1E+07	757636	1767	feb-2012
12	C16	Guyataloma	Quijos	Papallacta	9967063	820381	3783	ene-2014
13	C18	Puengasí	Quito	Itchimbía	9973816	779044	2971	ene-2005
14	C19	El Troje	Quito	Turubamba	9963118	775644	3123	may-2000
15	M5021	Yuracfaccha Oyacachi	Chaco	Oyacachi	9979154	821647	3710	jun-2007
16	M5025	La Virgen Papallacta	Quijos	Papallacta	9963068	811859	4020	jun-2007
17	M5026	Cotopaxi Control Norte	Mejía	Machachi	9937618	784573	3670	jun-2007
18	M5027	Loma Hurco Ilizizas	Mejía	Chaupi	9936280	760317	3727	jun-2007
19	M5028	Hcda Prado Miranda	Quito	Pintag	9946524	790439	3526	ene-2010
20	M5029	El Carmen	Quito	Pintag	9944491	796826	4100	ene-2010
21	M5031	Chumillos	Cayambe	Cangahua	9989510	810520	3750	ene-2010
22	M5074	Puntas	Quito	Checa	9981721	809389	4142	dic-2011
23	M5075	Itulcachi	Quito	Pifo	9967879	804574	4029	dic-2011
24	M5178	Tayango - Guagua Pichincha	Quito	Lloa	9978969	768518	4090	jul-2018
25	M5179	Paluguillo	Quito	Pifo	9966060	808109	3717	sep-2017
26	M5180	Atacazo Antenas CNT	Quito	La ecuatoriana	9964786	766935	3887	sep-2019

Una vez definidas las estaciones de interés, en la Figura 3.5 se muestra el procedimiento que se seguirá para realizar el análisis exploratorio de los mismos, la descarga se llevará a cabo de forma manual a través del portal de consultas en la página del Sistema de Estandarización de Datos Hidroclimáticos Crudos (SEDC).

A través del repositorio del SEDC se pueden acceder a los datos almacenados de cada estación y se puede realizar su descarga ingresando un periodo de tiempo, siempre y cuando los datos se encuentren disponibles se hará su descarga respectiva, cabe destacar la posibilidad de acceder a diversas variables ambientales, sin embargo, para el desarrollo del presente proyecto se hará uso únicamente de la variable “Radiación solar”.

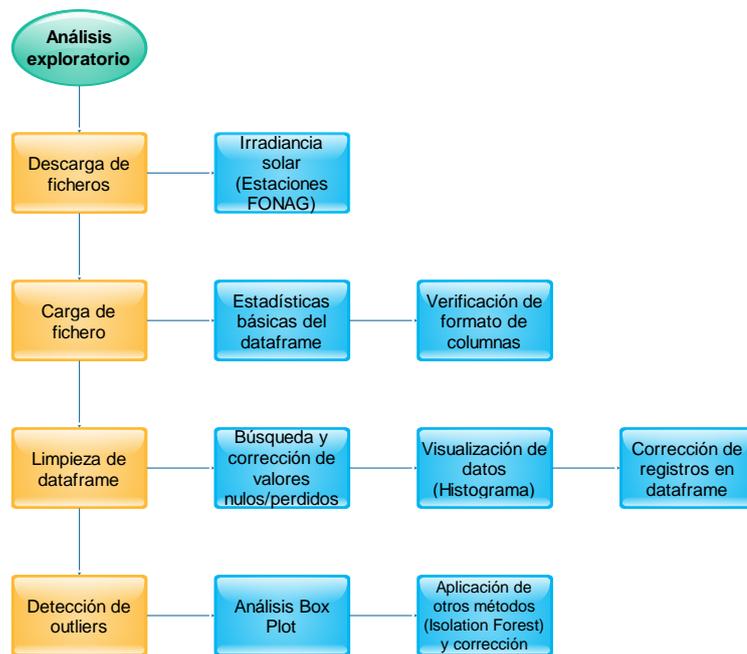


Figura 3.5. Flujo de proceso para el análisis exploratorio de irradiancia.

La carga de los ficheros se lo realizará utilizando el lenguaje Python y la plataforma Jupyter Notebook, posteriormente se aplicará estadísticas básicas a las variables del dataset para tener una percepción de los datos cargados, para el caso de los datos correspondientes a las fechas y horas de cada registro deben ser tratados como una variable de tipo “datetime”.

Como se había mencionado previamente, los datos procedentes de estaciones no mantienen una continuidad temporal constante y diversos factores externos pueden afectar a los datos almacenados, por lo tanto, es necesario detectar y corregir los valores perdidos y nulos del dataset, de la misma manera identificar valores atípicos que puedan encontrarse asociados a fallos en el sensor.

Para tener una perspectiva más amplia la Figura 3.6 muestra visualmente los registros que se encuentran disponibles en las diversas estaciones meteorológicas en el período comprendido desde enero de 2015 hasta julio de 2022, los registros vacíos se encuentran representados en negro y los datos disponibles para su análisis se muestran en color.

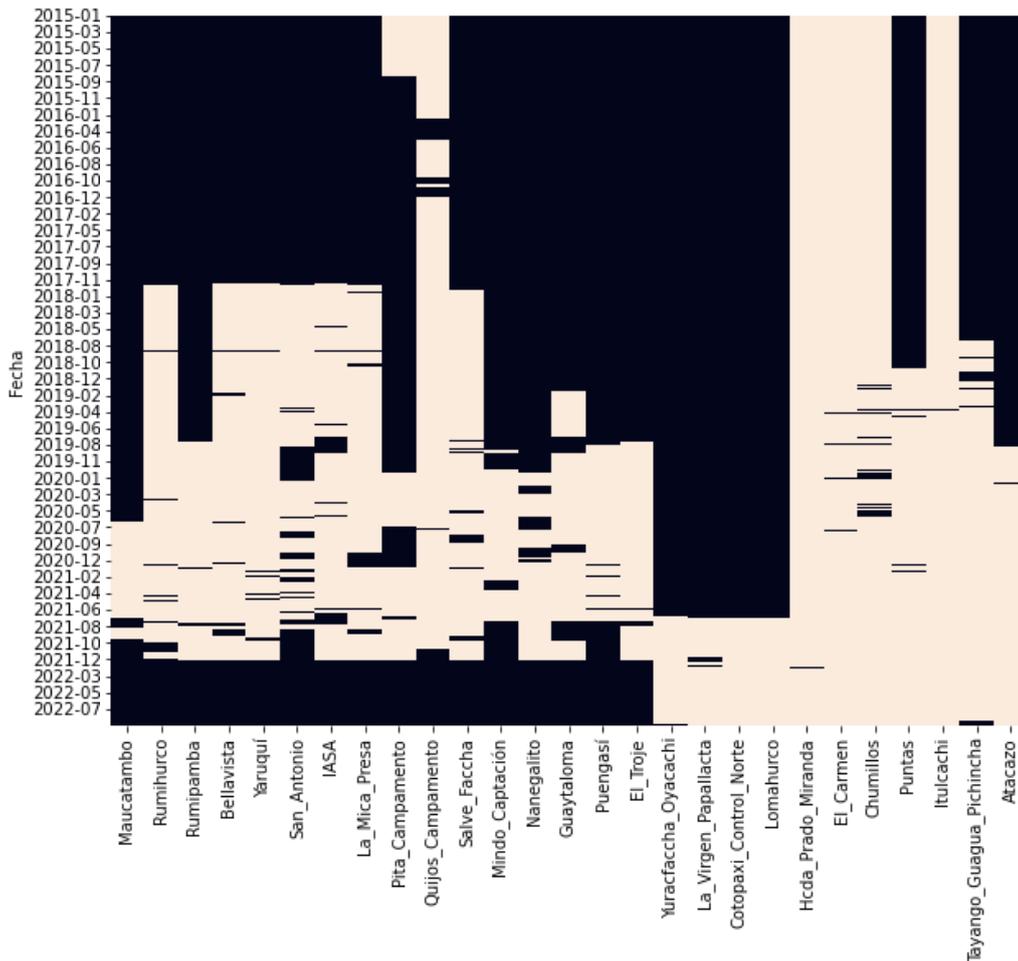


Figura 3.6. Representación de disponibilidad de datos desde 2015 hasta 2022 en las estaciones del DMQ.

Finalmente, es necesario identificar valores atípicos que no reflejen la naturaleza de la radiación solar e inclusive detectar daños o la acumulación de suciedad sobre el sensor, para la detección de outliers en series temporales se hará uso de Isolation Forest.

A continuación, se describirá el desarrollo del análisis exploratorio de la estación M5021 - Yurafaccha Oyacachi y que deberá ser replicado para las 26 unidades descritas en la tabla Tabla 3.2, a su vez, la extracción de datos satelitales debe ser realizada de forma individual utilizando como referencia la ubicación geográfica correspondiente a cada estación para su respectivo ajuste.

En la Tabla 3.3 se describe las estadísticas básicas de la estación M5021 - Yurafaccha Oyacachi, localizada en los límites de Pichincha con la provincia de Napo que mantiene registros desde julio de 2021 hasta la actualidad.

Dentro del resumen se puede visualizar la cantidad de registros por variable, el promedio calculado con todos los datos que componen la serie de tiempo asociado a cada variable ambiental, los valores mínimos y máximos, la cantidad de datos en función del respectivo cuartil siendo este último de utilidad para identificar la acumulación de datos dentro de la distribución de las series.

Tabla 3.3. Análisis de variables ambientales disponibles en la estación M5021.

Propiedad	Humedad	Presión	R. Global	R. Reflejada	Temperatura	Dirección V.	Velocidad V.
Cantidad	36.990	36.990	36.990	36.990	36.988	31.832	31.832
Promedio	75,09	757,32	219,48	40,36	15,24	168,54	2,06
Desviación estándar	26,58	1,51	316,91	57,44	4,15	110,75	1,64
Valor mínimo	-3,00	749,40	0,00	0,00	3,00	1,00	0,00
25%	55,70	756,40	0,00	0,18	12,20	98,00	1,00
50%	82,85	757,50	7,28	1,85	14,00	134,00	1,60
75%	100,00	758,40	397,93	73,95	18,50	285,00	2,60
Valor máximo	100,00	762,10	1.236,67	249,78	34,00	360,00	12,80

Como se había mencionado previamente, cada variable ambiental se mide de forma individual e independiente a otras variables y se ve reflejada en la cantidad de registros en un mismo intervalo de tiempo, asimismo, se pueden detectar anomalías en los registros cuando se verifica la existencia de límites que no se podrían dar en condiciones normales, indicando una posible avería en los sensores de la estación.

Dentro de las tareas de verificación de calidad de los datos, es importante identificar valores perdidos y nulos, a nivel general la Tabla 3.4 muestra la cantidad de valores perdidos en el dataset de la estación M5021 con respecto a 36.990 registros en la mayoría de variables, se puede identificar una gran cantidad de

valores faltantes en la velocidad y dirección del viento, indicando la necesidad de imputación de valores perdidos.

Tabla 3.4. Valores perdidos en estación M5021

Variable	Cantidad de valores perdidos	Porcentaje respecto al total de registros
Fecha	0	0%
Humedad	3	0.01%
Presión	3	0.01%
R. Global	3	0.01%
R. Reflejada	3	0.01%
Temperatura	5	0.01%
Dirección V.	5161	16.21%
Velocidad V.	5161	16.21%

Para tener una perspectiva más amplia de los datos, mediante el empleo de histogramas se puede visualizar la distribución de los datos con respecto a cada variable como se muestra en la Figura 3.7, la concentración de los datos puede ayudar a identificar comportamientos anómalos en el funcionamiento de sensores o a su vez, describir la naturaleza como en el caso de la irradiancia solar que muestra una alta cantidad de valores cercanos o iguales a cero asociados con las mediciones nocturnas donde no existe la incidencia de luz solar, indicando la importancia de tratar los datos de tal forma que se evite el sesgo durante las etapas de ajuste y pronóstico planteadas previamente.

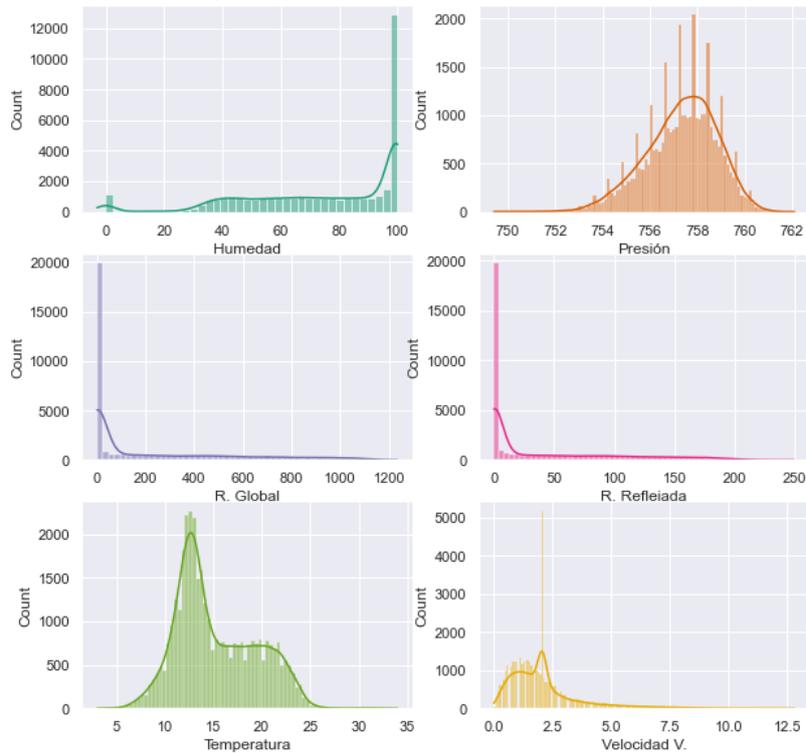


Figura 3.7. Gráfica de distribución de datos en variables ambientales

Para tener una percepción estadística detallada del comportamiento de datos de irradiancia se elabora un diagrama de caja únicamente de las variables de radiación solar como se muestra en la Figura 3.8, se puede enfatizar en la clara acumulación de datos en torno a cero correspondiente a ciclos nocturnos donde se encuentra la mediana tanto para la radiación global y reflejada, sin embargo, se puede visualizar que se registran datos sobre el extremo del bigote superior y debido a la naturaleza del potencial solar no se puede considerar como outliers debido a que dichos datos son los que se deben mantener para su respectivo análisis.

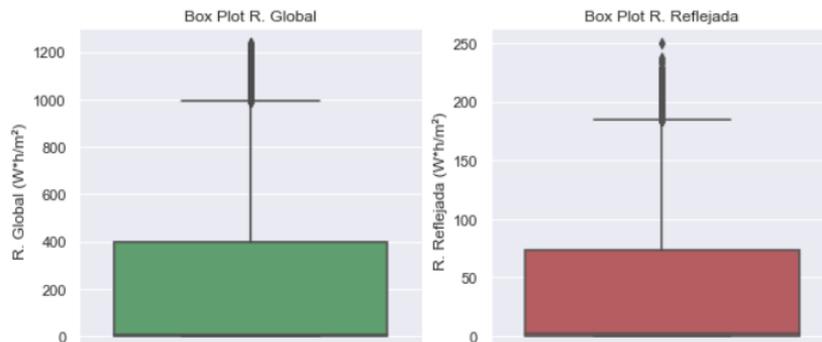


Figura 3.8. Box Plot de radiación global y reflejada.

Finalmente, para identificar valores atípicos se utilizará el método basado en árboles de clasificación y regresión (Isolation Forest) que permite identificar anomalías en una serie de tiempo para su posterior tratamiento como se muestra en la Figura 3.9.

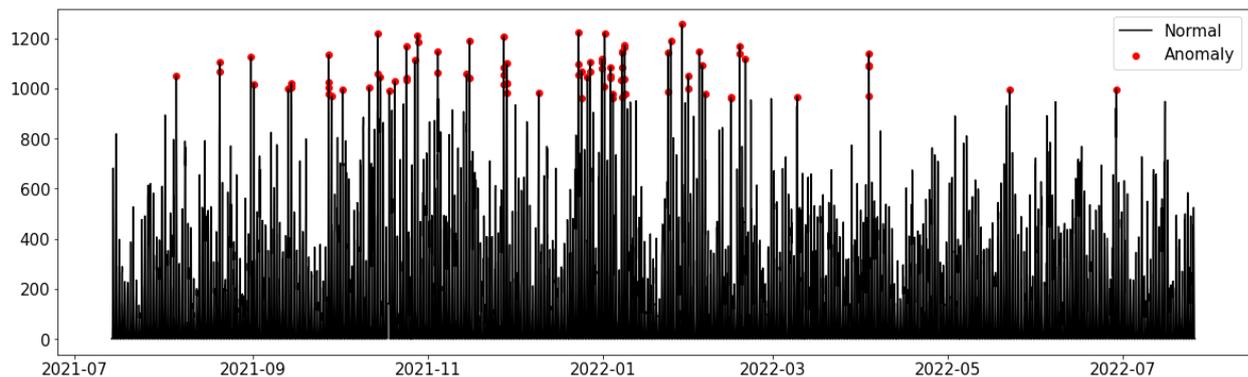


Figura 3.9. Isolation Forest para detección de outliers.

3.1.4 Descarga de datos satelitales

Una vez que se ha identificado la cobertura de las estaciones de tipo meteorológicas del DMQ y sus respectivas ubicaciones, se puede verificar la disponibilidad de datos en gran parte de la superficie de la provincia de Pichincha la misma que será considerada en su totalidad para la descarga de datos desde la base de datos satelital del repositorio NREL.

El dataset considerado para la descarga de datos es “USA & Americas (30, 60 min / 4 km / 2021)”, como se puede observar en su nomenclatura el conjunto de datos abarca Norteamérica y parte de Sudamérica aproximadamente hasta el límite sur de Bolivia esta información es proporcionada por el satélite GOES East, los registros se encuentran disponibles en intervalos de 30 y 60 minutos, finalmente, la resolución espacial por pixel de la imagen satelital es 4 km en la superficie.

Teniendo en cuenta las especificaciones del dataset de referencia, se dividirá la extensión de la provincia de Pichincha en puntos equidistantes cada 4 km dando como resultado una matriz comprendida por 1080 puntos coordenados

construidos a partir de la combinación de 27 latitudes y 40 longitudes como se muestra en la Figura 3.10, los límites considerados en el área de interés son:

- Límite norte: 0.3595
- Límite sur: -0.7391
- Límite este: -77.8176
- Límite oeste: -79.3953

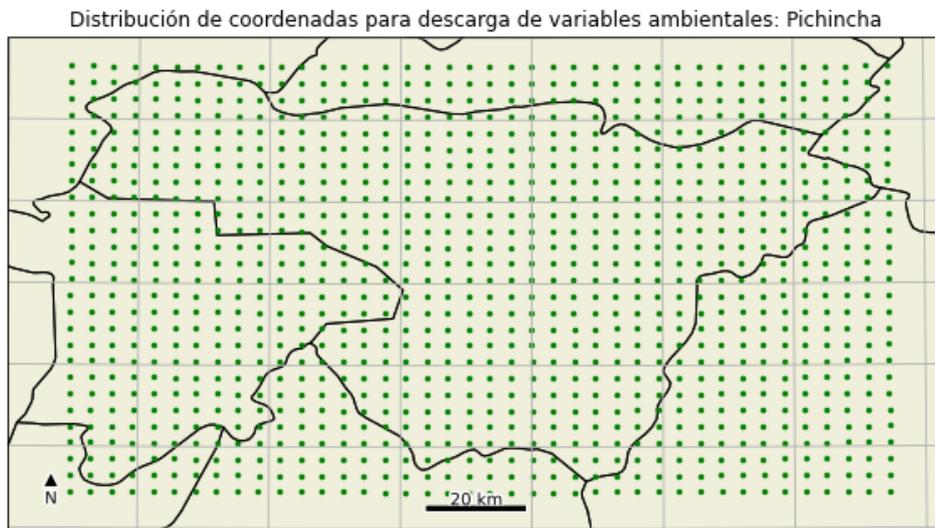


Figura 3.10. Distribución de puntos coordinados para descarga de variables ambientales.

Para continuar con la descarga de datos se puede realizar de forma manual a través del portal oficial de NSRDB que permite seleccionar ubicaciones por puntos o especificar un área de interés en concreto, sin embargo, este método se encuentra limitado impidiendo realizar la descarga de áreas aproximadamente igual a 1600 km², el área que se requiere descargar equivale a 17.280 km².

Para abarcar áreas extensas se dispone del acceso a los datos a través de la API propia de NREL que puede ser consultado a través de Python, para ello es necesario solicitar un permiso de acceso a través de un correo y una clave, hay que tener en cuenta que existen restricciones en cuanto a la cantidad de solicitudes que se pueden realizar.

De acuerdo a la guía de descarga de datos para el uso de la API, cada usuario se encontrará limitado a realizar 5.000 solicitudes cada 24 horas, se acepta únicamente una solicitud por segundo con un máximo de 20 solicitudes en cola. Dependiendo de la cantidad de atributos por solicitud (variables) el peso máximo de descarga por solicitud es 5 Mb, los datos son almacenados en formato CSV (NREL, 2022).

Debido a la amplia cantidad de puntos que requieren ser analizados y el tiempo a ser empleado para la descarga de los mismos, es un factor a tener en cuenta por tal motivo solamente será considerado el año 2021 para el desarrollo del presente proyecto.

3.1.5 Análisis descriptivo de datos satelitales

Mediante el repositorio NREL se puede acceder a múltiples variables ambientales acorde a una ubicación geográfica definida por un punto o un área determinada, dicha información es descargada en ficheros individuales (CSV) y posteriormente combinados en un solo archivo (netCDF) que puede aprovechar su capacidad de representar información geoespacial multivariable y mantener la dimensión del tiempo, se dispone de la siguiente estructura de datos mostrado en la Tabla 3.5.

Tabla 3.5. Atributos de fichero netCDF

Dimensiones		
Tiempo: 17520	Latitud: 27	Longitud: 40
Coordenadas		
Time	Datetime64	2021/01/01 hasta 31/12/2021
Lat (Latitud)	Float32	-0,71 hasta 0,33
Lon (Longitud)	Float32	-79,38 hasta -77,82
Variables ambientales		
DHI (Irradiancia directa horizontal)	Float32	(time, latitude, longitude)
DNI (Irradiancia normal directa)	Float32	(time, latitude, longitude)
GHI (Irradiancia global horizontal)	Float32	(time, latitude, longitude)

Temperatura	Float32	(time, latitude, longitude)
Presión	Float32	(time, latitude, longitude)
Humedad relativa	Float32	(time, latitude, longitude)
Dirección del viento	Float32	(time, latitude, longitude)
Velocidad del viento	Float32	(time, latitude, longitude)
Precipitación	Float32	(time, latitude, longitude)

Gráficamente, la estructura del formato netCDF se muestra en la Figura 3.11, donde se puede visualizar una matriz de puntos definidos por una latitud y longitud (componente geográfica) y en cada celda el registro de la variable de interés, en el ejemplo propuesto la gráfica representa la variable “temperatura del aire” en tres periodos de tiempo diferentes y por fines de visualización han sido segmentados por color, no obstante, la variable de interés puede cambiar bajo ciertas condiciones (cambio de temperatura en función de la altitud, estación del año, hora del día, etc.) o a su vez, en el mismo punto y tiempo, tener acceso a mediciones de múltiples variables dando origen a una matriz de 4 dimensiones.

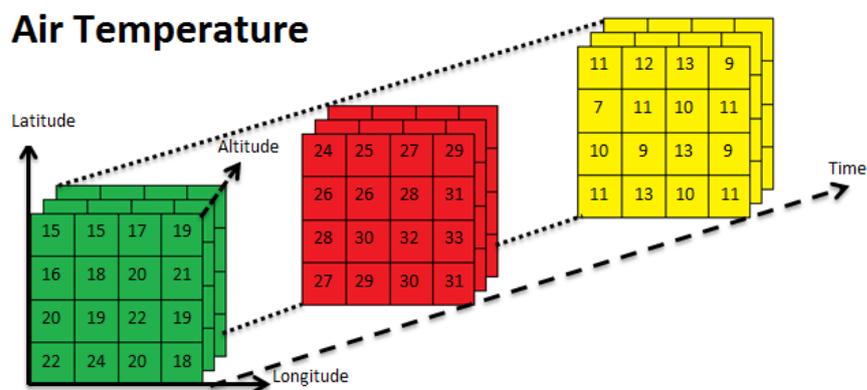


Figura 3.11. Estructura matriz 4D de archivos netCDF. (GeoSolutions, 2022)

De acuerdo a la estructura detallada en la Tabla 3.5 el archivo netCDF generado dispone de una malla de puntos comprendida entre 27 latitudes y 40 longitudes (1080 coordenadas) cuyo centro específico se encuentra localizado en la provincia de Pichincha como se muestra en la Figura 3.12 debido a que los datos que se utilizarán de las estaciones terrestres se encuentran ubicadas en dicha provincia.



Figura 3.12. Centro de análisis de archivo netCDF.

En la Figura 3.13 se representan los 1080 puntos de la malla del archivo netCDF y su respectiva representación en el mapa del Ecuador, la distancia existente entre puntos es aproximadamente 4 kilómetros lo que facilitará la diferenciación de zonas de alto potencial solar dentro de la provincia de Pichincha, para la ilustración en concreto se muestra únicamente la irradiancia directa horizontal (DHI) captada por satélite el 02/01/2021 a las 13H30, téngase en cuenta que existen otras 8 variables ambientales (DNI, GHI, temperatura, presión, humedad relativa, dirección del viento, velocidad del viento, precipitación) y cada una posee su respectiva representación espacial (1080 puntos geográficos) y su variable temporal (17520 registros horarios).

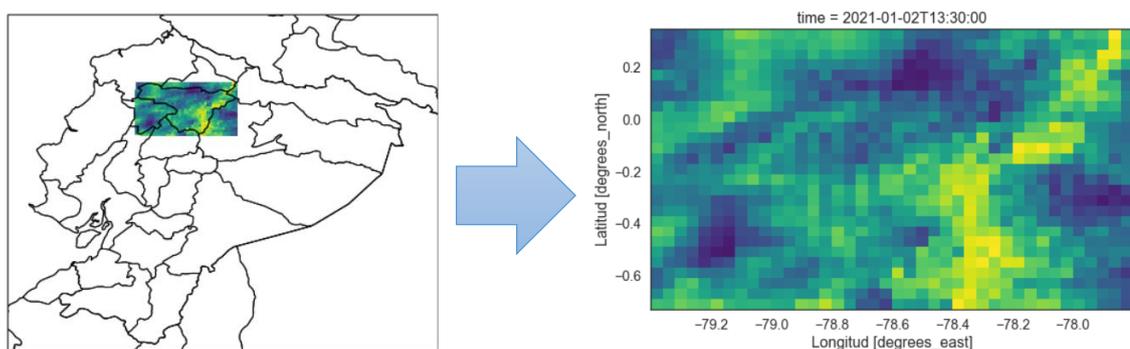


Figura 3.13. Visualización de irradiancia solar de archivo netCDF.

De acuerdo a la estructura mostrada previamente, se tienen 1080 puntos repartidos en la superficie del país y cada punto tiene almacenado 17520 registros correspondientes a las variables ambientales de temperatura, radiación solar, humedad relativa y presión atmosférica medidas desde el 01/01/2021 hasta 31/12/2021, de esta forma se pueden integrar fácilmente múltiples variables asociadas a un punto en el espacio y conservar su evolución en el tiempo.

En la Figura 3.14 se muestra la representación visual de los datos geográficos definidos en la Tabla 3.5, teniendo en cuenta que los datos se registran de forma horaria se puede notar que se han graficado los datos de irradiancia de treinta minutos correspondientes al día 02/01/2021, de esta manera se puede tener una visión general del tipo de información que se encuentra contenida en el archivo obtenido del repositorio NREL.

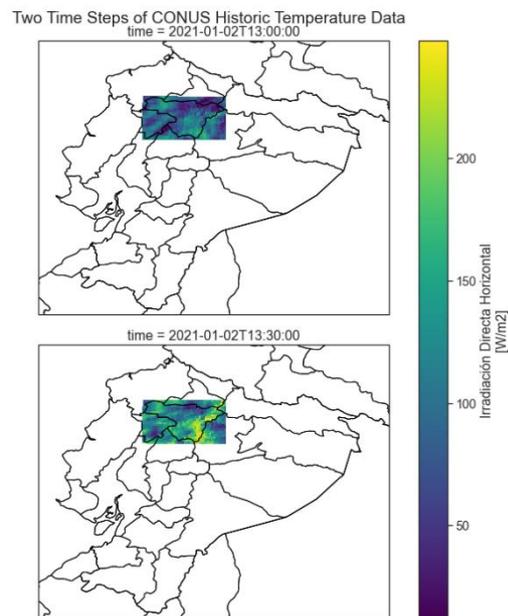


Figura 3.14. Representación de irradiancia a partir de archivo netCDF.

Teniendo en cuenta los puntos geográficos de las estaciones que se requieren analizar, poseen su respectiva correspondencia con una latitud, longitud e intervalo de tiempo en la cual se puede extraer la información siempre y cuando se encuentren contenido dentro de los límites geográficos que se encuentran dentro de la descripción de la Tabla 3.5.

Finalmente, el proceso de extracción de información correspondiente a un punto se puede completar y analizar cada una de las variables que se encuentran guardadas, en la Figura 3.15 se muestra un fragmento de los registros de irradiancia solar que se asemeja a los datos obtenidos por las estaciones meteorológicas y cuyo análisis exploratorio se realizará utilizando las mismas consideraciones para datos medidos en tierra.

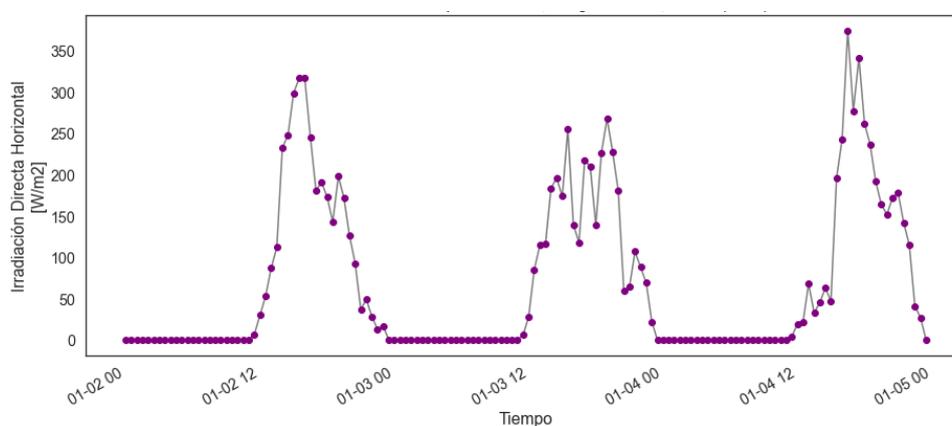


Figura 3.15. Serie de irradiancia solar correspondiente a un punto.

3.2 Ajuste de datos locales y satelitales

Con la finalidad de analizar el comportamiento de la radiación solar en la provincia de Pichincha se han definido dos fuentes de datos: estaciones meteorológicas en tierra y datos satelitales, sin embargo, los datos medidos en tierra pese a su alta confiabilidad la mayoría de los registros se encuentran incompletos y en contraste, datos satelitales aunque se encuentran completos y existe la posibilidad de acceder a diferentes periodos en el tiempo no son precisos lo que implica correr un alto riesgo de asumir mediciones con una tolerancia elevada con respecto a los datos reales.

Para aprovechar las ventajas de cada fuente de datos, se pretende construir un nuevo dataset a partir de la implementación de regresión basada en técnicas de Machine Learning donde las variables predictoras serán todas las medidas de variables ambientales captadas mediante satélite y la variable a predecir será la

radiación solar captada por estaciones meteorológicas, considerando el mismo intervalo de tiempo y la misma ubicación geográfica, la Figura 3.16 muestra el modelo propuesto.

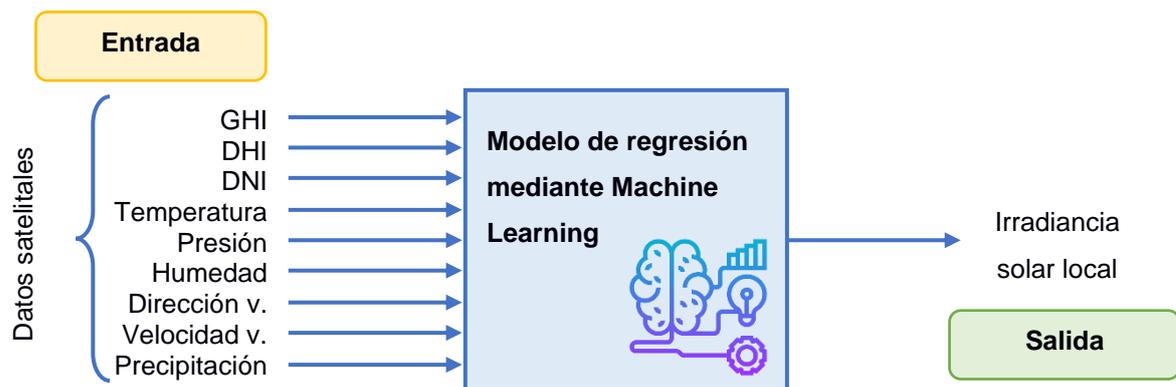


Figura 3.16. Diagrama de bloques para regresión de irradiación solar mediante Machine Learning.

La metodología propuesta será implementada para las 26 estaciones disponibles en el área de estudio para el año 2021, dando como resultado la misma cantidad de modelos diferentes asociados a cada estación y su respectiva ubicación geográfica, posteriormente, los modelos serán empleados para ajustar la radiación solar del resto de ubicaciones dentro del área de interés a través de interpolación usando como referencia los modelos de estaciones aledañas.

Los ajustes de datos satelitales para cada ubicación serán almacenados individualmente para crear un dataset artificial que permita representar la realidad de la luz solar en cada punto del área de estudio durante el año 2021, el formato a emplear será netCDF debido a la capacidad de almacenar las series de tiempo resultantes asociado a cada punto geográfico del área de estudio. Finalmente, este nuevo conjunto de datos será utilizado para el entrenamiento de modelos enfocados únicamente al pronóstico de irradiación solar.

Para el estudio de algoritmos y modelos enfocados al pronóstico de irradiación solar se propone la metodología mostrada en la Figura 3.17, partiendo de la construcción

de un dataset artificial utilizando las variables ambientales satelitales como predictoras de la irradiancia solar captada en tierra, de esta forma se puede asegurar un conjunto de datos que mantenga su continuidad en el tiempo gracias a los datos satelitales y a su vez, los registros captados sean cercanos a los valores reales obtenidos a partir de sensores especializados.

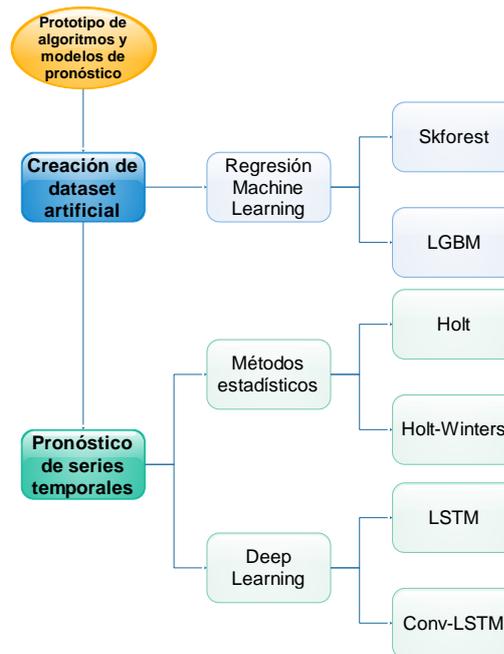


Figura 3.17. Metodología de ajuste de datos y pronóstico.

3.3 Prototipos de métodos y modelos para elaboración de dataset artificial de irradiancia solar

Para la creación del dataset artificial se empleará directamente métodos basados en Machine Learning para regresión, debido a la disponibilidad de diferentes variables predictoras, la facilidad de implementación, el uso de pocos recursos computacionales, alto rendimiento en proyectos similares para análisis de irradiación solar y la rapidez de entrenamiento de modelos.

De acuerdo a la revisión bibliográfica, existen varias alternativas basadas en Machine Learning que han dado buenos resultados para la regresión de series temporales multivariadas, para el desarrollo del presente proyecto se hará uso del modelo LGBM que se caracteriza por utilizar como base árboles de decisión que mejoran considerablemente su rendimiento a la vez que su entrenamiento es

relativamente veloz, también permite visualizar la importancia de las variables exógenas que fueron introducidas al modelo.

El modelo Skforest es otra alternativa basada en árboles de decisión con la particularidad de adaptar modelos de la librería Scikit-Learn para el pronóstico de series de tiempo, además se proporciona una amplia gama de herramientas de procesamiento de datos durante el entrenamiento de los modelos permitiendo explorar eficientemente diversas configuraciones de hiperparámetros para mejorar el pronóstico de irradiación.

Los modelos basados en Machine Learning pueden integrar múltiples variables que pueden usadas como predictores de otra variable y así mejorar su rendimiento. Para este caso de estudio, las variables usadas como predictoras serán los registros satelitales que corresponden a: DHI, DNI, GHI, temperatura ambiental, presión atmosférica, humedad relativa, nivel de precipitación, dirección y velocidad del viento.

La variable a predecir será la irradiancia solar medida por estaciones meteorológicas, siempre y cuando los datos hayan sido limpiados previamente y el entrenamiento del modelo se realice en el mismo intervalo de tiempo que se dispone en la estación en tierra, la segmentación utilizada será 80% para datos de entrenamiento y el 20% restante para evaluar el modelo de regresión como se muestra en la Figura 3.18.

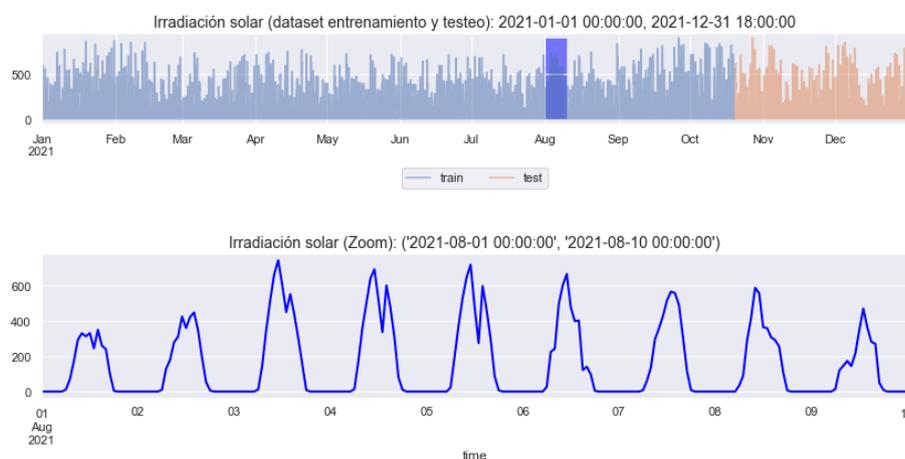


Figura 3.18. Segmentación de dataset para entrenamiento de modelos de regresión.

Hay que notar que la segmentación de los datos ha sido considerada de forma secuencial para mantener la integridad de la serie de tiempo, es por ello que no se puede realizar una selección de forma aleatoria como en otro tipo de aplicaciones.

3.3.1 Modelo Skforest

Skforest es una librería basada en Scikit-Learn que fue desarrollada específicamente para emplear modelos de regresión aplicada a modelos de pronóstico, con la particularidad de incorporar herramientas para el preprocesamiento de la serie temporal previo a su entrenamiento.

Dentro de la librería se encuentra disponible diferentes tipos de validaciones cruzadas de la serie temporal previa a su entrenamiento, siendo las más representativas: reajuste con origen fijo, reajuste con tamaño de entrenamiento fijo y validación sin reajuste.

Para el desarrollo del presente proyecto se hará uso del método de reajuste con origen fijo que es similar a una validación cruzada estándar, cuya diferencia radica en la selección de los datos para el entrenamiento no se realiza de forma aleatoria, la serie de datos es segmentada y el conjunto de entrenamiento crece secuencialmente manteniendo el orden temporal de los datos permitiendo el uso de todos los datos disponibles, una representación gráfica del método se muestra en la Figura 3.19.

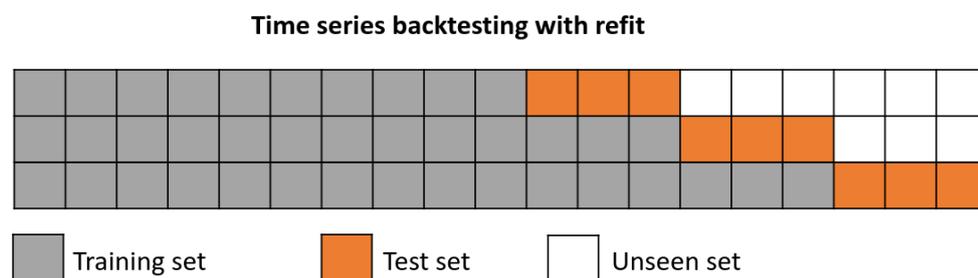


Figura 3.19. Reajuste Skforest con origen fijo. (Amat, 2021)

Una vez establecido el método de validación de los datos, se presenta la metodología de trabajo mediante el uso de Skforest en la Figura 3.20, utilizando principalmente la librería *ForecasterAutoreg* que convierte cualquier regresor

disponible en Scikit-Learn en un pronosticador recursivo a través del atributo *regressor*, para este caso de estudio será *RandomForestRegressor* y el último parámetro necesario *lags* define la cantidad de observaciones pasadas utilizadas como predictores.

En la etapa de entrenamiento se especificará la variable a predecir y las variables predictoras (exógenas), posteriormente se generará un modelo que permitirá estimar predicciones en el intervalo de testeo donde se aplicarán las métricas error absoluto medio (MAE), error cuadrático medio (RMSE) y R cuadrado (R2) para evaluar el rendimiento del modelo con respecto a los valores reales.

La evaluación del conjunto de métricas deberá ser optimizado a través del ajuste continuo de hiperparámetros del modelo utilizado, de esta forma obtener un modelo que se encuentre apto para generar una regresión confiable.

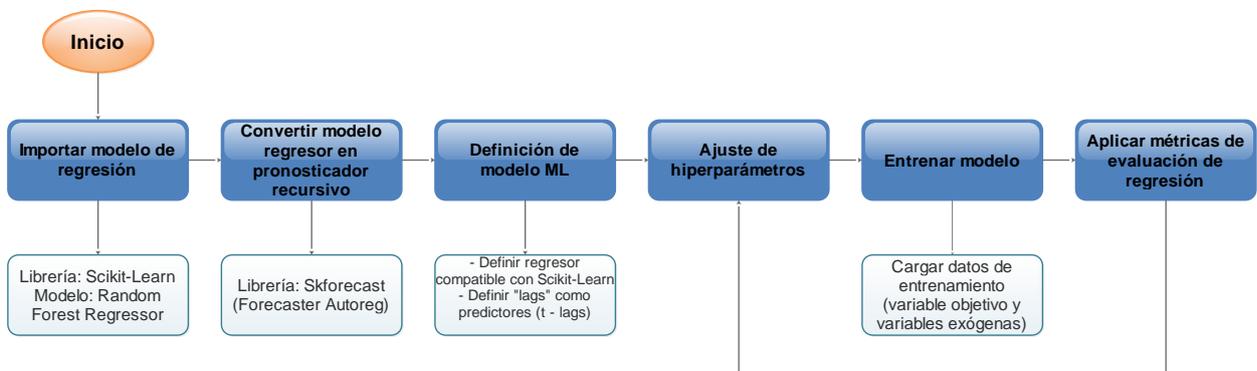


Figura 3.20. Metodología para modelo Random Forest mediante Skforecast

Una vez definido el procedimiento para el entrenamiento del modelo de Machine Learning es importante ajustar los hiperparámetros que permitan obtener el mejor rendimiento del modelo, para ello se realizará múltiples iteraciones con diferentes configuraciones de la cual se tomará únicamente el mejor modelo, todo esto se puede automatizar a través de la herramienta *grid_search_forecaster* disponible dentro de la librería de *Skforecast*. Los parámetros necesarios son:

- Forecaster: Modelo de pronóstico.

- `y`: variable a predecir.
- `exog`: conjunto de variables predictoras.
- `param_grid`: Conjunto de configuraciones correspondientes al modelo a ser probadas mediante iteraciones.
- `lags_grid`: Cantidad de observaciones pasadas usadas para pronóstico.
- `steps`: Espacio de observaciones a futuro a predecir.
- `refit`: Activar validación cruzada (True o False)
- `metric`: Métrica de evaluación utilizada para elegir el mejor modelo.
- `initial_train_size`: Cantidad de datos para entrenamiento a partir del 80% de datos destinado para entrenamiento.
- `return_best`: Almacenar únicamente el mejor modelo (True o False)
- `verbose`: Reporte del estado del entrenamiento (False)

Dentro de la función `grid_search_forecaster` existe una amplia cantidad de parámetros, pero los más importantes para este caso se han especificado previamente, de acuerdo al diagrama de flujo mostrado en la Figura 3.21 se muestra la metodología de trabajo empleada, partiendo de la definición del modelo de regresión y un listado de parámetros en el cual se realizará la exploración para el desarrollo del proyecto se ha considerado un conjunto de retrasos que serán utilizados como periodos de pronóstico, la cantidad y la profundidad de los árboles de decisión.

Todos los parámetros serán integrados en las instrucciones de la función de búsqueda donde se pueden realizar una amplia variedad de ajustes, especificar el modelo, el conjunto de datos, el conjunto de parámetros, habilitar la validación cruzada y establecer la métrica que será utilizada como referencia para identificar al mejor modelo de forma automática.

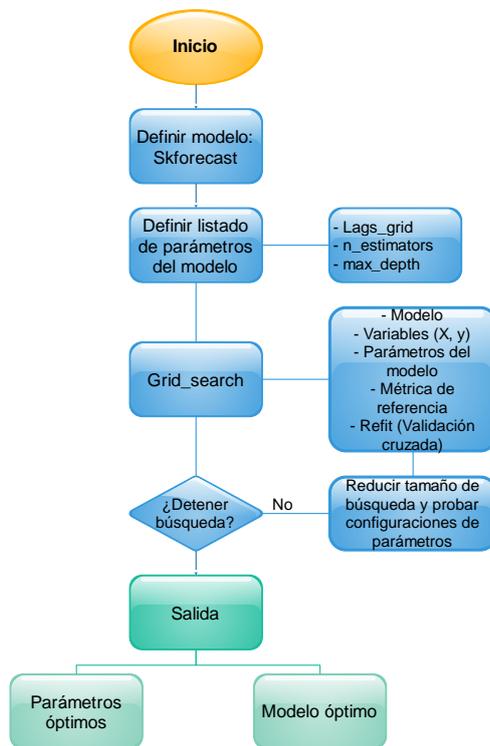


Figura 3.21. Descripción de funcionamiento de la herramienta grid_search_forecaster

Una vez finalizado el entrenamiento, la función almacena en un espacio de memoria denominado *forecaster* el mejor modelo obtenido, no es necesario asignar una variable nueva para almacenar el modelo. A partir de este punto se puede evaluar los resultados del modelo a partir de los datos de prueba y guardar el modelo en un fichero para utilizarlo posteriormente o inclusive continuar el entrenamiento con otros hiperparámetros.

3.3.2 Modelo LGBM

Otro modelo que ha sido ampliamente utilizado para el pronóstico de series temporales es LGBM debido a su precisión y velocidad de entrenamiento, siendo este último un factor importante a considerar debido a la cantidad de puntos asociados a cada estación en tierra en la provincia de Pichincha que serán analizados individualmente.

La estructura de los datos se mantendrá igual a la utilizada con Skforest, es decir, 80% de datos empleados para entrenamiento y el 20% restante para la validación

al igual que la validación cruzada de reajuste con origen fijo que puede ser reutilizada para el entrenamiento de LGBM como se mostró previamente en la Figura 3.18.

El modelo LGBM se encuentra disponible en la librería *lightgbm*, en la cual se puede configurar ciertos hiperparámetros básicos para dar inicio al entrenamiento del modelo como:

- *num_leaves* define la cantidad de hojas asociados al modelo.
- *max_depth* la profundidad del árbol de decisión.
- *learning_rate* la tasa de aprendizaje que se puede ir adaptando en función del entrenamiento.
- *n_estimators* que define la cantidad de árboles a ser entrenados.
- *min_child_samples* establece la cantidad mínima de datos necesarios por hoja.
- *max_bin* establece la cantidad de contenedores de memoria habilitados para almacenar datos de las variables predictoras.

La metodología propuesta se muestra en la Figura 3.22 que en esencia es similar a Skforest con la diferencia que la definición del modelo se realiza directamente desde la función *LGBMRegressor* donde se especificarán los hiperparámetros mencionados previamente, de esta manera se realiza un entrenamiento cuyo modelo resultante permitirá obtener predicciones en el periodo de testeo.

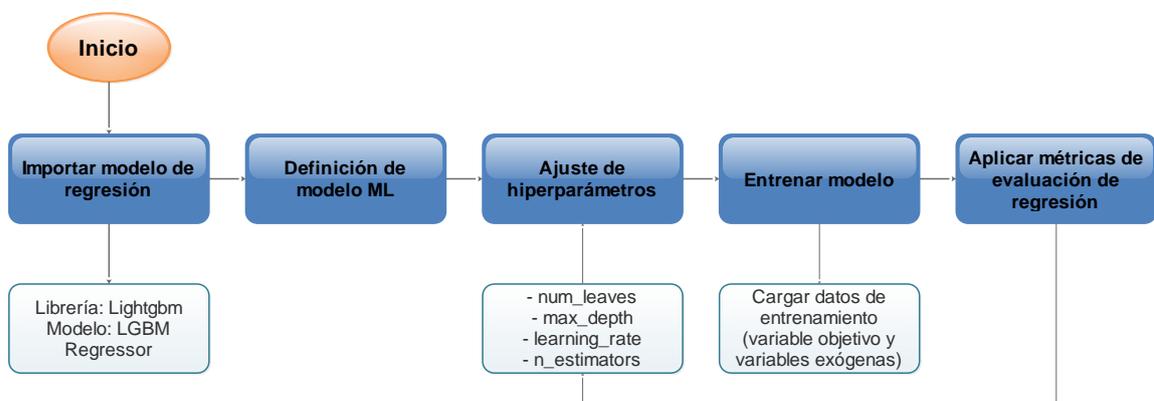


Figura 3.22. Metodología para modelo LGBM

Otra característica de LGBM es su capacidad de calcular el aporte de cada variable en relación con el pronóstico de la serie temporal, esto permitirá identificar las variables ambientales relevantes para el pronóstico de irradiación solar.

Para explotar el potencial del modelo, el ajuste de hiperparámetros se realizará mediante la herramienta Optuna que permite explorar eficientemente cada configuración hasta encontrar una combinación óptima en un espacio reducido de tiempo y aprovechando el recurso computacional.

Para la implementación de Optuna se empleará la metodología propuesta como se muestra en la Figura 3.23, el proceso inicia con la segmentación de los datos, como se había revisado previamente se requiere de la definición del modelo a entrenarse en cual será alimentado por los datos para entrenamiento hasta generar un modelo de salida y un resultado de dicho entrenamiento de acuerdo a una métrica, durante el desarrollo del proyecto se usará el error absoluto medio (MAE).

Posteriormente, el modelo resultante será alimentado con el conjunto de datos de validación y sus predicciones serán evaluadas de acuerdo a la métrica MAE, este resultado será tomado como referencia por Optuna que buscará minimizar esta evaluación en las siguientes iteraciones.

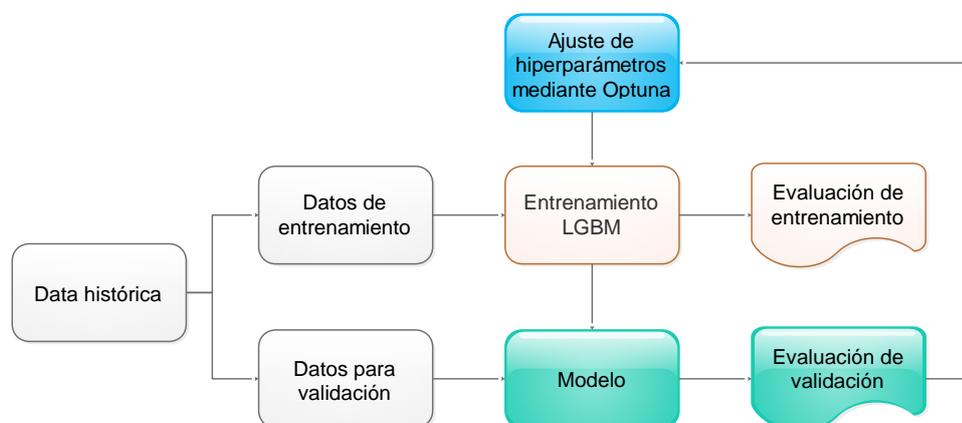


Figura 3.23. Metodología para implementación de Optuna

3.4 Prototipos de métodos y modelos para pronóstico de series temporales

3.4.1 Método Holt

Al hablar de radiación solar se puede asociar la cantidad de luz solar que recibe una ubicación determinada con factores como la hora del día, la época del año, ubicación geográfica, incidencia de nubosidad, entre otras variables que pueden agregar características como tendencia y estacionalidad que pueden ser analizadas a través de métodos estadísticos como Holt que permite analizar los valores más recientes y su capacidad de respuesta ante los cambios que se pueden presentar en la tendencia de una serie.

Para implementar el método Holt, es importante analizar las componentes de la serie temporal para determinar la viabilidad de aplicación del método dependiendo de la complejidad de los datos.

A modo de ejemplo, en la Figura 3.24 se presenta los registros de irradiancia solar de la estación M5021 - Yurafaccha Oyacachi que conforma el grupo de estaciones del DMQ y de forma general muestra el comportamiento de la luz solar durante los últimos meses del año 2021.

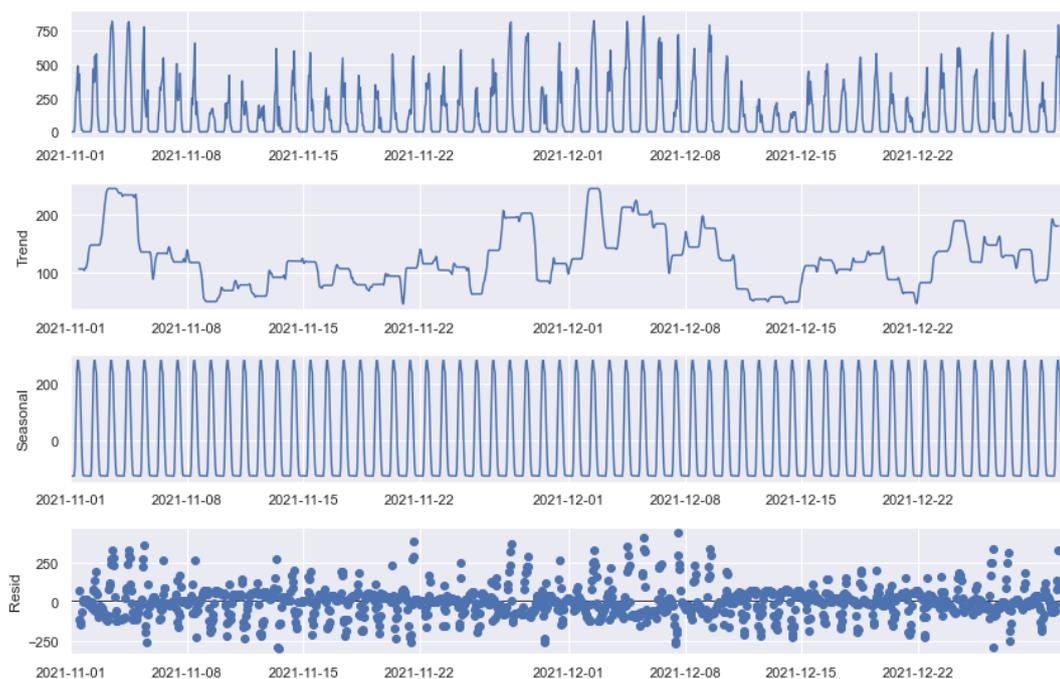


Figura 3.24. Componentes de radiación solar de la estación M5021 en 2021.

De acuerdo a los resultados mostrados en la Figura 3.24 se puede notar que el comportamiento irregular de la luz solar no mantiene una tendencia definida, no obstante, en los meses comprendidos entre diciembre a mayo se puede ver una clara reducción de la cantidad de luz solar debido al invierno y en junio a noviembre hay un ligero incremento de la luz solar debido a la temporada seca en la región sierra en la cual se encuentra ubicada la estación estudiada.

Con respecto a la estacionalidad, se puede visualizar un patrón constante en el tiempo asociado a la cantidad de luz en función de la hora del día, aunque la existencia de una gran cantidad de residuos indica que se puede extraer más información de la serie, sin embargo, un punto referencial para el pronóstico de series será el método Holt que se encuentra disponible en la librería *statsmodels*.

Previo a la aplicación del método estadístico, la cantidad de observaciones destinadas para entrenamiento y validación han sido segmentadas en una proporción de 80% y 20% de los datos, respectivamente.

Posteriormente, para entrenar el modelo se hará uso del método Holt y los parámetros representativos son “smoothing_level” y “smoothing_trend” que corresponden a las constantes de suavización α (Alpha) y β (Beta) en la ecuación de pronóstico de tiempo revisada en la sección “Línea base para predicción de series temporales – Holt Winters”, donde:

- α determinará la importancia de los datos en el tiempo, es decir, un α elevado dará énfasis a datos recientes, mientras que un α bajo dará más peso a los datos más antiguos.
- β será el parámetro asociado a la sensibilidad con respecto a la tendencia de la serie, es decir, un β elevado proporcionará una respuesta rápida a los cambios en la tendencia mientras que un β bajo suavizará la tendencia.

Para aplicar el método Holt de acuerdo a la metodología propuesta en la Figura 3.25 requiere del uso de la librería *statsmodels* para crear el modelo y especificar los parámetros de nivel de suavizado y tendencia que serán considerados durante el entrenamiento, finalmente el pronóstico será evaluado y en el caso de ser necesario se realizará la corrección de los hiperparámetros, esto permitirá ajustar apropiadamente los hiperparámetros para tomar una línea base sólida que será tomada como referencia para el resto de métodos y modelos que se tratarán en los apartados siguientes.

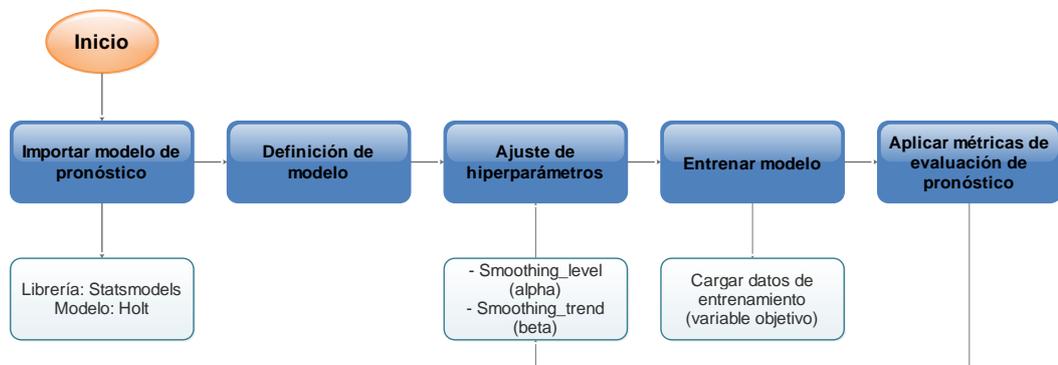


Figura 3.25. Metodología para pronóstico de radiación solar método Holt.

3.4.2 Método Holt-Winters

Las técnicas de pronóstico exponencial permiten obtener excelentes aproximaciones de las series de tiempo. En general, estas técnicas asignan un mayor peso a las observaciones recientes y un menor peso a las observaciones más antiguas. Como parte de las técnicas de suavizado exponencial uno de los principales referentes es Holt Winters al ser intuitivas, computacionalmente eficientes y se pueden generalizar a una amplia gama de series de tiempo.

La ventaja con respecto a otros métodos es su capacidad de adaptarse a medida que se va ingresando información nueva, otra característica es su aplicabilidad a datos que poseen una tendencia y estacionalidad definida, para el caso de irradiancia solar ayudará a considerar los ciclos de luz y sus cambios en el transcurso de los días.

El método de suavizado exponencial y su aplicación mostrada en la Figura 3.26 puede ser importado desde la librería *statsmodels* que permitirá la creación del modelo y la capacidad de modificar los hiperparámetros disponibles, para su ajuste se recomienda el empleo de una búsqueda de tipo *grid_search* hasta encontrar la combinación que proporcione el mejor rendimiento del método.

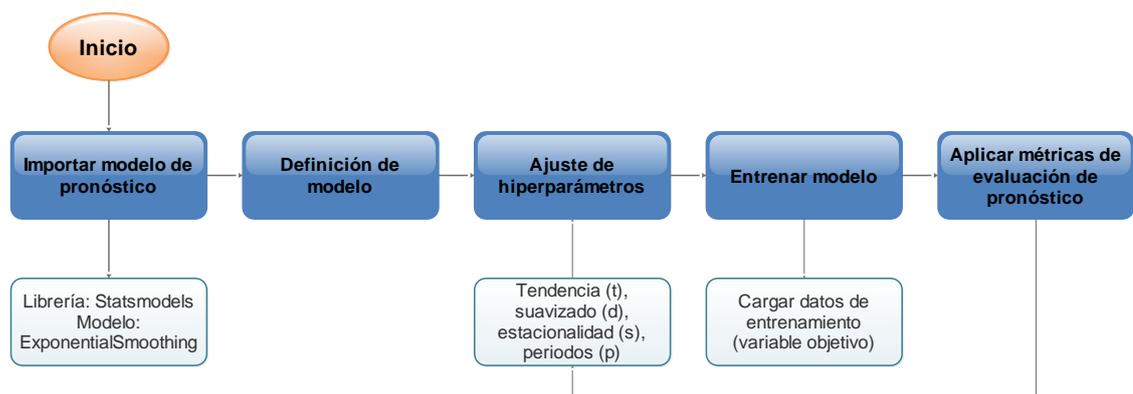


Figura 3.26. Metodología para pronóstico de radiación solar método Holt-Winters.

3.4.3 Redes LSTM one-step para series de tiempo univariadas

Del conjunto de datos generado a partir de los registros satelitales, se puede extraer una serie univariada que describe el comportamiento de la irradiancia solar en un conjunto de puntos previamente definido que abarca toda la provincia de Pichincha, de esta manera, se requiere generar pronósticos a futuro a partir de estos datos para ello se empleará las redes recurrentes LSTM debido a su capacidad para gestionar información relevante a corto y largo plazo.

Para realizar predicciones se empleará la metodología one-step que consiste en ajustar el modelo empleando todos los datos de entrenamiento y posteriormente, se generan predicciones individuales para cada paso de tiempo en el conjunto de pruebas, debido a su facilidad de implementación y capacidad de integrar un enfoque dinámico en base se adquieren más datos (Brownlee, 2017).

Del conjunto de datos se extraerá la serie de tiempo del punto coordinado de interés, teniendo en consideración que no se puede aplicar el modelo para múltiples ubicaciones de forma simultánea, para el entrenamiento de la red se

eliminarán los ciclos nocturnos que no aportan información relevante al modelo, por lo tanto, se considerarán únicamente los registros comprendidos desde las 08H00 hasta las 18H00.

Los datos deben ser procesados para mejorar el entrenamiento y su compatibilidad con el modelo LSTM, para ello se propone la siguiente metodología a implementar:

- 1) La serie de tiempo debe ser transformada a una serie estacionaria, es decir, no debe mantener ninguna característica de tendencia o estacionalidad de esta manera se facilita el modelado y el rendimiento del pronóstico se puede mejorar.
- 2) La estructura de los datos requeridos para entrenar el modelo LSTM debe incluir una entrada (X) y su respectiva salida (y), donde la entrada representa las observaciones del último paso de tiempo ($t - 1$) y la salida es el tiempo actual (t).
- 3) Los datos de la serie de tiempo serán segmentados siendo el 80% destinado para el entrenamiento del modelo, el 10% consecutivo será empleado para la validación y el 10% restante para el testeo, cabe destacar que la secuencia de los datos es importante al dividir el conjunto de datos.
- 4) Posteriormente, los datos deben ser escalados de acuerdo a la compatibilidad con la función de activación del modelo, para las redes LSTM se emplea por defecto la función tangente hiperbólica (\tanh) que se encuentra comprendida entre -1 y 1, los coeficientes para el escalado serán calculados únicamente en el conjunto de datos de entrenamiento y aplicados para la validación y testeo para evitar que el modelo conozca los rangos en los que se encuentran comprendidos los datos que no han sido observados por el modelo.
- 5) Una vez que se han preparado los datos, se desarrolla el modelo LSTM y ejecutan pruebas de ajuste de acuerdo a diversas configuraciones que se pueden implementar y entrenar el respectivo modelo.
- 6) Transformar las predicciones del modelo de acuerdo a las operaciones inversas de escalado y diferenciación para recuperar el estado original de los datos.

- 7) Generar pronóstico por cada paso de tiempo disponible en el conjunto de datos de testeo y agruparlos para poder evaluar el rendimiento del modelo con respecto a los datos reales de acuerdo a las métricas que se han establecido previamente para ser usadas en este proyecto.

Para tener una perspectiva más clara del proceso en general, la preparación de la serie de tiempo e implementación de una red LSTM a partir del procedimiento detallado previamente, en la Figura 3.27 se muestra la metodología propuesta para el desarrollo de un modelo de pronóstico y las librerías más relevantes para desarrollar cada etapa.

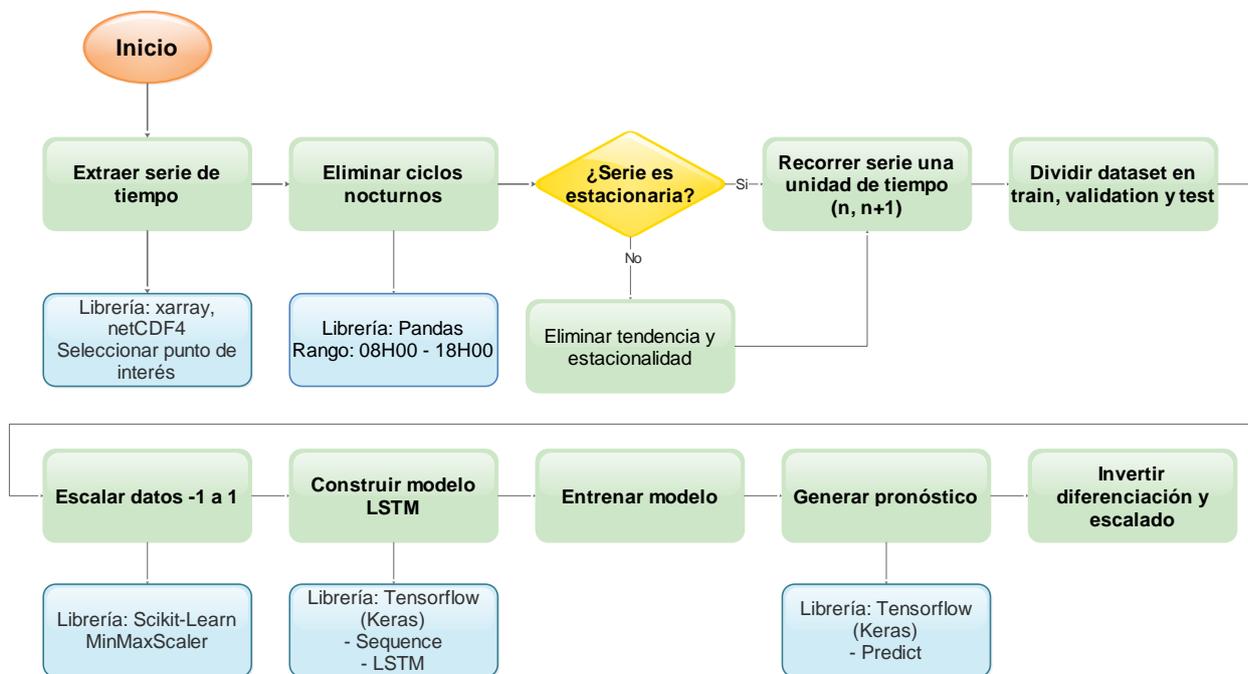


Figura 3.27. Metodología para entrenamiento de red LSTM

3.4.4 ConvLSTM

En la sección anterior se detalla la metodología a seguir para el pronóstico de irradiancia a partir de un punto seleccionado dentro del área de estudio, sin embargo, se puede analizar el comportamiento de la luz solar por áreas de tal manera que se pueda generar una predicción global del área de estudio.

Una vez que se ha creado el dataset artificial de irradiancia los datos se encuentran almacenados en un archivo netCDF, los datos pueden ser visualizados individualmente por puntos para obtener una serie de tiempo estándar y por áreas, donde se proporciona una matriz 3D que abarca toda el área de estudio durante el periodo de tiempo disponible en el dataset.

Para preparar los datos, se realiza una búsqueda previa de valores máximos y mínimos en toda la matriz de irradiancia con la finalidad de emplear esta información para normalizar todas las observaciones disponibles a través del método Min-Max y tener todas las medidas en la escala de 0 a 1, la ecuación empleada es:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.1)$$

Teniendo en cuenta que el modelo Conv-LSTM admite imágenes como entrada de datos, es necesario convertir el contenido de la matriz de irradiancia a imágenes, para ello se extrae cada paso de tiempo individualmente de forma secuencial y se aplica una conversión de la escala actual (0 a 1) hacia el rango de representación de imágenes (0 a 255), posteriormente, los resultados son guardados localmente como una imagen de un canal, el procedimiento a seguir se muestra en la Figura 3.28.

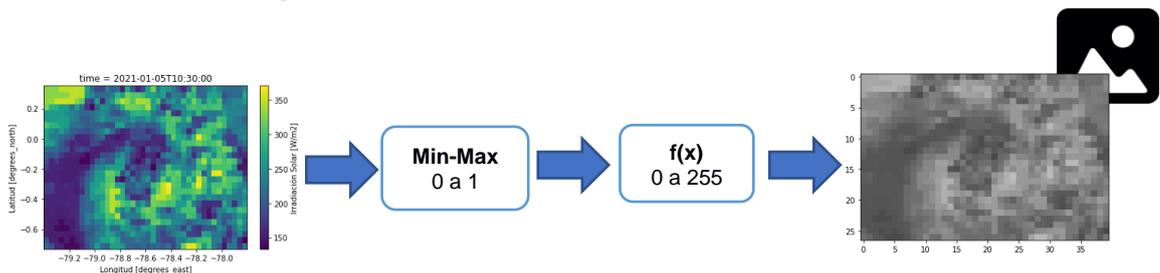


Figura 3.28. Preparación de registros de irradiancia solar.

Se ha considerado únicamente los periodos donde existe aporte solar superior a 50 W/m^2 , es decir, desde las 08H00 hasta las 17H00, de esta manera se eliminan los ciclos nocturnos y medidas de baja irradiancia que no aportan información, se

generaron 7227 imágenes de los cuales el 80% se empleará para entrenamiento, 10% para validación y 10% para realizar pruebas del modelo.

Para cargar los datos en el modelo ConvLSTM es necesario segmentar toda la secuencia de imágenes en lotes, debido a la cantidad de procesos requeridos y la limitación de hardware para el entrenamiento del modelo, cada lote comprenderá los datos de un día (frames) y apilados consecutivamente para conformar una secuencia, de esta manera se obtiene un objeto de tipo array estructurado como (número de secuencia, frames, alto de imagen, ancho de imagen).

Previo al entrenamiento del modelo los datos deben ser previamente procesados de tal forma que el objeto array debe añadirse una dimensión adicional para emular la estructura de una imagen en escala de grises, este objeto será dividido para asignar el conjunto de datos para entrenamiento y validación del modelo, los dataset resultantes deben ser normalizados a 1 y desplazados en un fotograma con respecto a los datos reales para especificar el margen de pronóstico requerido, la metodología propuesta se encuentra descrita en la Figura 3.29.

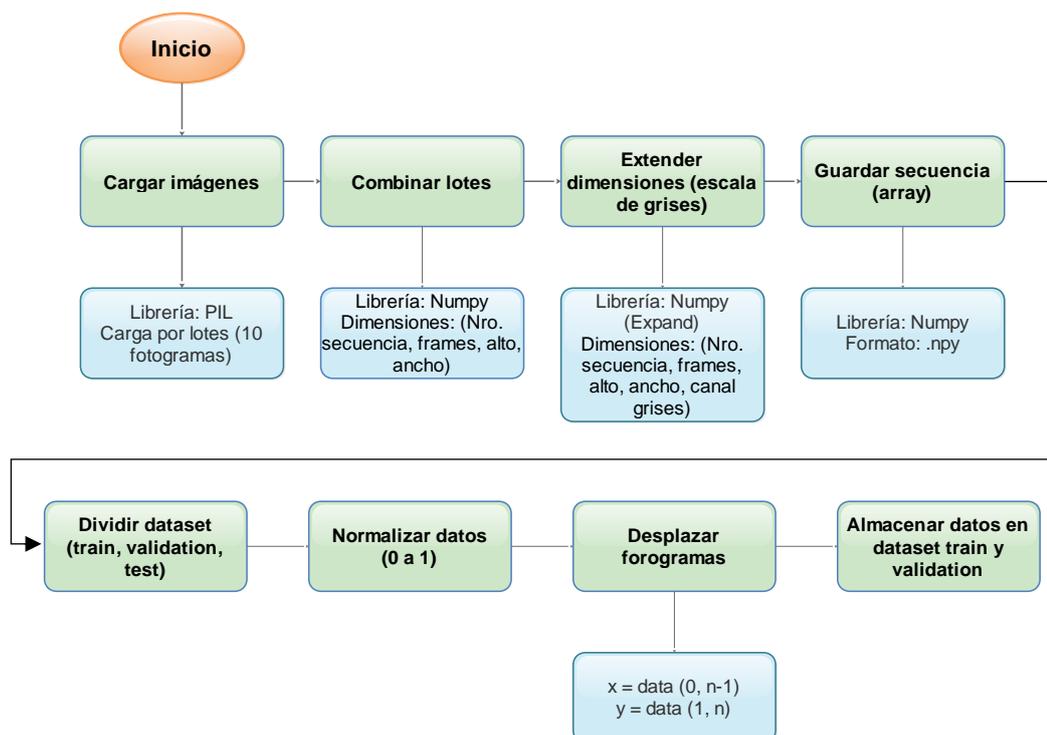


Figura 3.29. Preprocesamiento de secuencia de imágenes.

Una vez que los datos se encuentran adaptados correctamente se realizarán las pruebas para el entrenamiento del modelo ConvLSTM disponible en la librería *Keras* de *Tensorflow*, el tipo de modelo a implementar es *Sequential* en conjunto con convoluciones 2D y 3D, posteriormente se analizará su factibilidad como un método de pronóstico adecuado para el dimensionamiento de sistemas fotovoltaicos.

3.5 Infraestructura para el procesamiento y almacenamiento

Para el procesamiento de los datos y el entrenamiento de los modelos de inteligencia artificial se hará uso de los siguientes recursos descritos en la Tabla 3.6, teniendo en cuenta la gran cantidad de datos disponibles y el requerimiento de entrenamiento de modelos.

Tabla 3.6. Características técnicas de equipos y servidor

Equipo / Servidor	Características	Descripción
Computador portátil	Procesador: i7-9750H RAM: 16 GB Almacenamiento: 1 TB Tarjeta de video: Nvidia Gtx 1050 Ti 4 Gb	Descarga y limpieza de datos.
kobi.korigen.com (Google Collaboratory)	RAM: 8GB Disco: 160 Gb vCPUs: 4	Entrenamiento y almacenamiento de modelo de pronóstico

3.6 Diseño base de plataformas y prototipos de visualización

3.6.1 Pantalla de inicio y selección de ubicación

En esta vista mostrada en la Figura 3.30, el usuario tiene un panorama general de las herramientas disponibles del explorador que puede ser habilitado mediante el uso de botones de acuerdo al requerimiento y el nivel de detalle de la información que se puede obtener en cada una de ellas: calcular el ahorro en la cuenta de luz (propósito general) y calcular sistemas fotovoltaicos (avanzado).



Figura 3.30. Diseño propuesto de pantalla de inicio

Cada programa dentro del explorador solar requiere el acceso a una ubicación de referencia para poder realizar el cálculo respectivo o mostrar información relacionada a dicho lugar, para ello, el usuario será capaz de ingresar su ubicación mediante la búsqueda a través de un mapa en el cual pueda navegar aumentando o disminuyendo el nivel de detalle y seleccionar un punto de interés, mediante el ingreso manual de una coordenada mediante latitud, longitud o en su defecto, ingresar una cadena de texto con la dirección requerida como se muestra en la Figura 3.31.



Figura 3.31. Pantalla de definición de ubicación previa al inicio de cada programa.

3.6.2 Herramienta 1: Cálculo de ahorro eléctrico

Mediante el programa mostrado en la Figura 3.32, el usuario puede calcular de forma rápida el ahorro que podría obtener al instalar un sistema solar fotovoltaico teniendo en cuenta la superficie disponible en la ubicación previamente especificada y el consumo eléctrico que se puede obtener de las planillas de consumo eléctrico, de esta forma, la herramienta puede sugerir un estimado de la capacidad de la instalación, mostrar la potencia que se podría generar mensualmente con dicho sistema y a su vez, mostrar un pronóstico de potencia que se puede aprovechar a partir de la luz solar. Finalmente, mostrar el ahorro económico de acuerdo a la capacidad de generación estimada de forma anual.

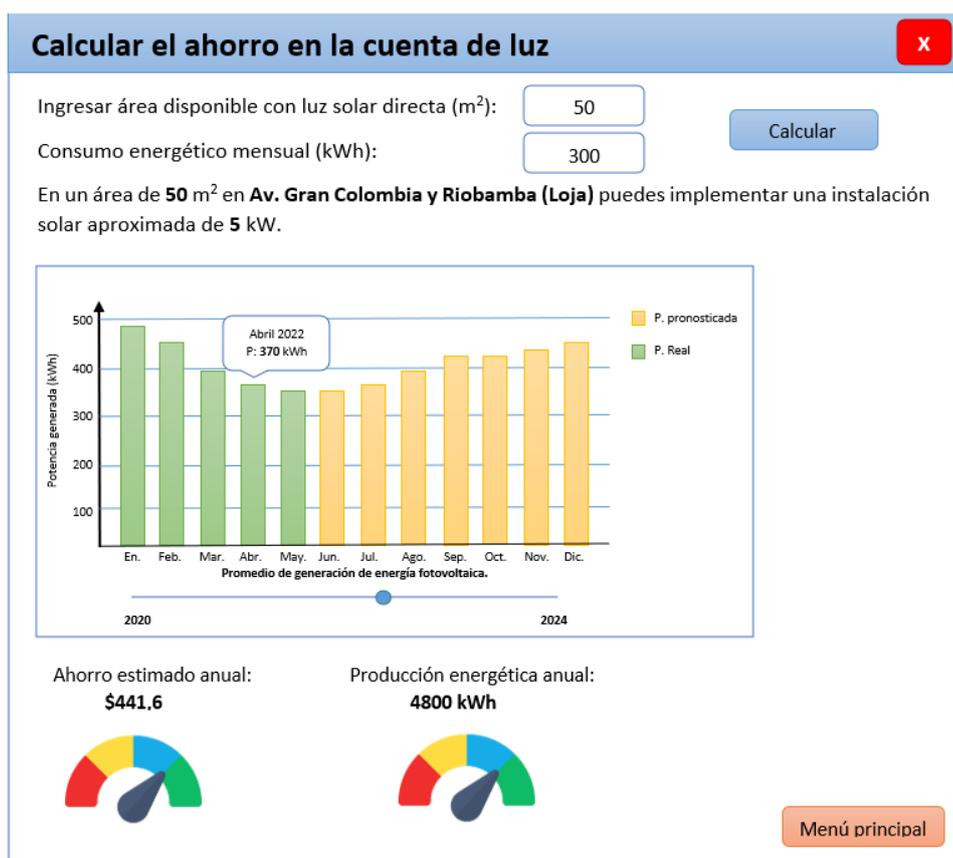


Figura 3.32. Pantalla para calcular el ahorro energético.

Para representar la potencia generada, se ha empleado el uso de un gráfico de barras para representar el promedio mensual teniendo en cuenta los valores basados en mediciones reales de irradiación real (marcados en verde) y los

valores estimados (marcados en amarillo) de esta forma poder mostrar la efectividad del sistema sugerido para satisfacer la demanda del usuario, además, existe la posibilidad de poder verificar valores estimados en función de los registros de radiación solar pasados y visualizar el comportamiento de la luz solar en diferentes años a través de un slider que comprende un rango desde el 2020 hasta el 2024 que correspondería a valores estimados y valor referencial que puede ser modificado durante el desarrollo del proyecto para asegurar siempre una alta confiabilidad de los valores pronosticados.

Para mostrar los beneficios del uso de sistemas solares, se ha introducido el concepto de ahorro económico que se puede calcular a partir de la potencia generada, teniendo en cuenta que la tarifa de consumo eléctrico a nivel nacional es aproximadamente 9,20 cUSD/kWh se puede transformar las unidades de potencia en su equivalente en dinero, de esta manera proporcionar una mejor perspectiva de la capacidad de auto sustentación energética esto mostrados mediante indicadores visuales que indican el nivel de ahorro que corresponde con la producción energética.

3.6.3 Herramienta 2: Calcular sistemas fotovoltaicos

A diferencia de la herramienta de cálculo de ahorro que muestra de forma general la eficiencia de un sistema solar, la opción de calculadora de paneles solares permite tener en consideración aspectos técnicos avanzados que son necesarios para el diseño de un sistema fotovoltaico teniendo en cuenta el comportamiento de la luz solar en una ubicación determinada en diferentes intervalos de tiempo, apoyándose de valores de irradiación que se mantengan constantemente actualizados y disponiendo de las predicciones estimadas por la herramienta para apoyar a la toma de decisiones y determinar la viabilidad del proyecto.

Para esta sección, los datos más importantes que se requieren conocer se representan mediante diferentes gráficas como se muestra en la Figura 3.33, la primera sección corresponde a un registro de la cantidad de irradiación anual representado mediante un gráfico de líneas para facilitar la visualización de

patrones de tendencia, el rango de visualización puede ser controlado por un slider para incrementar o disminuir el nivel de detalle del gráfico.

La gráfica de líneas tiene la capacidad de seleccionar un punto en concreto referente a un año de interés que puede interactuar con el resto de gráficas y actualizar las vistas de acuerdo al año de interés (por defecto el año actual), en la siguiente sección se puede mostrar el promedio de irradiación solar de acuerdo a los registros correspondientes a los meses que conforman el año de interés previamente seleccionado, además de proporcionar una vista del promedio mensual así como la menor irradiación registrada en dicho año, este dato es de vital importancia ya que permite establecer el aporte solar mínimo que el sistema fotovoltaico debe usar para satisfacer la demanda del usuario. Dentro de la gráfica se ha implementado un botón de recarga que permite volver a la vista por defecto de la herramienta.

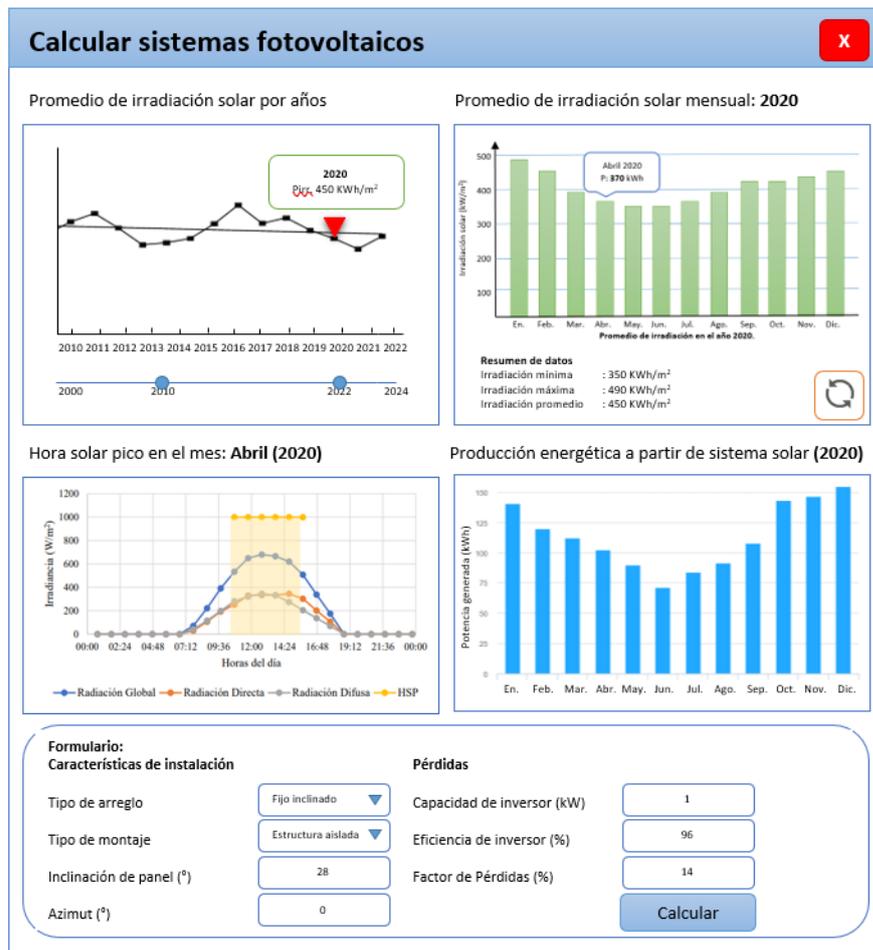


Figura 3.33. Pantalla para el cálculo de sistemas fotovoltaicos

Esta sección permite visualizar y seleccionar los promedios mensuales y mantiene una interacción con la gráfica correspondiente a las horas solares pico, que muestra la cantidad de luz solar efectiva que puede utilizar un panel solar para obtener su máximo rendimiento además de visualizar el comportamiento de la luz solar en el transcurso del día.

Finalmente, en la vista inferior se implementa un panel que permite el ingreso de información de un posible arreglo de paneles e inversores para determinar la capacidad energética estimada de acuerdo a los datos de irradiación y que se encuentra representado en la última sección, de forma similar a la gráfica mostrada en la herramienta de cálculo de ahorro con la ventaja de poder interactuar con los datos registrados en los diferentes años que comprende el dataset.

CAPÍTULO 4

4. ANÁLISIS DE RESULTADOS

4.1 Recolección de datos y estrategias para validación del proyecto

En los apartados anteriores se han podido definir las diversas fuentes de datos que se emplearán para el desarrollo del proyecto y su rol dentro de las diferentes etapas del mismo, siendo los componentes preliminares para la construcción de un dataset artificial que permita abarcar toda el área de estudio adoptando características como la continuidad temporal disponible desde los datos satelitales y la precisión en las medidas de irradiancia solar proporcionado por estaciones meteorológicas.

Se ha podido verificar las deficiencias existentes en diferentes estaciones a nivel nacional, lo que ha reducido el área de estudio a la provincia de Pichincha debido a la existencia de estaciones correspondientes al DMQ que abarcan gran parte del territorio de la provincia y mantienen protocolos de mantenimiento periódicos asegurando la confiabilidad de los datos disponibles.

Otro factor a tener en cuenta que ha surgido durante el desarrollo de la metodología para la descarga de datos satelitales, corresponde a las limitaciones propias de la API del repositorio NREL que dificulta la descarga de información para superficies extensas debido al tamaño de los ficheros y el tiempo requerido para su descarga, por tal motivo se considerará únicamente el año 2021.

Cabe recalcar que en la mayoría de estaciones en tierra se dispone la información completa de todos los meses para el año especificado, lo que permitirá estudiar datos relativamente actualizados y comparar los pronósticos generados con los datos correspondientes al año 2022.

4.1.1 Análisis de datos de estaciones meteorológicas

Teniendo en cuenta que cada estación meteorológica guarda los registros de irradiancia solar de forma horaria, para el año 2021 idealmente se estima tener 8.760 registros en cada estación, sin embargo, como se ha podido anticipar

muchas de las estaciones no satisfacen esta condición debido a que cada sensor se encuentra sometido a diversos factores que impiden su correcto funcionamiento, como se puede observar en la Figura 4.1 de las 26 estaciones consideradas apenas 8 tienen sus registros completos mientras que las 18 restantes presentan registros vacíos en el mismo intervalo de tiempo.

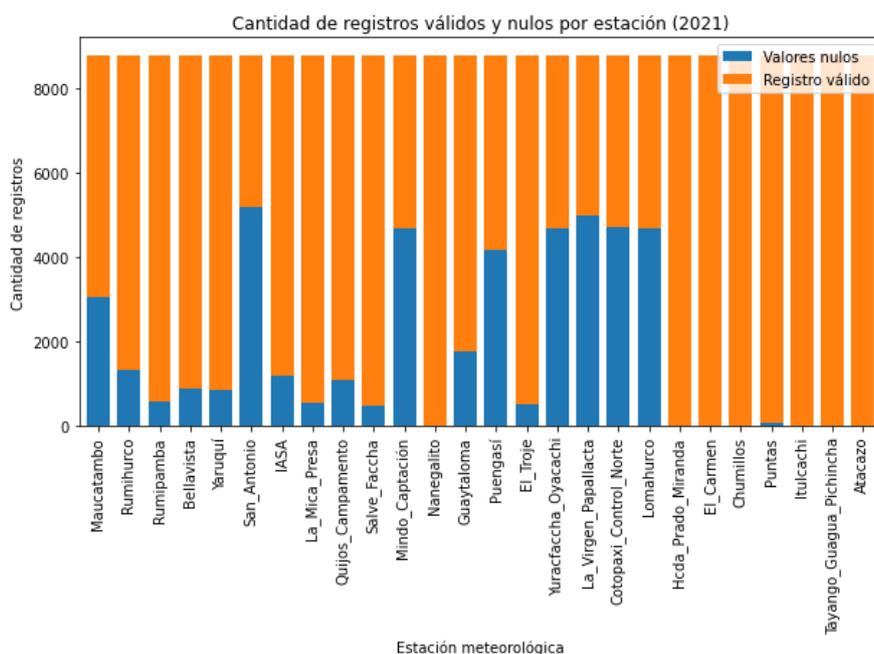


Figura 4.1. Revisión de registros nulos y válidos correspondientes a estaciones meteorológicas del DMQ.

Para tener una perspectiva más detallada de los datos disponibles por estación, la Tabla 4.1 muestra un resumen descriptivo de los registros de cada estación partiendo de la cantidad de registros disponibles ordenado de mayor a menor, el porcentaje que representa la cantidad de registros con respecto al total esperado para el año 2021, valores mínimos, máximos, desviación estándar y cuartiles.

Tabla 4.1. Resumen descriptivo de datos disponibles por cada estación meteorológica.

Estación	Cantidad	Porcentaje	Promedio	Desv. Std.	Mínimo	0,25	0,5	0,75	Máximo
Hcda Prado Miranda	8760,00	100,00	155,13	243,17	0,00	0,00	7,30	244,72	1196,25
Chumillos	8760,00	100,00	165,56	245,75	0,00	0,00	7,70	280,18	1195,22

Itulcachi	8760,00	100,00	127,00	188,13	0,00	0,00	7,02	215,46	909,44
Tayango Guagua Pichincha	8760,00	100,00	126,10	187,38	0,00	0,00	7,29	209,03	1027,54
Nanegalito	8759,00	99,99	132,44	181,93	2,10	16,45	39,98	189,56	1321,13
Atacazo	8759,00	99,99	156,00	237,36	0,00	0,00	8,80	251,86	1199,35
El Carmen	8756,00	99,95	188,78	277,03	0,00	0,00	9,56	328,23	1291,93
Puntas	8701,00	99,33	188,71	273,39	0,00	0,00	10,45	326,83	1147,32
Salve Faccha	8296,00	94,70	119,01	177,79	0,00	0,00	5,50	209,08	1231,08
El Troje	8240,00	94,06	183,60	278,32	1,56	2,00	7,17	296,75	1178,09
La Mica Presa	8207,00	93,69	174,83	262,71	0,00	0,00	7,42	292,55	1240,30
Rumipamba	8184,00	93,42	133,10	233,17	0,00	1,83	2,17	166,23	1183,91
Yaruquí	7902,00	90,21	193,63	276,59	0,00	0,00	7,70	359,01	1122,45
Bellavista	7867,00	89,81	167,83	254,61	0,00	0,00	6,00	272,63	1119,50
Quijos Campamento	7656,00	87,40	69,01	111,67	0,00	0,00	2,73	104,05	724,20
IASA	7567,00	86,38	165,98	256,81	0,00	0,00	5,18	272,41	1153,08
Rumihurco	7419,00	84,69	164,03	247,34	0,00	0,00	8,05	259,66	1140,17
Guaytaloma	6996,00	79,86	113,77	173,47	0,00	0,00	5,52	197,41	1182,05
Maucatambo	5701,00	65,08	164,15	248,56	0,00	0,00	6,44	269,34	1197,63
Puengasí	4591,00	52,41	193,50	282,50	1,61	2,00	8,17	339,03	1169,67
Mindo Captación	4091,00	46,70	108,07	193,48	0,00	0,00	4,90	136,28	1152,79
Yuracfaccha Oyacachi	4087,00	46,66	147,83	232,18	0,00	0,00	4,90	232,87	1225,41
Lomahurco	4068,00	46,44	157,92	245,88	-1,00	0,00	5,50	245,19	1132,25
Cotopaxi Control Norte	4042,00	46,14	184,42	277,21	-1,00	0,00	8,04	305,21	1209,83
La Virgen Papallacta	3796,00	43,33	139,24	220,09	0,00	0,00	5,29	218,72	1226,53
San Antonio	3584,00	40,91	193,91	291,32	0,00	0,00	5,33	326,44	1203,42

De los datos presentados, se puede verificar que las primeras 18 estaciones disponen más del 79% de datos disponibles para ser analizados, mientras que las estaciones “Mindo Captación”, “Yuracfaccha Oyacachi”, “Lomahurco”, “Cotopaxi Control Norte”, “La Virgen Papallacta” y “San Antonio” tienen disponible alrededor del 40% de los datos esperados durante el año lo que puede dificultar la creación de un modelo de regresión a partir de dichas estaciones.

Otra consideración a tener en cuenta corresponde a los registros de irradiancia en torno a los ciclos nocturnos donde la ausencia de luz solar bajo condiciones normales sugiere una medición igual a cero, no obstante, cinco estaciones

presentan un mínimo diferente a este valor lo que puede ser indicio de una posible descalibración del sensor durante el periodo de estudio, las estaciones que presentan esta condición son: “Lomahurco”, “Cotopaxi Control Norte”, “El Troje”, “Puengasi” y “Nanegalito”.

Para tener una perspectiva más amplia de la distribución de los datos se elabora un diagrama de caja del nivel de irradiancia solar captada por las 26 estaciones correspondientes al DMQ como se muestra en la Figura 4.2, donde se puede notar la clara tendencia de los datos a mantenerse cercanos a cero esto debido a los ciclos nocturnos que abarcan del 46% hasta el 50% de los datos por cada estación.

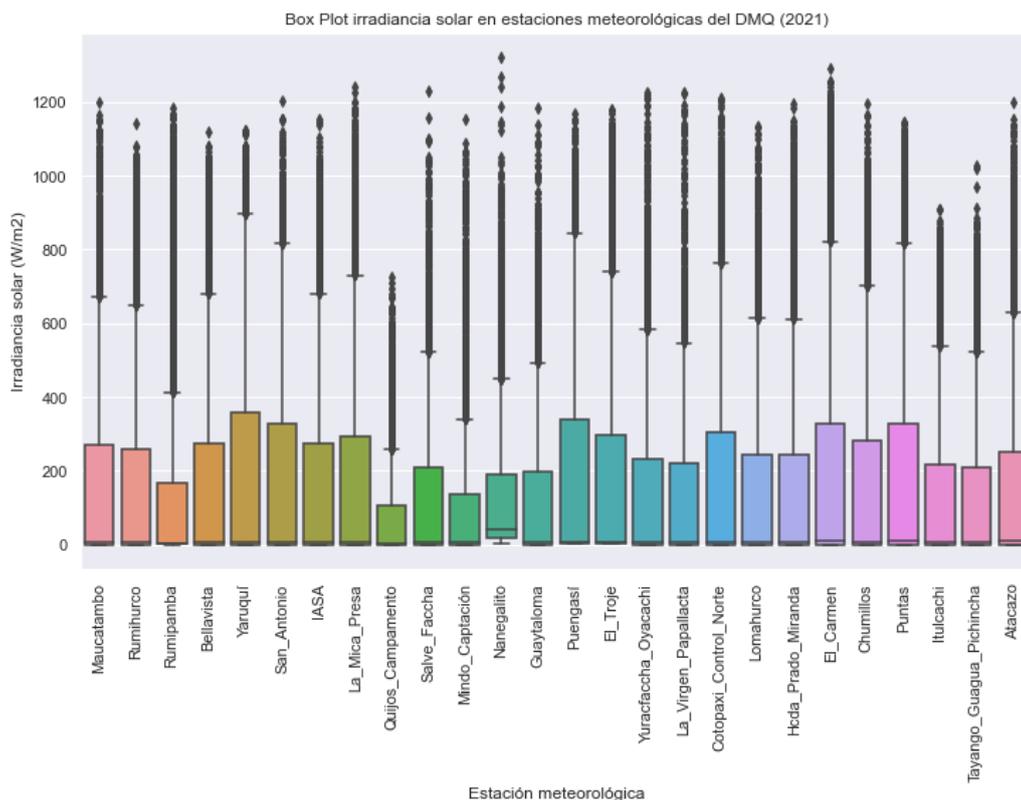


Figura 4.2. Box Plot para irradiancia solar en estaciones meteorológicas del DMQ.

Referente a las estaciones “Cotopaxi Control Norte” y “Lomahurco” presentan registros inferiores a cero que abarcan el 8.31% y 28.88% de la totalidad de los datos respectivamente, en la Figura 4.3 se puede visualizar el comportamiento de

la luz solar por horas del día y la línea que marca la irradiancia promedio donde se descarta un mal funcionamiento del sensor.

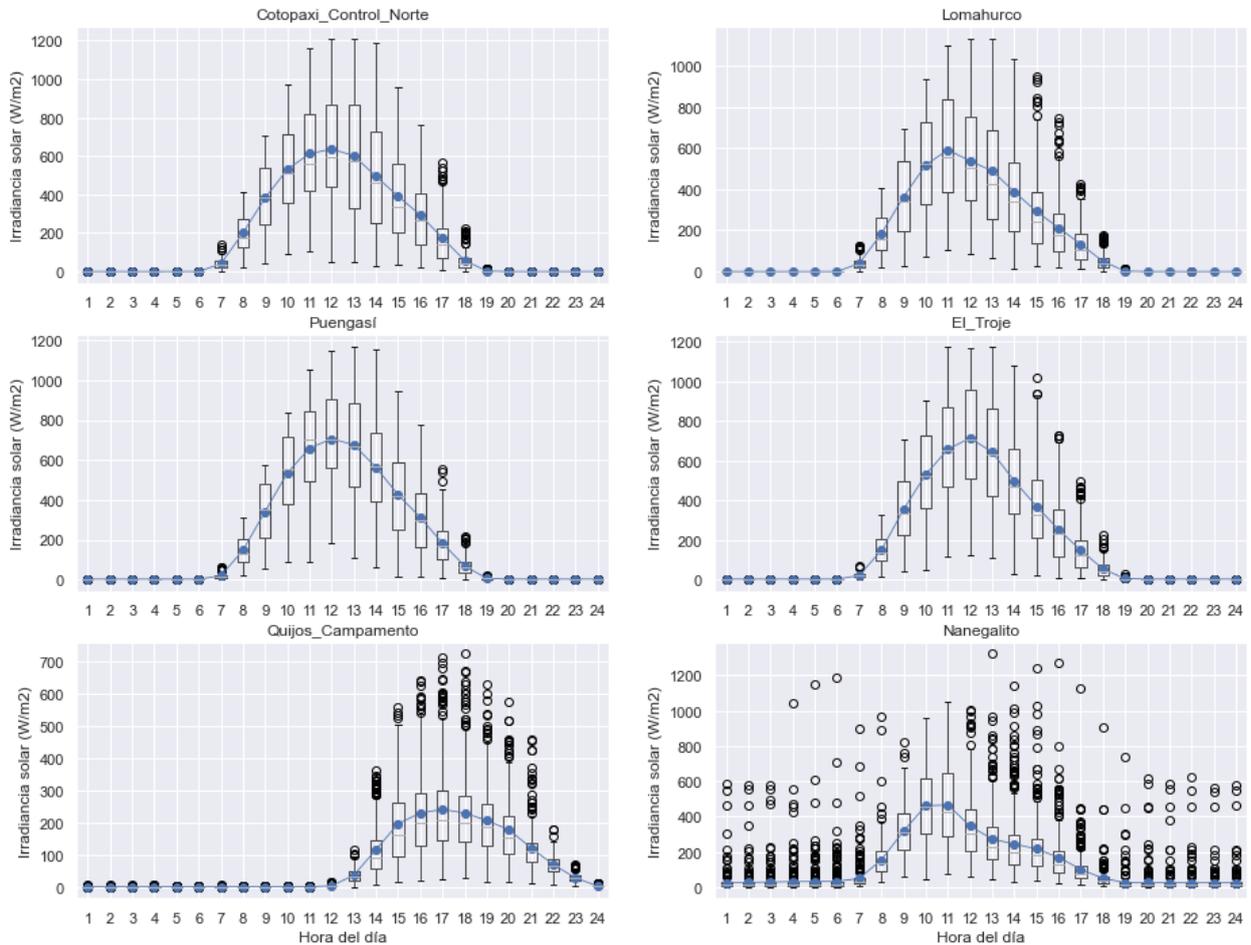


Figura 4.3. Análisis de diagrama de cajas para irradiancia solar por horas.

Las estaciones “Puengasí” y “El Troje” se encuentran desfasadas debido a que sus registros nocturnos no corresponden a una irradiancia igual a cero, este desfase abarca el 45.96% y 46.38% los datos respectivamente, no obstante, de acuerdo a la Figura 4.3 se puede visualizar una repuesta adecuada en los horarios donde se espera una mayor cantidad de luz solar.

Durante el análisis por horas, se pudo verificar otro tipo de error que no se puede detectar mediante los diagramas de caja mostrados en la Figura 4.2 correspondiente al desfase horario, siendo en la estación “Quijos Campamento” donde se puede notar un error entre los registros de irradiancia y la hora.

Finalmente, la estación “Nanegalito” presenta inconsistencias en sus datos que se pueden evidenciar en los diagramas de caja de la Figura 4.2 donde se puede notar que la mediana de los datos registrados es 39.97 W/m^2 que no corresponde a la tendencia revisada en el resto de equipos y el análisis por hora mostrado en la Figura 4.3 indica la poca veracidad de los datos debido a condiciones inherentes al funcionamiento del sensor de la estación.

Para el desarrollo del proyecto no se emplearán los datos de las estaciones “Nanegalito” (C15) y “Quijos Campamento” (C12).

4.1.2 Tratamiento de valores atípicos en datos de estaciones meteorológicas

Una vez revisados los datos disponibles de las estaciones en tierra y sus estadísticas, en los diagramas de caja presentados en la Figura 4.2 se han podido observar datos que se encuentran fuera de la distribución esperada y marcados como valores atípicos, sin embargo, no pueden ser eliminados debido a la amplia cantidad de registros marcados por estación, se puede observar que las mediciones se encuentran cercanos al margen superior de 1.2 KW/m^2 en la mayoría de las estaciones, indicando el funcionamiento correcto de las estaciones con excepción a las unidades que han presentado incidencias y posteriormente han sido descartadas.

De acuerdo a Vaca & Ordóñez (2020) en el “*Mapa solar del Ecuador 2019*” se estima que para la provincia de Pichincha en promedio se recibe aproximadamente entre 3.9 a 4.8 KWh/m^2 lo que indica la capacidad energética disponible en la provincia, teniendo en cuenta este valor, se determina la cantidad de irradiación solar a partir del cálculo del área bajo la curva de las gráficas de irradiancia por horas de las estaciones meteorológicas cuyo resultado se muestra en la Figura 4.4, obteniendo un promedio general de $3,82 \text{ KWh/m}^2$ asegurando que los datos captados se encuentra acorde al potencial solar esperado en la zona.

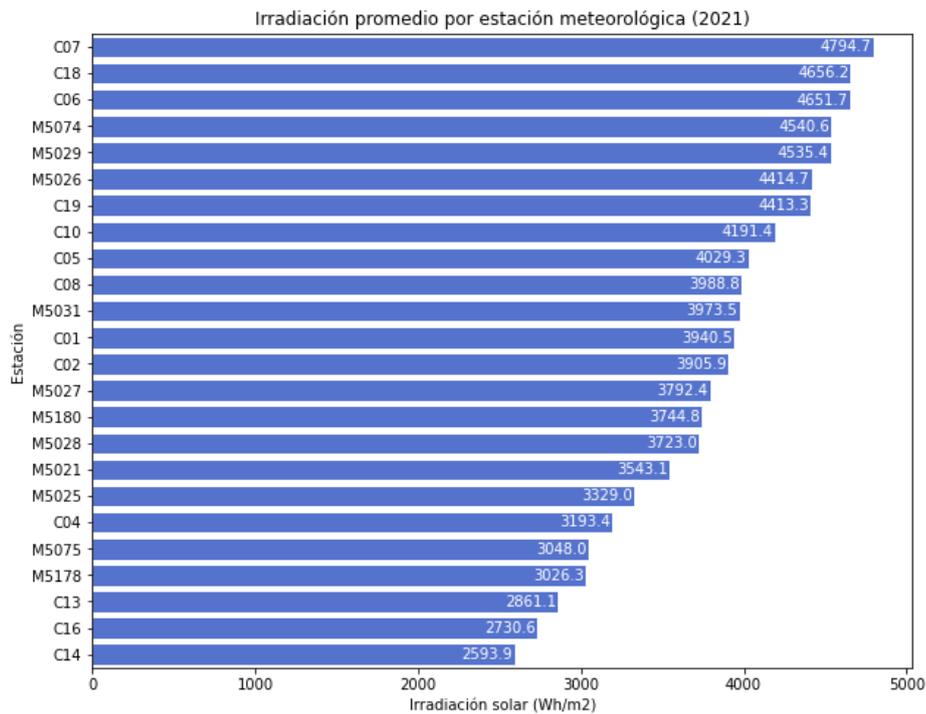


Figura 4.4. Análisis de irradiación solar por estación.

Para realizar la detección de valores atípicos dentro de las series de tiempo correspondiente a la radiación solar por cada estación, se implementó el método *Isolation Forest* basado en árboles de decisión y regresión (CART), al ser un método de detección no supervisado utiliza como premisa la proporción de anomalías que pueden estar presentes en el dataset y las selecciona directamente mediante evaluaciones de prueba y error (Bajaj, 2022).

El método de detección incorpora identificadores para valores normales (1) o anómalos (-1), facilitando las tareas de análisis y corrección de outliers presentes en el dataset, en la Figura 4.5 se muestra un ejemplo de la detección realizada en los datos correspondientes a la estación El Carmen (M5029) utilizando un factor de 1% donde se pudieron identificar 88 valores atípicos, este análisis y correcciones posteriores deben ser replicados para el resto de estaciones.

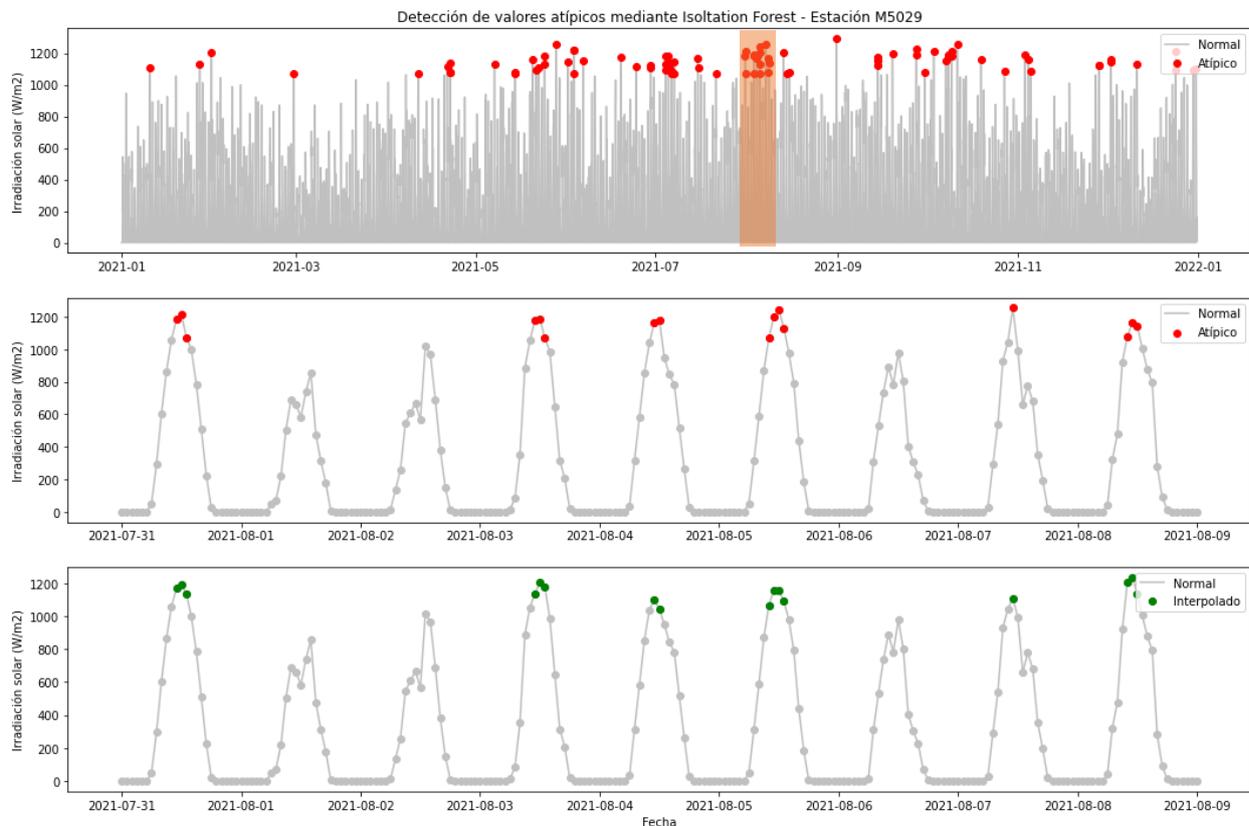


Figura 4.5. Detección de outliers en datos de radiación de la estación “El Carmen”

Se puede destacar que el método se enfoca en las mediciones cercanas a los valores máximos registrados dentro de la serie, para el ejemplo mostrado es cercano a 1.2 KW/m^2 siendo pocos los registros que cumplen esta condición y requieren ser tratados, no obstante, debido a la resolución temporal por horas dificulta el análisis de los datos al no poder visualizar a detalle la evolución de la radiación solar especialmente en los cambios bruscos de irradiancia en intervalos pequeños de tiempo.

De los resultados obtenidos por el detector de outliers, se puede notar que los puntos considerados representan registros elevados de irradiancia a nivel general dentro de la serie de tiempo, teniendo en cuenta esta característica se empleó una interpolación de outliers de tipo *spline* con polinomios de segundo grado para ajustar los datos y emular el comportamiento natural de la radiación solar sin alterar drásticamente los picos de irradiancia que posteriormente deben ser pronosticados.

De los resultados obtenidos tras la detección de outliers por cada estación, se puede observar en la Figura 4.6 la cantidad de registros que fueron identificados como atípicos, teniendo en cuenta que las series analizadas disponen de 3584 hasta 8755 elementos, las estaciones que tienen mayor incidencia de outliers son: El Carmen (M5029), Tayango Guagua Pichincha (M5178), Itulcachi (M5075), Puntas (M5074), Hacienda Prado Miranda (M5028) y Atacazo (M5180).

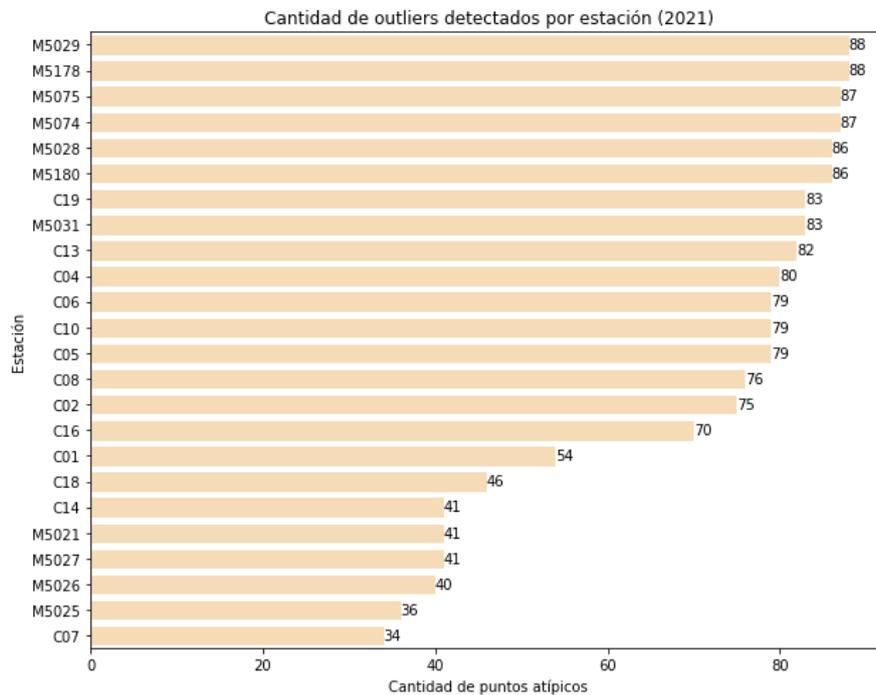


Figura 4.6. Registro de outliers captados por estación.

Existe una amplia variedad de alternativas para el tratamiento de valores atípicos dentro de series de tiempo, siendo el suavizado una alternativa a considerar para mitigar el ruido en datos provenientes de sensores, para los datos de radiación disponibles la atenuación de la señal reduce considerablemente además de agregar variaciones en los ciclos nocturnos con valores negativos como se muestra en la Figura 4.7 que pueden afectar el rendimiento de los modelos a emplearse para la creación del dataset artificial.

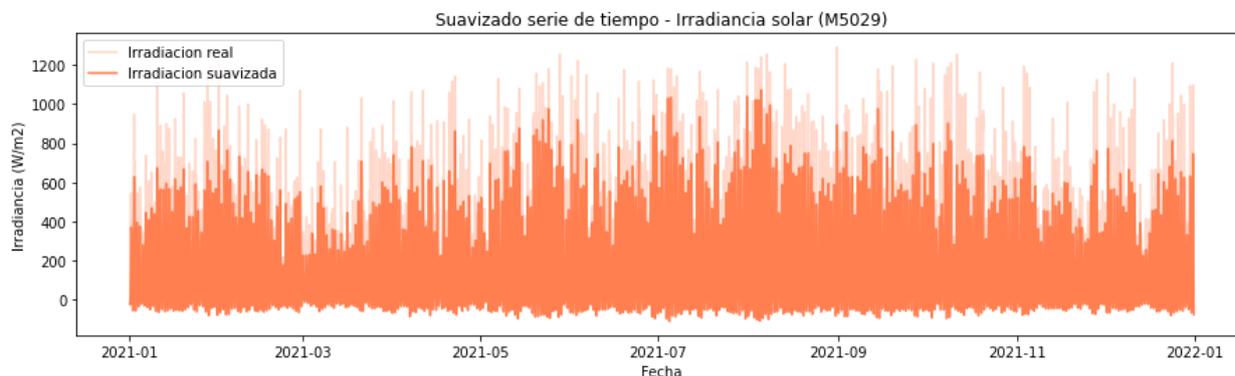


Figura 4.7. Suavizado outliers en serie de tiempo de irradiancia solar.

4.1.3 Análisis de datos satelitales

Durante la sección 3.1.5. *Análisis descriptivo de datos satelitales* se ha definido el tipo de formato que será empleado para almacenar datos ambientales en función de las coordenadas geográficas asociadas al área de estudio, por tal motivo, es necesario realizar un análisis de las variables ambientales disponibles que han sido obtenidas desde el repositorio de datos satelitales NREL.

El dataset satelital consta de 1080 puntos de referencia cuyo historial de cambios considerado en este proyecto se mantiene desde 01/01/2021 hasta 31/12/2021 segmentados en intervalos de 30 minutos, para facilitar la compatibilidad entre datos satelitales y locales se han agrupado todos los datos de forma horaria.

Para tener una perspectiva general de la calidad de los datos satelitales correspondiente a todas las variables ambientales consideradas en torno a la provincia de Pichincha, se ha seleccionado el punto central dentro de la superficie de estudio para su respectivo análisis, teniendo en cuenta que los datos disponibles han sido obtenidos de un único sensor satelital (GOES East) no existen irregularidades en las medidas de los 1080 puntos establecidos previamente.

Se utilizará el punto con coordenadas -0.19, -78.5 (latitud, longitud) ubicado en la provincia de Pichincha como referencia para el análisis exploratorio de los datos satelitales, cada punto geográfico dispone de 9 variables ambientales

representadas como una serie temporal de 17.509 elementos, la Tabla 4.2 muestra el resumen estadístico para el punto considerado.

Tabla 4.2. Resumen descriptivo de datos satelitales en el punto (-0.19, -78.5).

Variable	Cantidad	Promedio	Desv. Std.	Mínimo	25%	50%	75%	Máximo
DHI	17509	102,84	147,24	0,00	0,00	5,00	171,00	609,00
DNI	17509	151,70	267,78	0,00	0,00	0,00	180,00	1054,00
GHI	17509	205,70	291,62	0,00	0,00	5,00	377,00	1098,00
Temperatura	17509	10,22	3,17	2,20	7,70	9,40	13,00	18,30
Presión	17509	707,33	1,26	704,00	706,00	707,00	708,00	711,00
Humedad Relativa	17509	92,92	11,23	49,10	86,90	100,00	100,00	100,00
Dirección del viento	17509	248,25	94,52	0,00	177,00	295,00	314,00	360,00
Velocidad del viento	17509	1,73	1,07	0,20	0,80	1,40	2,70	4,90
Precipitación	17509	1,99	0,22	1,10	1,90	2,00	2,10	2,50

Como se puede observar, los registros satelitales se encuentran completos a diferencia de los datos locales no presentan ningún valor vacío y cada variable ambiental presenta un comportamiento diferente entre sí.

Para facilitar la comprensión de los datos mostrados en el resumen estadístico, en la Figura 4.8 se presenta un gráfico de violín con la finalidad de observar la densidad de los datos en conjunto con las estadísticas descriptivas representadas de forma gráfica con respecto a cada variable ambiental.

Se puede notar que las componentes de irradiancia solar (directa, normal y global) presentan un comportamiento similar a los valores de radiación medidos por estaciones de tipo meteorológico indicando un claro sesgo alrededor del tercer cuartil, esto se debe a la acumulación de registros alrededor a cero obtenidos de ciclos nocturnos, el alba y el ocaso donde el nivel de luz solar tiende a ser bajo.

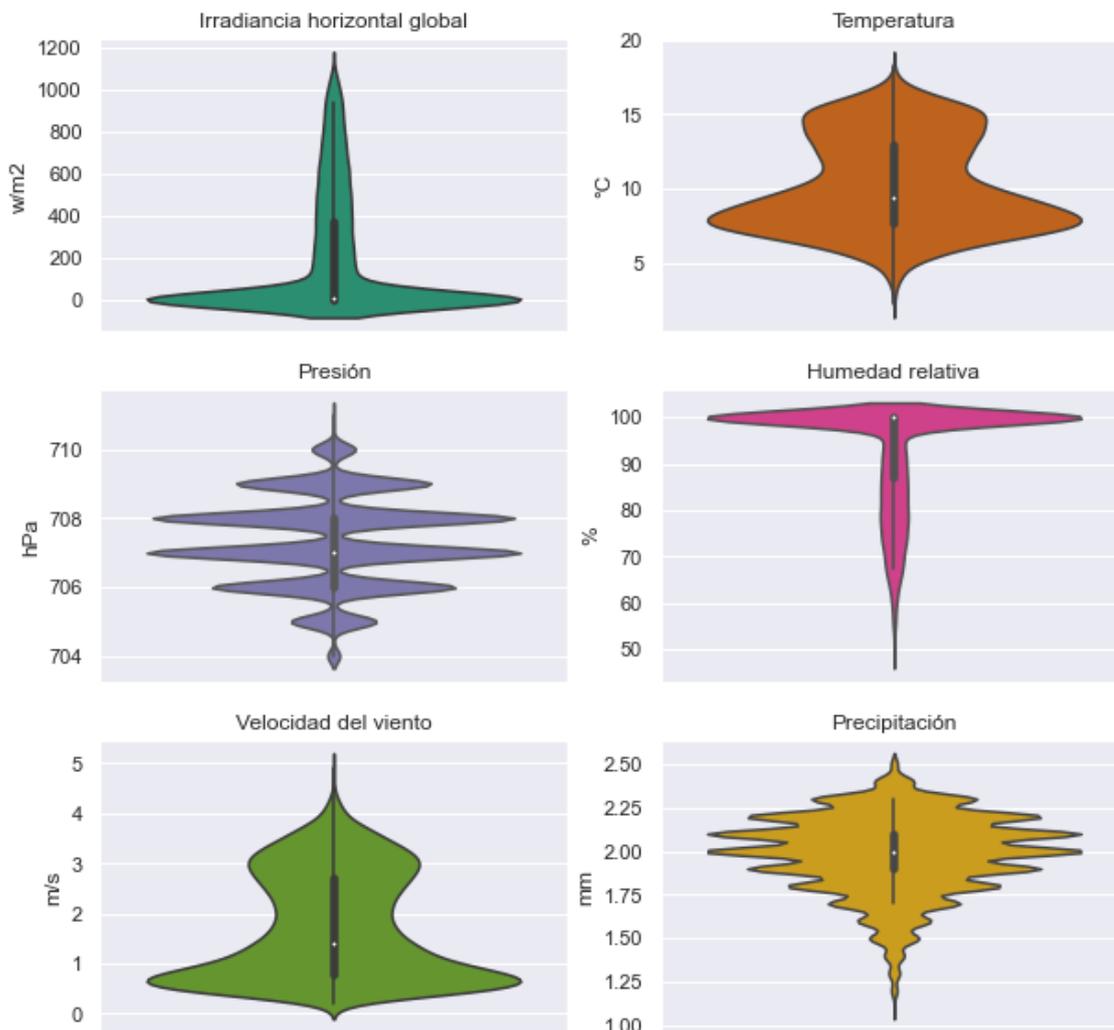


Figura 4.8. Gráfico de violín para variables ambientales de tipo satelital.

La variable de temperatura por otra parte, se ha registrado variaciones desde 2°C hasta 18°C y una acumulación de registros alrededor de los 9.4°C, teniendo en cuenta que se trata de una zona alta dentro de la sierra los niveles de temperatura son normales.

La variable presión atmosférica presenta una serie de variaciones en el rango comprendido entre 704 hPa y 711 hPa, sin embargo, este comportamiento se debe a la baja resolución de las mediciones al considerarse únicamente valores enteros dentro del dataset, en la Figura 4.9 se muestra a detalle la distribución de los datos para cada variable.

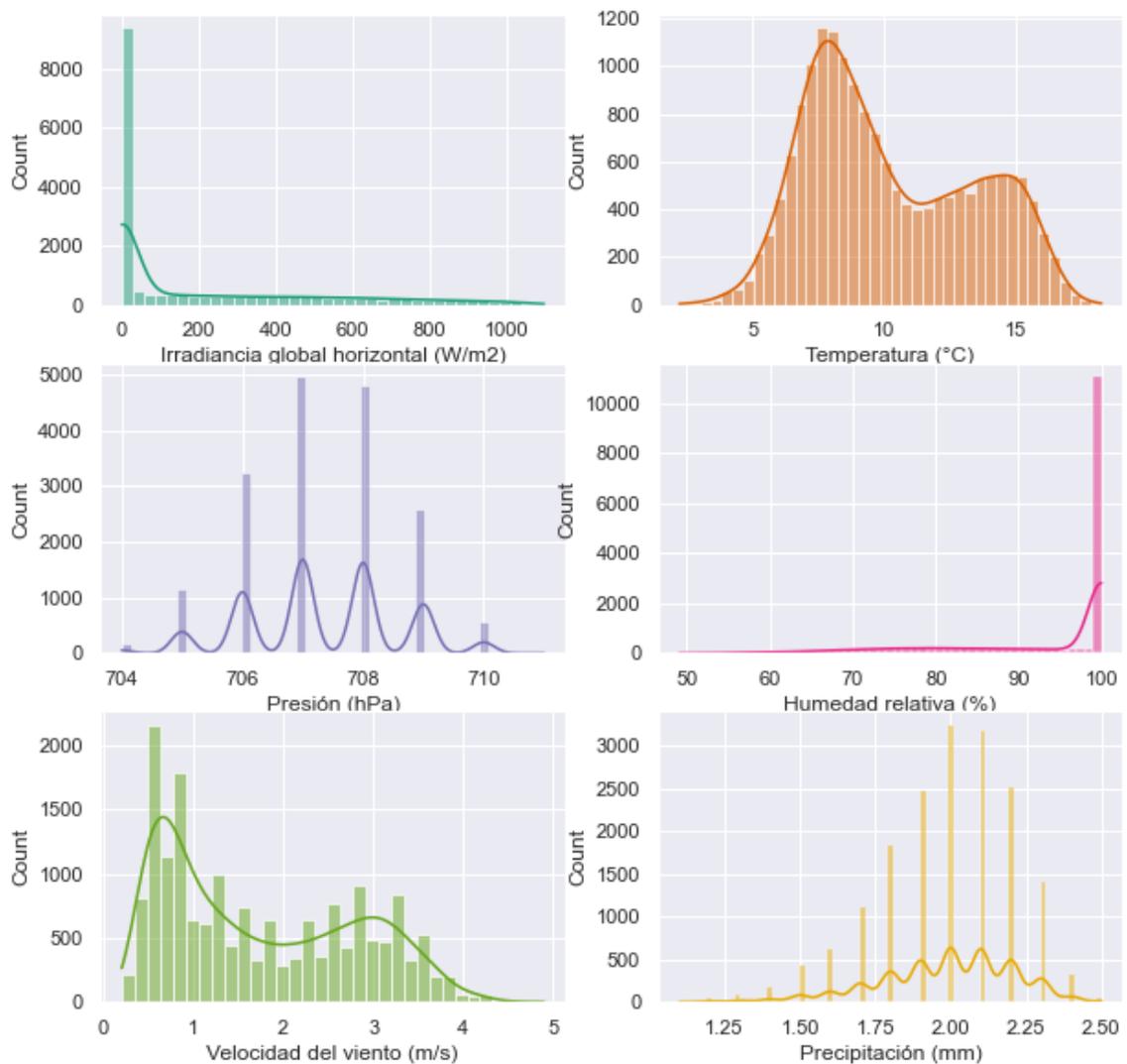


Figura 4.9. Histograma para variables ambientales de tipo satelital.

La humedad relativa dentro del punto analizado se puede visualizar que se encuentra contenido entre 49.10% y 100% de humedad, pero la mayor cantidad de datos se encuentran acumulados a partir del segundo cuartil indicando ser una zona con un elevado nivel de humedad.

Otra variable disponible es la velocidad del viento de acuerdo a los datos durante el año 2021 se han registrado velocidades de 0.20 a 4.90 m/s, indicando una tendencia de ráfagas de viento con velocidades inferiores a 1.40 m/s.

Finalmente, el nivel de precipitación presenta un problema de baja resolución similar al observado en la presión atmosférica lo que se ve representado como una serie de variaciones en el intervalo de 1.10 a 2.50 mm, pero se puede observar una tendencia a mantener niveles altos de precipitación a partir del segundo cuartil.

Para visualizar el comportamiento de los registros de radiación solar desde los datos satelitales, la Figura 4.10 muestra un análisis por horas mediante diagramas de caja donde se puede observar la línea de tendencia promedio de irradiancia, valores mínimos, valores máximos, cuartiles y se puede contrastar con el comportamiento registrado en las diferentes estaciones en tierra, hay que tener en consideración que el formato de hora debe ser ajustado a la zona horaria en Ecuador (UTC-5) ya que los datos satelitales por defecto se encuentran en formato UTC 0.

De los datos mostrados, se puede verificar que los valores más bajos de irradiancia se encuentran desde las 6 y 7 horas, asimismo, desde las 18 y 19 horas y el intervalo donde existe mayor aporte solar promedio se encuentra entre las 11 horas hasta las 14 horas, además, dependiendo de las condiciones estos niveles de luz pueden variar siendo de vital importancia definir correctamente la irradiancia solar especialmente para aplicaciones de generación eléctrica.

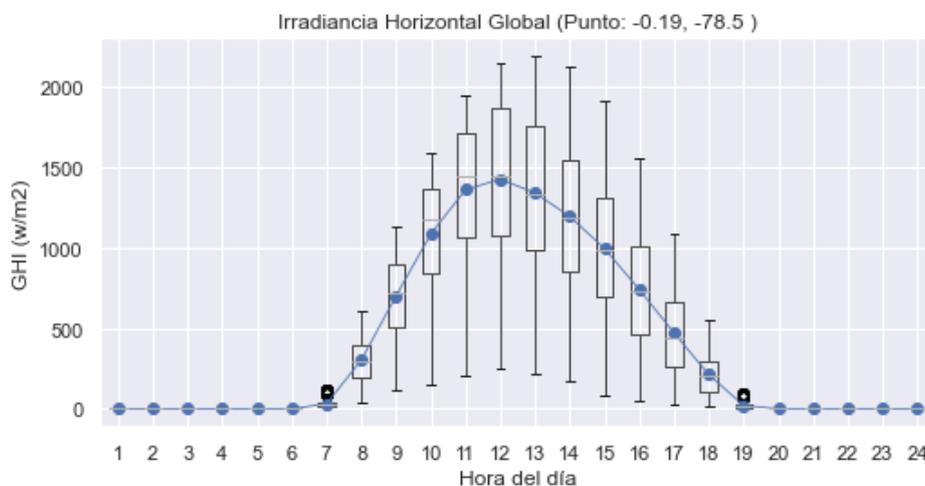


Figura 4.10. Análisis de irradiancia GHI satelital por horas.

De los datos mostrados, se puede verificar que los valores más bajos de irradiancia se encuentran desde las 6 y 7 horas, asimismo, desde las 18 y 19 horas y el intervalo donde existe mayor aporte solar promedio se encuentra entre las 11 horas hasta las 14 horas, además, dependiendo de las condiciones estos niveles de luz pueden variar siendo de vital importancia definir correctamente la irradiancia solar especialmente para aplicaciones de generación eléctrica.

Como se ha podido observar, los datos satelitales no presentan discontinuidades en sus mediciones independientemente del periodo de análisis, lo que permite evaluar de forma equilibrada la cantidad de irradiancia registrada por mes durante el año 2021, en la Figura 4.11 se muestra el resumen estadístico correspondiente a cada mes pudiéndose notar los meses con los registros más altos de radiación solar son enero, febrero, marzo, abril, octubre, noviembre y diciembre que coinciden con la temporada de invierno para la región sierra.

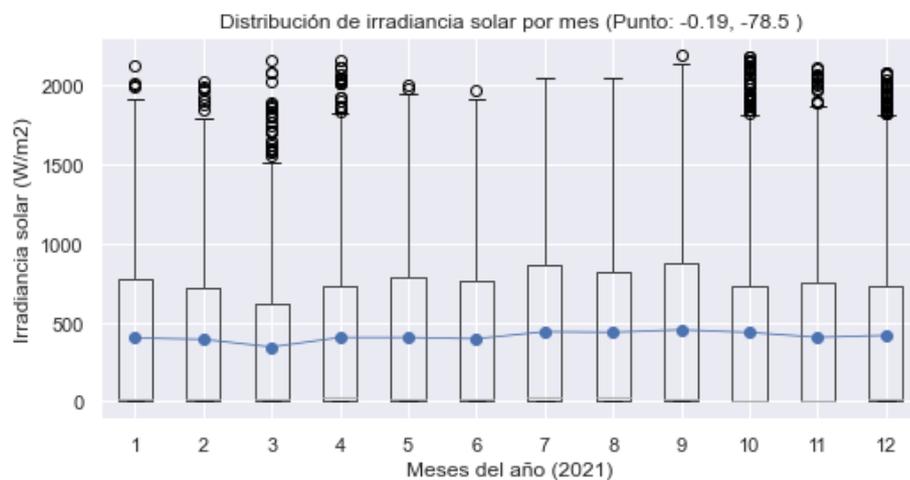


Figura 4.11. Análisis de irradiancia GHI satelital por meses.

Hasta este punto se ha analizado solamente una ubicación de las 1080 coordenadas que se encuentran disponibles, cada una de ellas se encuentra conformada por la misma cantidad de variables que refleja la realidad en el lugar durante el año 2021, la Figura 4.12 muestra un fragmento del dataset de la irradiancia horizontal global (GHI) que abarca toda la superficie de estudio donde se puede visualizar el comportamiento de la luz solar en dos horas diferentes

mediante la variación de la escala de color y de esta manera, se disponen en total de 17.509 matrices para la variable GHI, es por ello que se requieren de métodos especializados que permitan conocer el comportamiento de la luz solar partiendo de una amplia cantidad de datos como Machine Learning y Deep Learning.

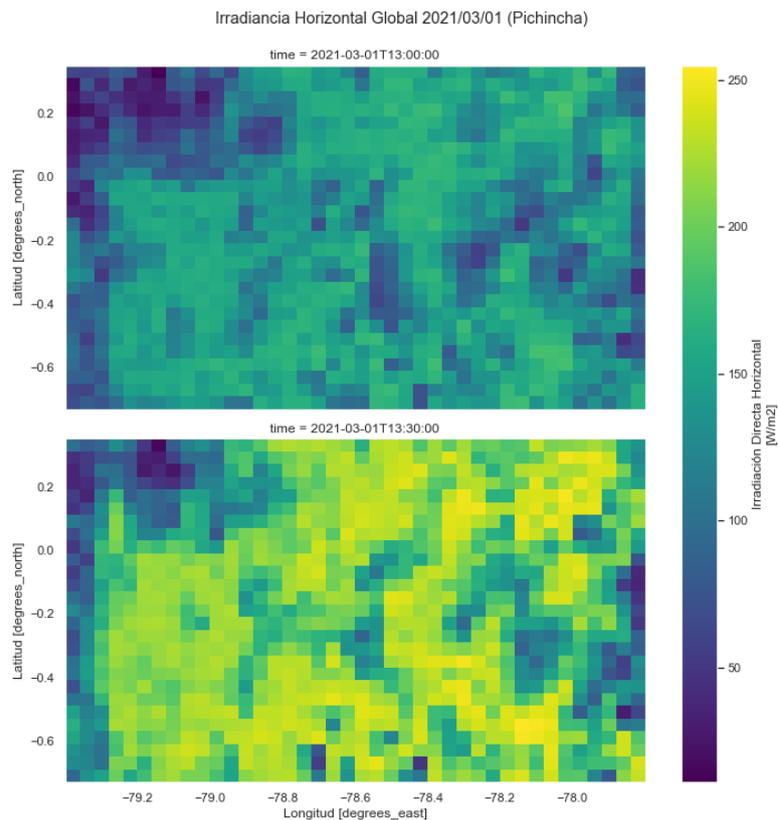


Figura 4.12. Ejemplo de matrices de irradiancia horizontal global.

Los datos satelitales presentados para el entrenamiento del modelo de regresión serán considerados sin efectuar ningún cambio debido a que el repositorio realiza un procesamiento de los datos obtenidos directamente desde los diferentes satélites previo a su publicación.

4.2 Puesta en marcha y funcionamiento

4.2.1 Modelos de regresión para construcción de dataset de irradiancia solar

Una vez que se ha realizado el análisis exploratorio y limpieza de los datos, de acuerdo a la metodología propuesta se plantea la construcción de un dataset artificial en base a los datos locales y satelitales aplicando modelos de Machine

Learning: Skforest y LGBM, de los cuales será considerado el tiempo de entrenamiento y la evaluación de los modelos de acuerdo a las métricas: error cuadrático medio (RMSE), error absoluto medio (MAE) y el coeficiente de determinación R2.

Para el ajuste de hiperparámetros se empleó la herramienta Optuna que permite realizar pruebas en intervalos de tiempo más cortos, optimizar los recursos de cómputo y abarcar rangos apropiados de cada parámetro con la finalidad de mejorar el rendimiento de los modelos.

A primera instancia, en función de la calidad de los datos y la compleción de los registros durante el año 2021 se seleccionó la estación Itulcachi (M5075) de forma referencial para realizar las respectivas pruebas de entrenamiento de modelos, ajuste de hiperparámetros y el tiempo de entrenamiento, donde los resultados obtenidos se muestran en la Tabla 4.3, teniendo en cuenta que el procedimiento descrito debe ser efectuado en las estaciones restantes.

Tabla 4.3. Resultados de entrenamiento de modelos SkForest y LGBM.

Modelo	Hiperparámetros	Iteraciones	RSME	MSE	MAE	R2	Tiempo ejecución (segundos)
Skforest (RF)	n_estimators max_depth steps	9	63.77	227.88	32.29	0.89	7787.53
LGBM	n_estimators learning_rate num_leaves max_depth min_child_samples max_bin	1000	58.14	227.88	29.71	0.91	74.06

El modelo Skforest al encontrarse basado en un regresor de tipo Random Forest tradicional los hiperparámetros que tuvieron mayor impacto durante el entrenamiento fueron *n_estimators*, *max_depth* y *steps* donde se pudieron ejecutar 9 iteraciones al cabo de 7787.53 segundos (aprox. 2.16 horas) a diferencia del modelo LGBM que pudo ejecutar 1000 iteraciones en 74.06 segundos permitiendo explorar una amplia variedad de combinaciones para ajustar adecuadamente cada hiperparámetro.

En base a los resultados obtenidos, en la Figura 4.13 se presenta la comparativa de los datos reales y sus respectivas predicciones generadas por los modelos Skforecast y LGBM a partir de los datos satelitales de entrada, donde se puede visualizar un comportamiento similar para ambos modelos siendo el único diferenciador el tiempo de entrenamiento.

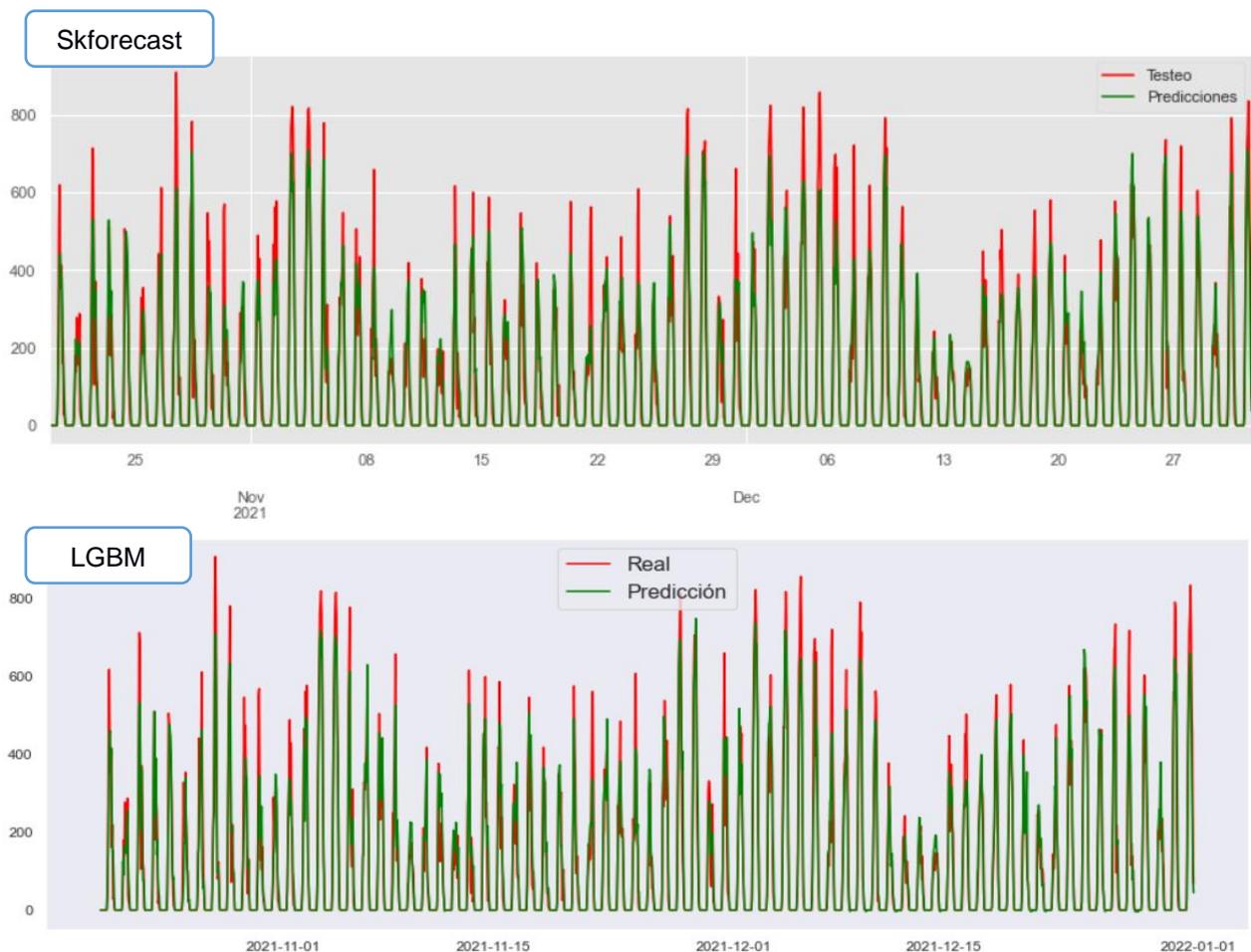


Figura 4.13. Resultados de regresión de irradiación solar en estación Itulcachi.

Durante la revisión bibliográfica se puede destacar la aplicación de estrategias de filtrado de datos de irradiancia solar para mejorar los resultados de pronóstico, de acuerdo a Pasion et al., (2020) y Fouilloy et al., (2018) se menciona el sesgo ocasionado por los ciclos nocturnos que ocupan gran porcentaje de los datos y su influencia durante el entrenamiento de modelos de Machine Learning ocasionado

un sobreajuste de los mismos, para atenuar su efecto Ordoñez et al., (2019) recomienda la eliminación de registros de radiación solar inferiores a 50 W/m².

Con la finalidad de verificar el rendimiento de la regresión a partir de datos satelitales, se aplica un filtro para eliminar los datos nocturnos dando como resultado un nuevo dataset con 3712 registros que representa el 42.39% del total del dataset inicial que será empleado para el entrenamiento del modelo LGBM debido a su mejor rendimiento y menor tiempo de ejecución.

Se realizaron pruebas con el modelo LGBM con todos los hiperparámetros configurados por defecto y el mejor modelo seleccionado por Optuna después de 1000 iteraciones, en la Tabla 4.4 se muestran los resultados obtenidos donde se puede destacar que existe un punto de partida con un rendimiento relativamente bajo incluso con el ajuste elaborado por Optuna, de acuerdo a las diferentes métricas presentadas el error absoluto medio alcanzó un puntaje de 64.99, error cuadrático medio de 88.62 y un coeficiente R2 igual a 0.78 que a diferencia de los resultados iniciales donde se incluyen los ciclos nocturnos se evidencia una disminución en el rendimiento del modelo.

Tabla 4.4. Resultados de entrenamiento LGBM sin irradiancia nocturna.

Modelo	Hiperparámetros	Iteraciones	RSME	MSE	MAE	R2	Tiempo ejecución (segundos)
LGBM (default)	Default	1	94.65	351.27	70.14	0.75	0.049
LGBM (Optuna)	n_estimators learning_rate num_leaves max_depth min_child_samples max_bin	1000	88.62	351.27	64.99	0.78	180.955

Gráficamente, en la Figura 4.14 se puede observar que el recorte efectuado por la aplicación del filtro de irradiancia afecta a la regresión propuesta por el modelo en intervalos de tiempo donde se esperan registros de radiación bajos, sumando con la resolución temporal por horas el modelo no puede representar eficazmente el comportamiento de la luz solar.

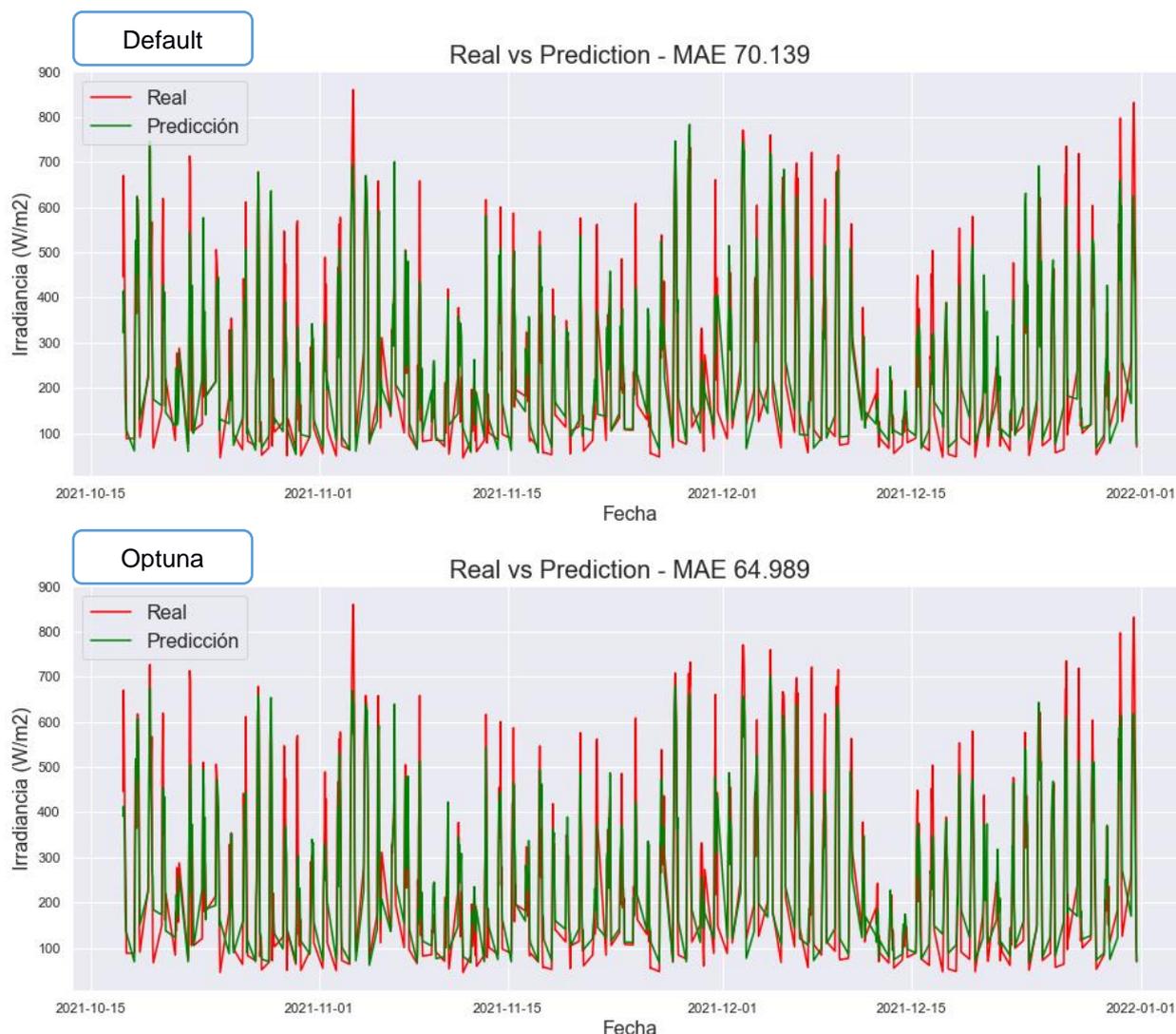


Figura 4.14. Resultados modelo LGBM sin considerar irradiación nocturna.

Una de las ventajas que se puede destacar de LGBM al ser un derivado de los árboles de decisión es su capacidad para medir la importancia de cada variable predictora y su influencia dentro del modelo resultante, trasladado al presente caso de estudio para el modelo seleccionado mediante Optuna se puede visualizar en la Figura 4.15 las variables predictoras más relevantes son: la Irradiancia Horizontal Global (GHI), velocidad del viento, hora del día, temperatura ambiental y la humedad relativa.

De estas variables se puede destacar que el modelo ha considerado la hora del día para realizar sus estimaciones de irradiancia que de forma práctica puede ser de utilidad, sin embargo, existen otras alternativas que han sido consideradas

previamente durante la selección de variables para la elaboración del dataset que podrían aportar positivamente para el cálculo de radiación solar.

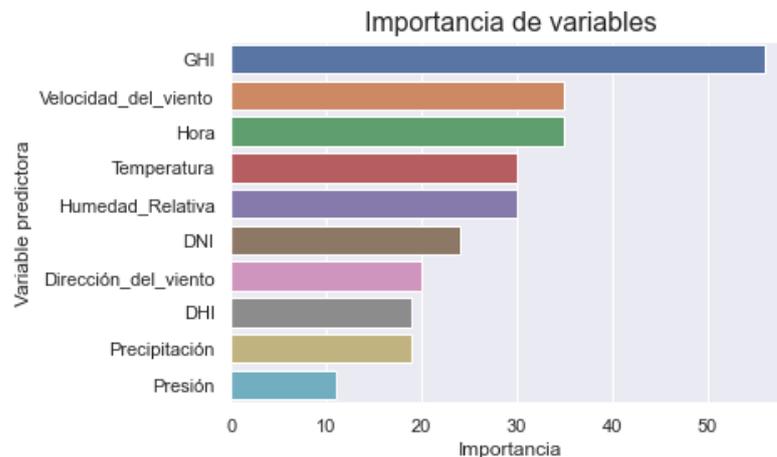


Figura 4.15. Importancia de variables para regresión de irradiancia solar sin considerar ciclos nocturnos.

Como se ha podido observar, debido a la baja resolución temporal el rendimiento del modelo LGBM ha disminuido al no poder representar correctamente la serie de tiempo de irradiancia solar, para el desarrollo del presente proyecto se mantendrán todos los valores de radiación solar.

De acuerdo a los resultados obtenidos del ajuste de hiperparámetros elaborado mediante Optuna se ha podido identificar de forma general las variables predictoras de tipo satelital con mayor relevancia en los modelos de regresión de irradiancia para todas las estaciones meteorológicas contempladas como se muestra en la Figura 4.16, siendo las principales:

- Irradiancia horizontal global (GHI)
- Temperatura ambiental
- Irradiancia directa horizontal (DHI)
- Velocidad del viento
- Dirección del viento
- Humedad relativa del ambiente

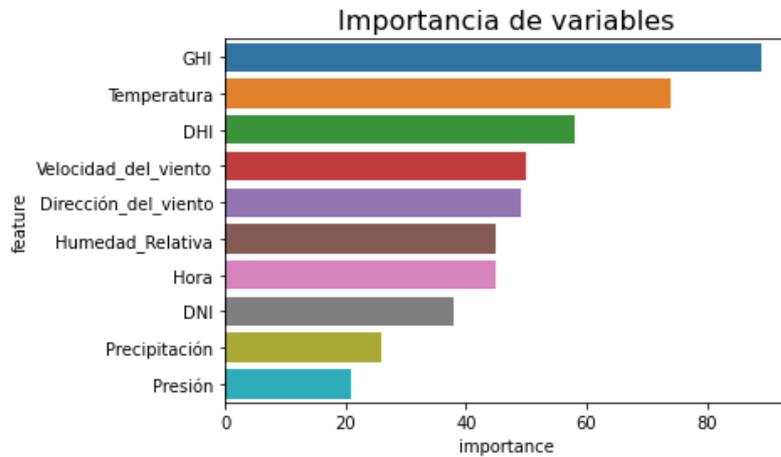


Figura 4.16. Importancia de variables para regresión de irradiancia solar.

Previamente se hizo referencia a los hiperparámetros con mayor impacto durante el entrenamiento de los modelos LGBM, cada uno de ellos fueron probados individualmente en función a la respuesta de los diferentes modelos, el incremento de su rendimiento y a través de Optuna se puede visualizar la importancia de cada hiperparámetro como se muestra en la Figura 4.17 donde se puede observar que `n_estimators`, `min_child_samples` y `max_bin` han sido determinantes para la regresión de irradiancia solar.

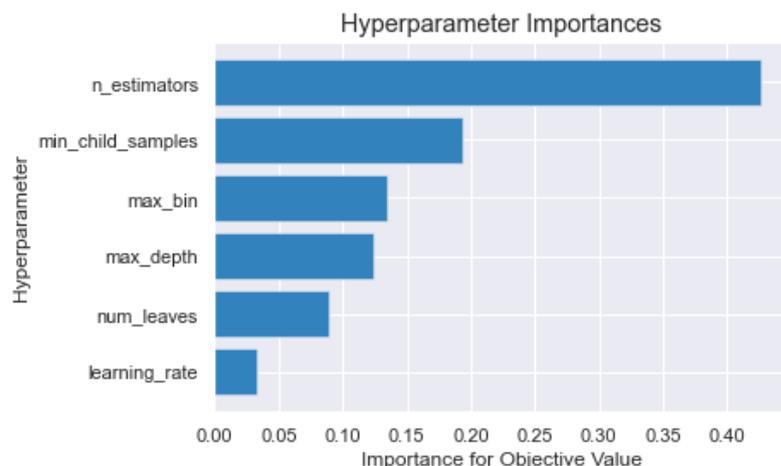


Figura 4.17. Importancia de hiperparámetros para regresión de irradiancia solar.

Como tal, LGBM dispone de una amplia variedad de hiperparámetros y valores que puede adoptar cada uno de ellos e influir en el rendimiento del modelo, Optuna por su parte permite realizar un ajuste de forma automática y visualizar la

convergencia de cada hiperparámetro de tal forma que permite al usuario intuir los rangos que se deben explorar para optimizar los resultados del modelo a entrenar, en la Figura 4.18 se muestra el comportamiento que tuvo el ajuste automático de acuerdo a cada parámetro a medida del incremento de iteraciones, al tratarse de conjuntos de datos similares debido a que los puntos de referencia son contiguos los rangos de exploración se mantuvieron constantes para determinar los respectivos modelos para cada estación.

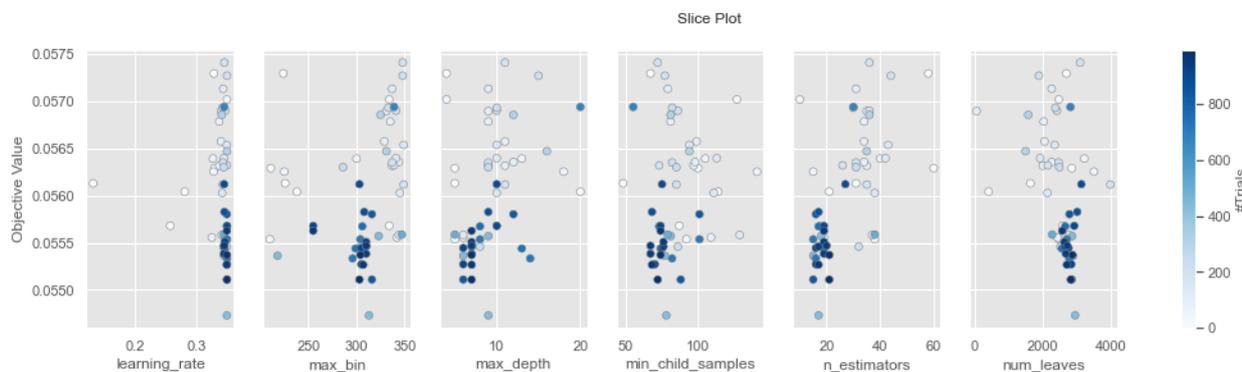


Figura 4.18. Ajuste de hiperparámetros por iteración.

Para cada estación se obtuvo un modelo de regresión de radiación solar utilizando como referencia el modelo LGBM por defecto y el ajuste realizado mediante Optuna, de esta forma se puede realizar una comparativa de rendimiento entre modelos, los resultados obtenidos se muestran en la Tabla 4.5.

Tabla 4.5. Evaluación de modelos de irradiancia solar por estación.

Estación	Modelo	MAE	MSE	RMSE	R2
C1	Default	39,336	310,487	83,603	0,895
	Ajustado	36,717	310,487	80,342	0,903
C2	Default	50,240	326,957	99,734	0,867
	Ajustado	49,835	326,957	98,484	0,870
C4	Default	36,981	271,180	80,901	0,882
	Ajustado	36,300	271,180	78,901	0,888
C05	Default	40,341	282,610	83,256	0,876
	Ajustado	39,792	282,610	79,518	0,887
C06	Default	42,945	320,422	85,980	0,894
	Ajustado	41,139	320,422	80,150	0,908
C07	Default	29,071	327,536	59,221	0,955
	Ajustado	28,471	327,536	56,656	0,959
C08	Default	32,225	317,695	64,583	0,942

	Ajustado	31,873	317,695	61,130	0,948
C10	Default	48,127	344,922	96,302	0,887
	Ajustado	46,657	344,922	90,097	0,901
C13	Default	44,423	262,516	94,672	0,814
	Ajustado	43,102	262,516	90,266	0,831
C14	Default	36,529	303,622	83,309	0,898
	Ajustado	34,236	303,622	75,387	0,916
C16	Default	35,044	166,430	68,321	0,753
	Ajustado	35,154	166,430	64,852	0,777
C18	Default	30,904	366,245	59,593	0,960
	Ajustado	29,624	366,245	56,285	0,964
C19	Default	35,886	323,452	73,776	0,927
	Ajustado	36,305	323,452	70,637	0,933
M5021	Default	39,175	302,164	74,722	0,914
	Ajustado	36,014	302,164	67,547	0,930
M5025	Default	40,694	243,308	83,185	0,833
	Ajustado	36,526	243,308	74,521	0,866
M5026	Default	51,938	303,226	104,750	0,832
	Ajustado	50,376	303,226	98,028	0,853
M5027	Default	37,087	302,034	71,880	0,920
	Ajustado	34,111	302,034	66,804	0,931
M5028	Default	42,987	232,945	89,958	0,791
	Ajustado	42,537	232,945	87,488	0,803
M5029	Default	43,787	302,271	90,509	0,870
	Ajustado	43,880	302,271	88,063	0,877
M5031	Default	39,099	259,492	77,885	0,872
	Ajustado	37,576	259,492	73,374	0,887
M5074	Default	52,416	320,086	102,826	0,851
	Ajustado	51,177	320,086	100,688	0,857
M5075	Default	32,156	227,879	62,589	0,893
	Ajustado	30,341	227,879	58,826	0,906
M5178	Default	32,605	206,599	63,334	0,865
	Ajustado	31,904	206,599	60,867	0,875
M5180	Default	43,454	268,918	84,588	0,862
	Ajustado	41,439	268,918	80,419	0,875

Como se puede observar, los cambios entre los resultados del modelo por defecto y la optimización de hiperparámetros muestran ligeras mejoras en las métricas adoptadas, siendo el Error absoluto medio el que presenta más variaciones entre los diferentes modelos obtenidos como se muestra en la Figura 4.19, dado que el valor más alto corresponde a la estación M5074 con un valor de 51.177 indicando

un posible rendimiento bajo del modelo de regresión y la estación que ha presentado mejores resultados en su modelo es la C07 con un valor de 28.471.

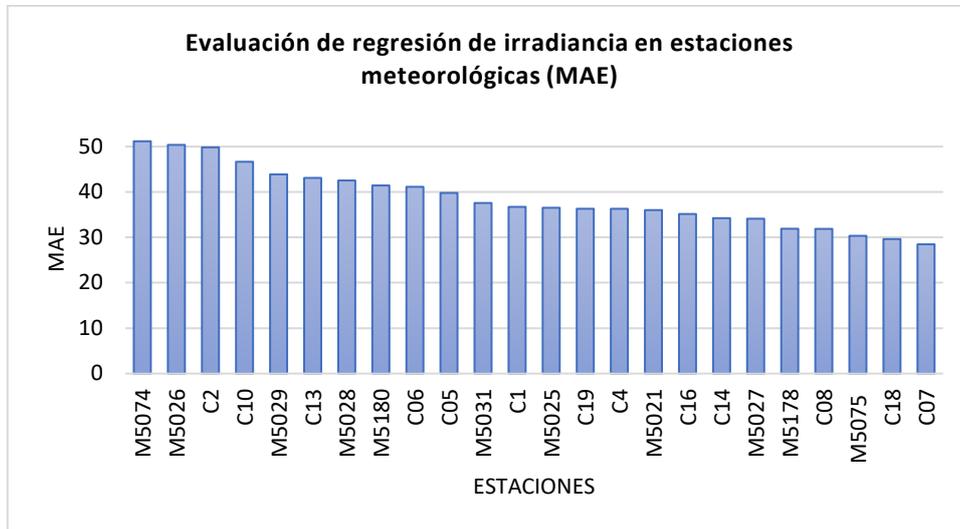


Figura 4.19. Evaluación de regresión de irradiancia en estaciones meteorológicas según métrica MAE.

Teniendo en consideración los resultados de otras métricas, en el caso del coeficiente de determinación R^2 permite cuantificar el ajuste del modelo a los datos reales, siendo el modelo de la estación C18 el que mejor rendimiento presenta a nivel general con un valor de 0.964 y el mínimo correspondiente a la estación C16 con un valor de 0.777, a partir de este punto la expectativa de cada modelo va incrementando.

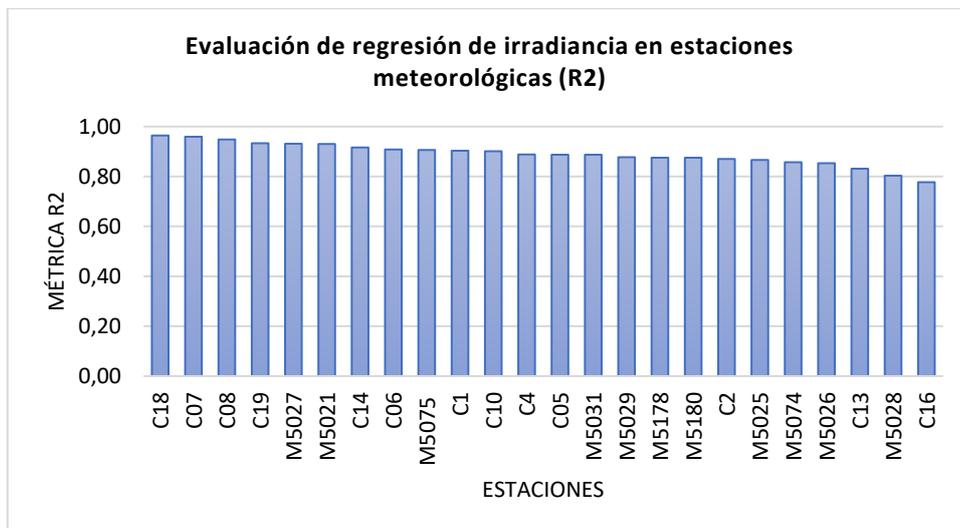


Figura 4.20. Evaluación de regresión de irradiancia en estaciones meteorológicas según métrica R2.

4.2.2 Aplicación de modelo de regresión

Una vez que se han obtenido los diferentes modelos de regresión para cada estación, se requiere llevar a cabo su implementación para aproximar los datos satelitales dentro del área de interés que comprende toda la superficie de la provincia de Pichincha, para realizar este procedimiento se realizaron diversas pruebas con la finalidad de emplear el o los modelos más cercanos al punto de interés en un radio de 4 km.

Teniendo en cuenta que el área de interés es pequeña se empleó el cálculo de distancias entre dos puntos en el plano 2D y se seleccionaron las distancias más cercanas para considerar seleccionar el modelo, en el caso que se contemplen superficies más amplias o cercanos a los polos es necesario considerar la curvatura de la tierra para tener una estimación precisa.

El área de interés se encuentra comprendida por 1080 puntos localizados cada 4 Km que se encuentran distribuidos en 40 puntos en el eje de longitud y 27 puntos en el eje de latitud que facilita su representación como una matriz de 40x27, donde se realizaron pruebas de selección de modelos utilizando como referencia 64 puntos ubicados por cuadrantes.

De las pruebas realizadas se clasificaron los resultados por categorías, siendo: bueno (verde), aceptable (azul) e inconsistente (rojo), con fines de visualización como se muestra en la Figura 4.21 se ha asignado un color a los resultados y ubicados en el punto empleado para las pruebas.

De acuerdo a la representación visual, se puede notar que el método empleado no satisface en su totalidad la realidad del comportamiento de la luz solar dentro del área de interés, apenas 21 puntos seleccionados presentaron resultados acordes a los criterios establecidos, mientras que 12 puntos presentaron desfases leves en los registros de radiación y pueden ser empleados, sin embargo, 31 puntos presentaron resultados desfavorables indicando que la metodología aplicada no es adecuada para ser implementada en el resto de puntos.

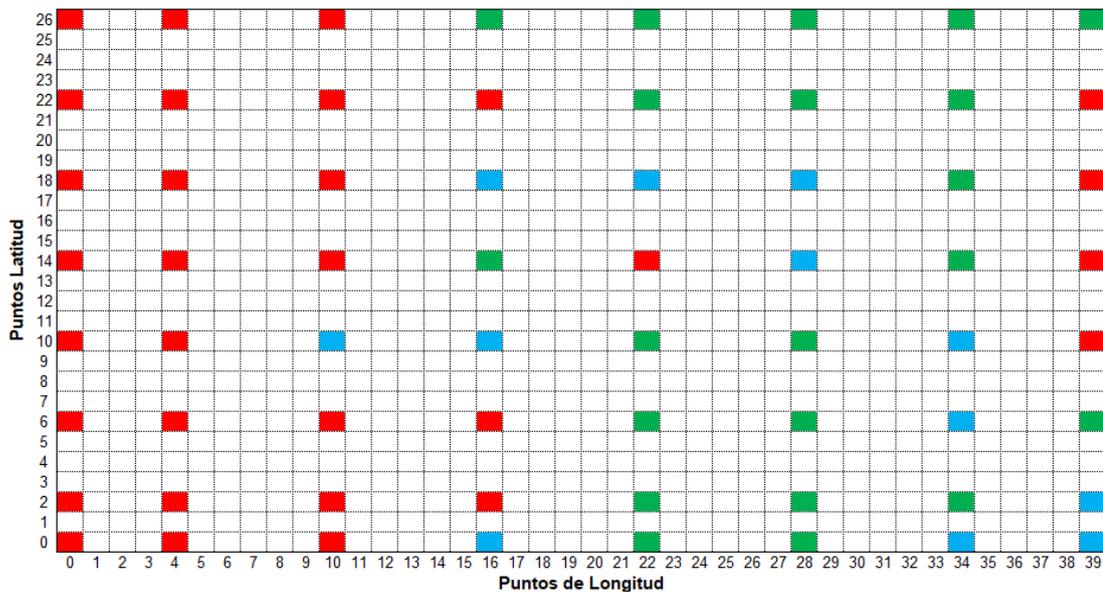


Figura 4.21. Resultados de evaluación de modelos de regresión de radiación solar.

Cabe destacar que los puntos seleccionados no disponen de información real para validar los resultados de la regresión generada a través de los modelos y los resultados se han clasificado de forma manual, se descartaron las regresiones que presenten desfases de la luz solar en horarios que no correspondan o presenten registros de irradiancia mayores a cero durante ciclos nocturnos.

Otra consideración a tener en cuenta es que los resultados clasificados como “buenos” se encuentran distribuidos en el rango de los puntos 22 a 28 correspondientes a las longitudes -78.5 a -78.25 respectivamente, y coinciden con la ubicación real de la mayoría de estaciones meteorológicas consideradas para el desarrollo del proyecto como se muestra en la Figura 4.22, otro aspecto importante es la altitud en la que se encuentran ubicadas las estaciones, a partir del gráfico de presión atmosférica se puede contrastar que el bajo desempeño en la regresión de irradiancia se dieron en zonas de altitud baja que en consecuencia presentan niveles altos de presión atmosférica.

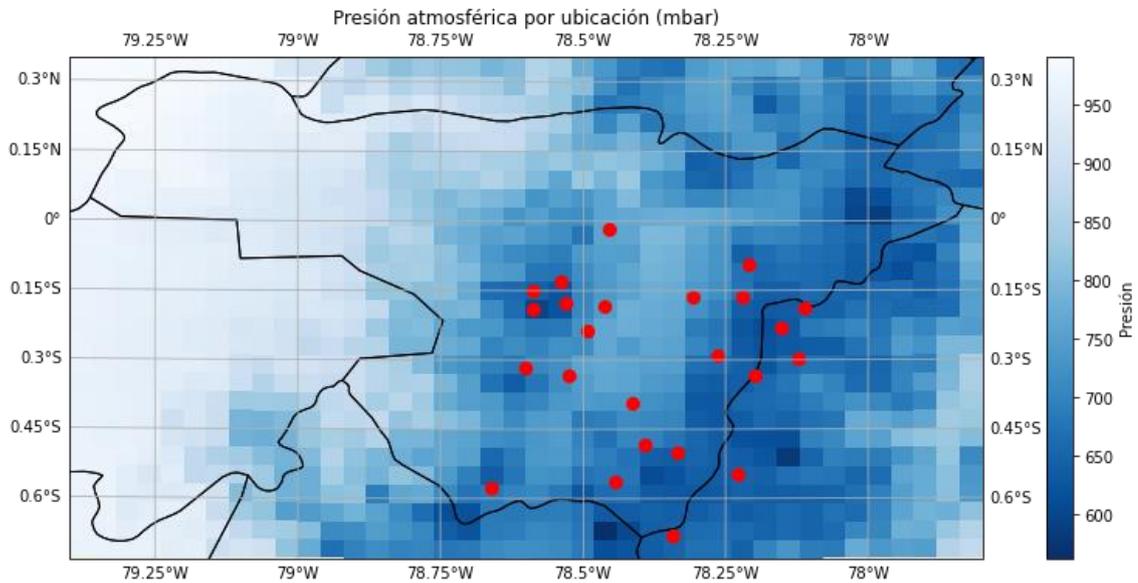


Figura 4.22. Presión atmosférica por ubicación en la provincia de Pichincha.

Teniendo en cuenta las limitaciones que conlleva realizar el ajuste de irradiancia en función a la cercanía de un punto dado a una estación meteorológica, se realizó una evaluación individual del rendimiento de cada modelo con respecto a las ubicaciones geográficas en las que se dispone de la suficiente información para validar cada regresión, para este caso de estudio los únicos puntos geográficos que cumplen con esta condición se encuentran asociados a las estaciones: “M5021”, “M5027”, “M5028”, “M5029”, “M5031”, “M5074”, “M5075”, “M5178” y “M5180”, cabe recalcar que cada estación se encuentra localizada en puntos independientes dentro de la provincia.

Los resultados de la aplicación de cada modelo para el conjunto de ubicaciones propuesto pueden verse representados en la Figura 4.23, donde se encuentran marcados los resultados del coeficiente de determinación (R^2) utilizado como métrica de evaluación, donde se puede evidenciar la existencia de modelos que presentan un bajo rendimiento fuera de su localización, no obstante, los modelos “C07”, “C18”, “M5075”, “M5027”, “C19” han podido generalizar satisfactoriamente los datos de irradiancia en las ubicaciones de prueba.

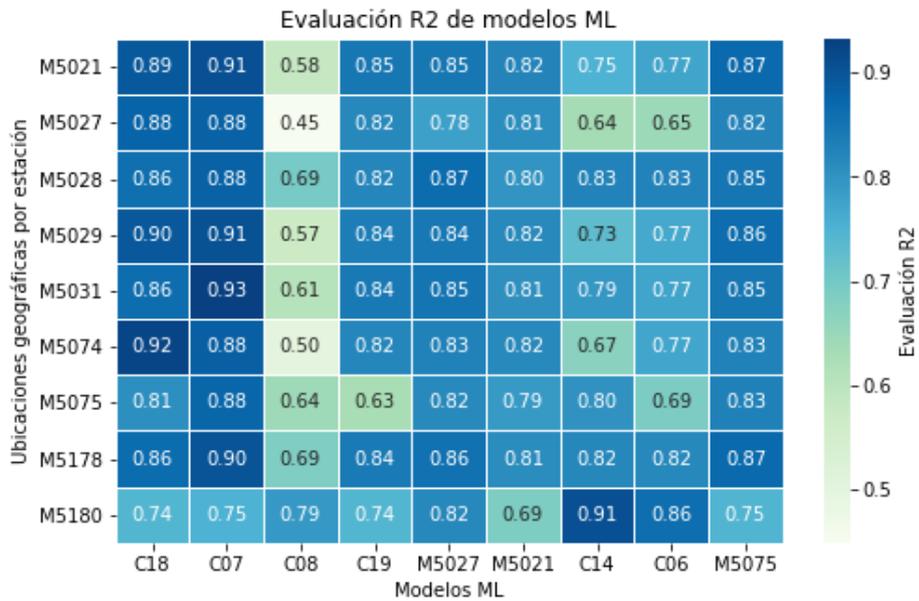


Figura 4.23. Evaluación de regresión de modelos ML para ubicaciones de prueba.

Posteriormente, se realizó la combinación de los cuatro modelos con mejor rendimiento para promediar sus regresiones y obtener un modelo resultante más robusto que pueda interpretar las variables exógenas ingresadas para acercarse a las mediciones reales, este modelo combinado fue aplicado a cada uno de los puntos del área de estudio para generar los datos de irradiancia que posteriormente fueron almacenados en un archivo netCDF, debido a la versatilidad para manejar datos espaciales y series de tiempo.

Del dataset resultante, se puede destacar la posibilidad de acceso a información detallada de la cantidad de irradiancia en la provincia de Pichincha para el año de estudio, en este caso 2021 contemplando la continuidad temporal para cada punto y la precisión cercana a los datos reales que se obtuvieron al combinar las fuentes de datos locales y satelitales, en la Figura 4.24 se muestra un fragmento de la serie de tiempo para dos puntos seleccionados de forma aleatoria dentro del área de estudio donde se puede verificar la serie de tiempo resultante.

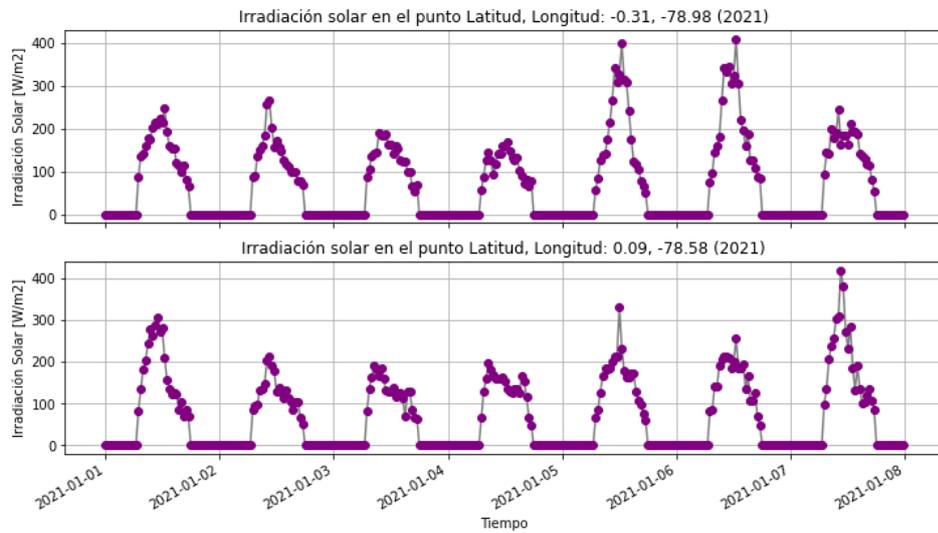


Figura 4.24. Visualización de serie de tiempo de radiación solar.

Con los datos disponibles, es posible calcular la irradiación anual a partir de la suma de la radiación solar experimentada en el año con la finalidad de identificar zonas de alto potencial solar en la provincia como se muestra en la Figura 4.25, este tipo de información es de vital importancia durante el dimensionamiento de sistemas fotovoltaicos. En el país los únicos registros de irradiación se encuentran en mapas solares que se encuentran desactualizados, donde la identificación de zonas de interés se debe realizar de forma manual basándose únicamente en las escalas de color.

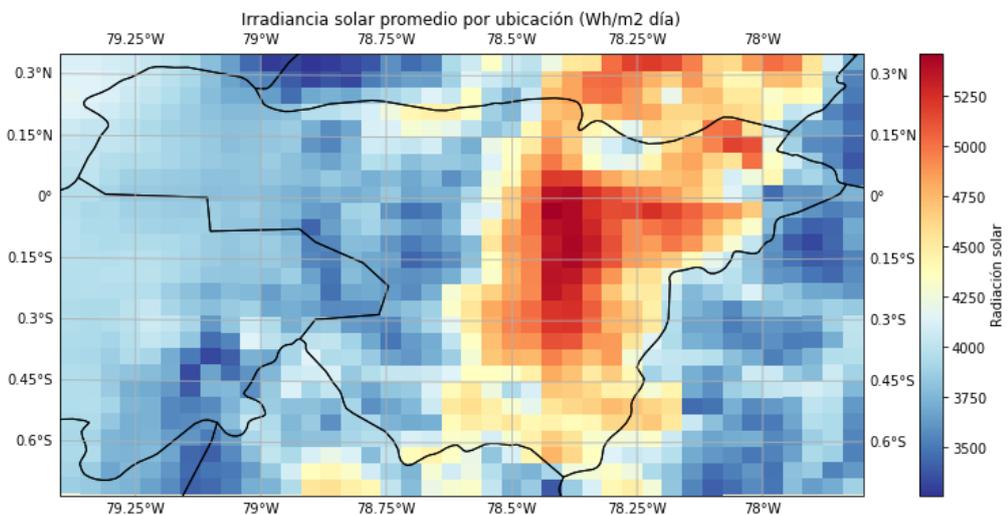


Figura 4.25. Irradiación solar anual en la provincia de Pichincha durante 2021.

4.2.3 Métodos estadísticos para pronóstico de irradiancia solar

Una vez que se ha consolidado un dataset de irradiancia solar que abarca toda el área de interés manteniendo su componente temporal para cada punto geográfico, se requiere analizar su comportamiento y poder pronosticarla varios pasos en el tiempo para facilitar el dimensionamiento de sistemas fotovoltaicos, se ha considerado como punto de partida la aplicación de los métodos Holt y Holt-Winters.

En las secciones 3.4.1 y 3.4.2 se ha detallado la metodología a ser implementada, cabe destacar que los pronósticos generados por los métodos estadísticos serán usados como un punto referencial a ser mejorado a partir de la implementación de modelos basados en Deep Learning.

A primera instancia, el método Holt se caracteriza por su capacidad de capturar el comportamiento de la tendencia de una serie temporal en función de la importancia que se dé a la antigüedad de los datos y la velocidad de cambio en la tendencia de la serie.

Para las pruebas planteadas se tomó un punto al azar dentro del área de estudio (Latitud: -0.31, Longitud: -78.26), del cual se extrajo la serie de tiempo y mediante la herramienta *seasonal_decompose* de la librería *statsmodels* se puede analizar rápidamente las componentes de la serie como se muestra en la Figura 4.26, se puede evidenciar que el comportamiento de la luz solar presenta una tendencia variable y no describe un patrón característico, no obstante, su estacionalidad sigue un patrón fácil de identificar asociado a los ciclos de luz que se repiten diariamente y se pueden modelar, finalmente, se puede observar una gran cantidad de residuos que no se han podido representar en la descomposición y son un indicio que se puede extraer más información de la serie.

Los datos de la serie de tiempo se han segmentado en una proporción de 80% para entrenamiento y el 20% restante para evaluar el ajuste obtenido del pronóstico generado a partir de los métodos propuestos.

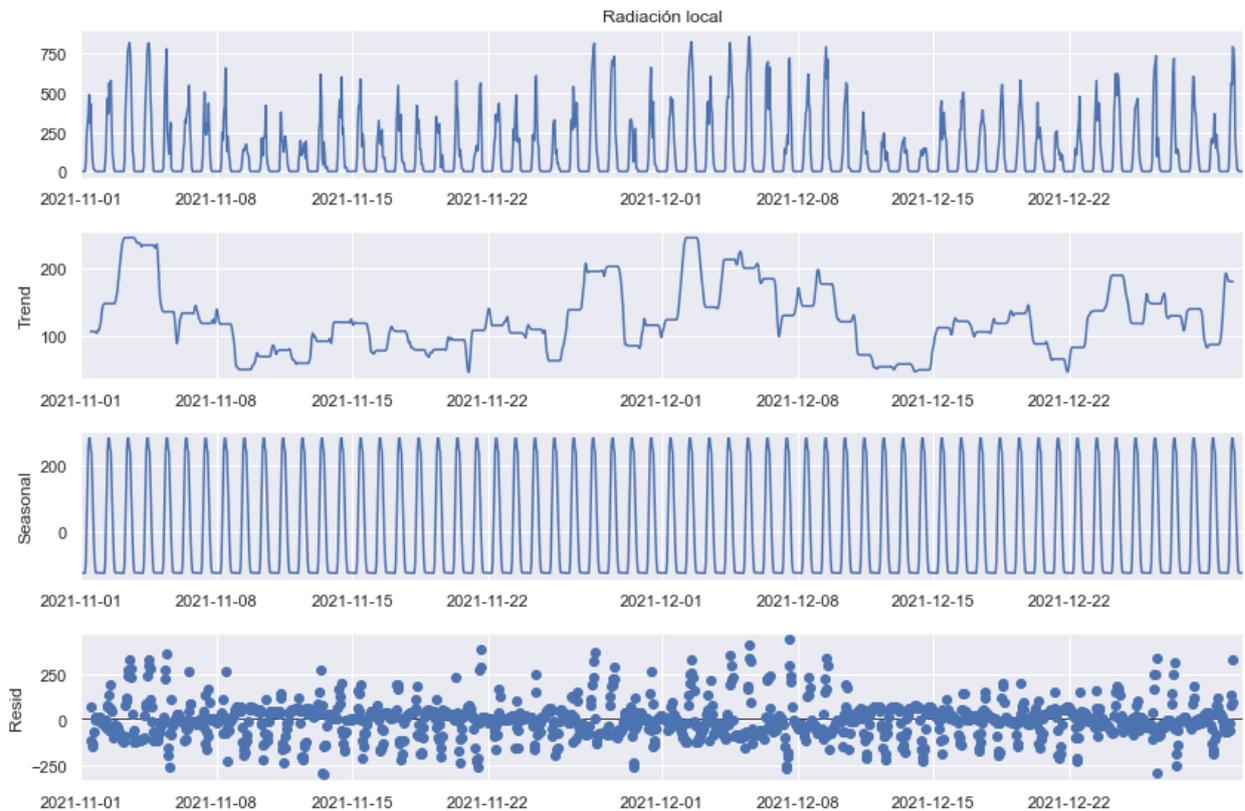


Figura 4.26. Descomposición estacional de serie de tiempo de irradiancia en el punto (-0.31, -78.26).

De las pruebas realizadas, el mejor ajuste obtenido a partir del método Holt se presenta en la Figura 4.27, donde se puede verificar que se ha modelado únicamente la tendencia donde los últimos datos del conjunto de entrenamiento han dado la base para generar el pronóstico, el método implementado no tiene la capacidad para modelar la estacionalidad de la serie.

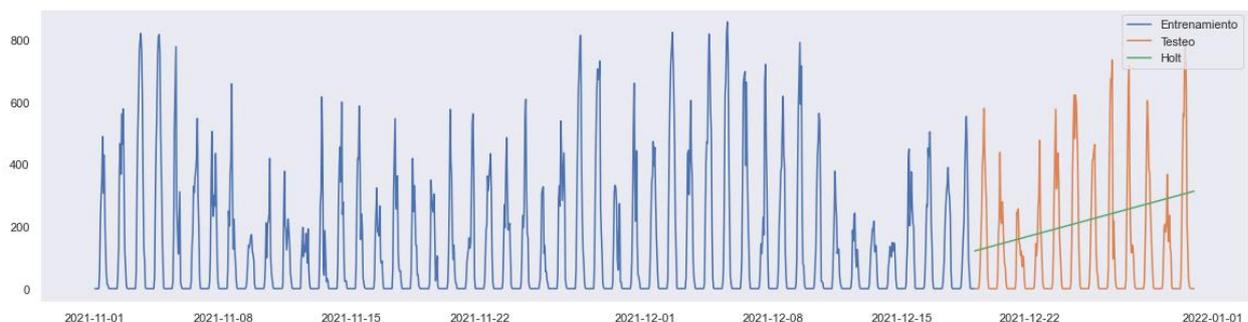


Figura 4.27. Pronóstico de irradiancia solar mediante método Holt.

Como se ha podido observar, la serie de tiempo de irradiancia presenta un nivel, tendencia variable y una estacionalidad marcada, entonces el método Holt-Winters es idóneo para el caso de estudio debido a que precisa de estos parámetros para desarrollar pronósticos.

Para el desarrollo de este proyecto se exploraron a través de una búsqueda de tipo *grid_search* la mejor configuración a partir de los siguientes parámetros:

- Tendencia (aditiva, multiplicativa, ninguna).
- Amortiguamiento.
- Estacionalidad (aditiva, multiplicativa, ninguna).
- Periodos de estacionalidad (considerados cada 24 y 48 observaciones).

De las pruebas realizadas se pudo determinar que el modelo con el mejor rendimiento fue generado sin considerar ningún tipo de tendencia, ausencia de nivel, estacionalidad de tipo aditiva y el análisis de la estacional cada 48 observaciones, el pronóstico obtenido se muestra en la Figura 4.28 donde se puede evidenciar la dificultad para modelar el comportamiento de la irradiancia solar debido a que los pesos de los datos van decreciendo a medida que los datos son más antiguos en el tiempo.

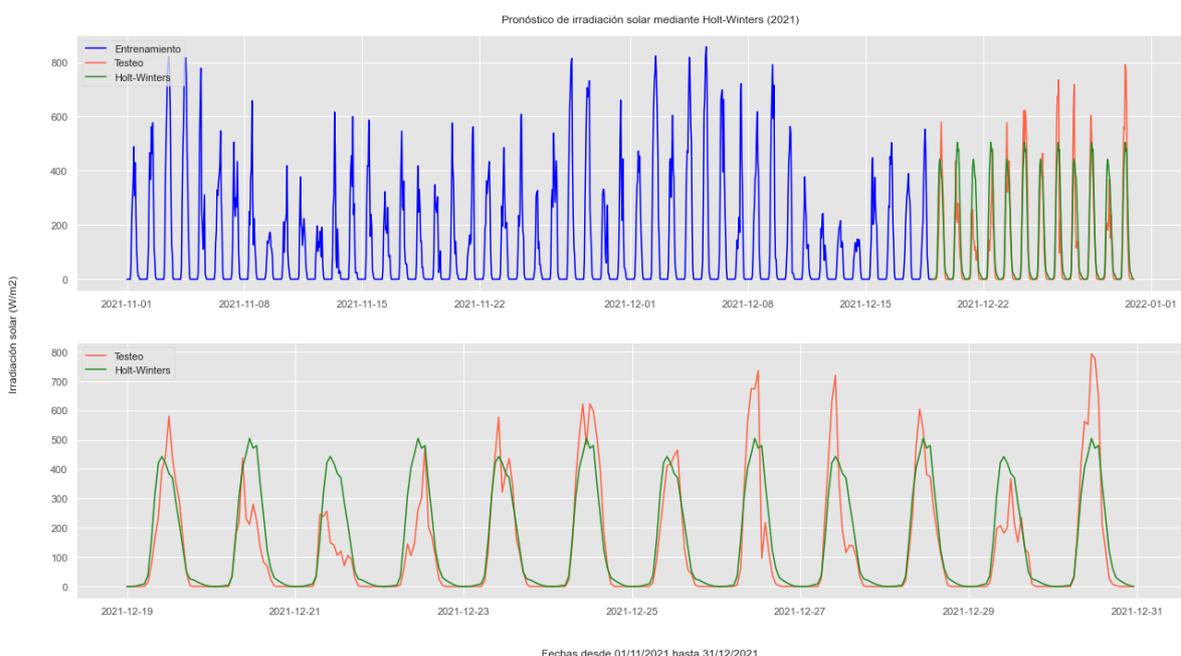


Figura 4.28. Pronóstico de irradiancia solar mediante método Holt-Winters.

De las pruebas realizadas mediante el método de Holt-Winters se exploraron diversas configuraciones y su ajuste con los datos reales, en la Tabla 4.6 se puede visualizar los parámetros de los cuatro mejores modelos obtenidos, donde se puede destacar que la serie de tiempo presenta una estacionalidad de tipo aditivo y el mejor periodo de análisis de la estacionalidad es 48 horas, amortiguamiento variable y la tendencia en la mayoría de las pruebas obtuvo un buen rendimiento con tendencia aditiva no obstante, el mejor modelo no adopta ninguna configuración de tendencia.

Tabla 4.6. Evaluación de pronóstico mediante método Holt-Winters.

Parámetros de entrenamiento				Métricas de evaluación			
Tendencia	Amortiguamiento	Estacionalidad	Periodo	MAE	MSE	RMSE	R2
Ninguno	Falso	Aditivo	48	39,31	5790,39	76,10	0,836
Aditiva	Verdadero	Aditivo	48	39,65	5839,80	76,42	0,835
Aditiva	Falso	Aditivo	48	39,78	5911,54	76,89	0,833
Aditiva	Verdadero	Aditivo	24	41,38	6151,80	78,43	0,826

Finalmente, de los métodos implementados se tiene su respectiva evaluación como se muestra en la Tabla 4.7, como se mencionó previamente el método Holt no es apropiado para representar la estacionalidad de la serie y se ve reflejado en su evaluación a diferencia del método Holt-Winters donde se pudo observar que sus predicciones se asemejan al comportamiento real de la irradiancia, aunque no es capaz de representar variaciones espontáneas de la luz solar debido a que considera el promedio de las observaciones más recientes.

Tabla 4.7. Evaluación de métodos estadísticos para pronóstico de irradiancia.

Método	MAE	MSE	RMSE	R2
Holt	190,88	225,98	212,88	-0,28
Holt-Winters	39,31	5790,39	76,10	0,836

La línea base para pronóstico para el presente proyecto será tomada a partir de la evaluación del método Holt-Winters.

4.2.4 Pronóstico de irradiancia solar por puntos

Mediante la implementación de los métodos estadísticos Holt y Holt-Winters se ha definido la línea base que se pretende mejorar mediante el uso de modelos de Deep-Learning, siendo las redes recurrentes LSTM la opción seleccionada debido a su capacidad para almacenar dependencias a largo plazo para ser empleadas para generar predicciones.

En la sección 3.4.3 se ha presentado la metodología a seguir teniendo en cuenta el procesamiento respectivo de la serie de tiempo previo al entrenamiento del modelo, se debe verificar si la serie tomada es estacionaria o no, para ello se empleará la prueba de Dickey-Fuller o prueba de raíz unitaria que permite determinar el efecto de la tendencia en una serie de tiempo (Brownlee, 2017), donde se plantea la siguiente prueba de hipótesis:

- **Hipótesis nula (H_0):** Si la hipótesis no se rechaza indica la existencia de raíces unitarias lo que implica que la serie de tiempo mantiene una estructura que depende del tiempo, es decir, la serie no es estacionaria.
- **Hipótesis alternativa (H_1):** Se rechaza la hipótesis nula, indicando que no existen raíces unitarias a partir de la serie de tiempo, por lo tanto, la serie es estacionaria.

Este test puede ser implementado fácilmente en Python a través de la librería *statsmodels* en su complemento *adfuller*, donde se puede evaluar una serie de tiempo a través de un valor estadístico ADF, prueba estadística p y valores críticos, se ha seleccionado un punto geográfico dentro del conjunto de datos para realizar la evaluación de la serie de tiempo en dicho punto, los resultados de muestran en la Tabla 4.8.

Tabla 4.8. Resultados test Dickey-Fuller para serie de tiempo de irradiancia solar.

Test Dickey-Fuller	
Estadística ADF	-12.734
Valor p	0.000
Valores críticos	
1%	-3.431
5%	-2.862
10%	-2.567

Con la prueba p se determinará la probabilidad que la hipótesis nula es verdadera, considerando un valor $p < 0.05$ (5%) se puede rechazar la hipótesis nula, es decir, la serie de tiempo estudiada es estacionaria, asimismo, un valor superior a 0.05 (5%) no se puede rechazar la hipótesis nula, es decir, la serie de tiempo no estacionaria y por lo tanto, se requieren realizar un procesamiento adicional de los datos (Brownlee, 2017).

De acuerdo a los resultados obtenidos, el valor estadístico ADF es -12.734 que es menor que el valor p (0.000) e incluso, considerando un nivel de significancia del 1% (-3.431) el resultado del estadístico ADF sigue siendo menor, lo que permite rechazar la hipótesis nula indicando que la serie ingresada es estacionaria debido a que no presenta ningún tipo de dependencia temporal.

Posteriormente, los datos deben ser adaptados para que sean compatibles con la entrada de la red LSTM, es decir, se requiere el ingreso a partir de una matriz tridimensional compuesta por el número de muestras, pasos de tiempo y las variables de pronóstico, estructuradas como: *(batch_size, time_steps, features)*.

La estructura de la red LSTM propuesta se muestra en la Figura 4.29, debido a que la serie temporal es univariada y previamente los datos han sido acondicionados, se empleará un modelo sencillo compuesto por dos capas LSTM y durante las diferentes pruebas realizadas se ha determinado la cantidad de unidades de memoria necesarias para poder representar el comportamiento de la luz solar a través del modelo de inteligencia artificial.

Tabla 4.9. Configuración de modelo LSTM.

Capas (Tipo)	Dimensiones de salida	Número de parámetros
Capa LSTM (10 unidades)	(1, 1, 10)	480
Dropout (0.1)	(1, 1, 10)	0
Capa LSTM (10 unidades)	(1, 10)	840
Dropout (0.1)	(1, 10)	0
Dense	(1, 1)	11
Total		1.331

Para construir el modelo se utilizará un esquema mostrado en la Tabla 4.9 basado en dos capas LSTM consecutivas compuestas por 10 unidades de memoria, después de cada capa LSTM se aplica una capa *dropout* para controlar el sobreajuste del modelo y finalmente se tiene una capa densa cuya dimensión de salida (1,1) proporciona el pronóstico generado del siguiente paso de tiempo, la función de activación empleada para esta capa se aplica la selección por defecto en Keras (lineal).

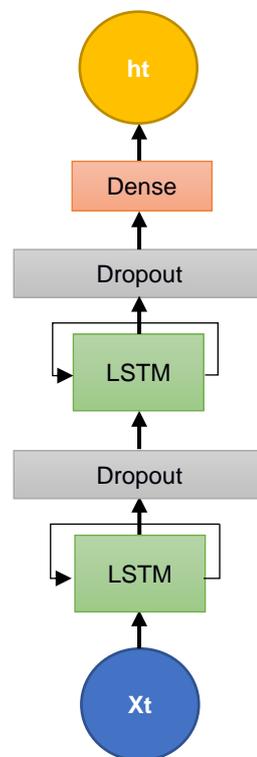


Figura 4.29. Estructura modelo LSTM para irradiación solar.

Una vez que se ha entrenado el modelo, se realizaron pruebas con el conjunto de datos de testeo que comprende 803 puntos en el futuro, teniendo en cuenta que cada punto representa 30 minutos en la escala real y se consideran únicamente los ciclos diurnos (10 horas al día) se tiene un pronóstico de aproximadamente 40 días a futuro, como se puede observar en la Figura 4.30 las predicciones realizadas por el modelo (naranja) pueden representar satisfactoriamente el comportamiento real de la irradiación solar (azul), cabe destacar que los

pronósticos son generados mediante pasos únicos (one-step) y posteriormente concatenados para generar la serie de salida.

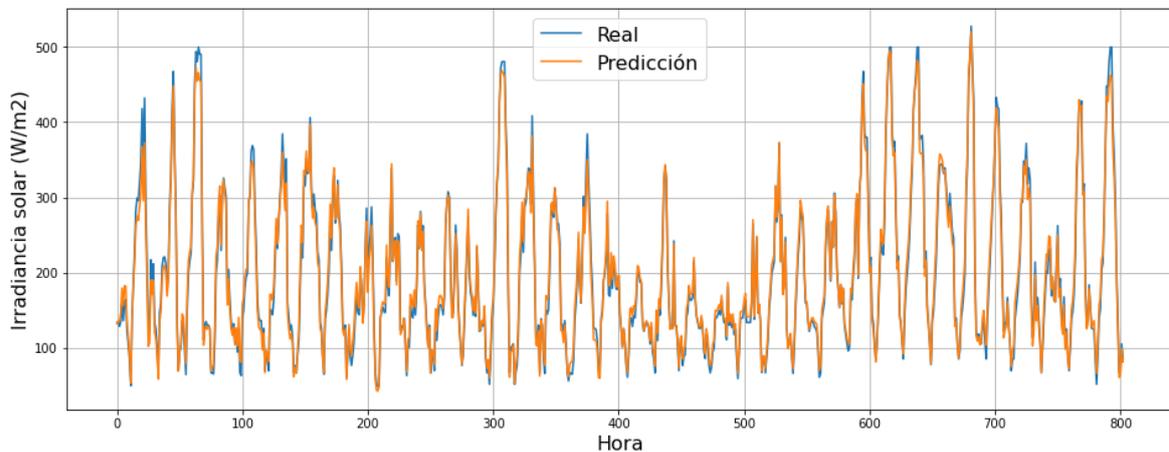


Figura 4.30. Comparativa de pronóstico y valores reales de irradiancia solar.

Finalmente, se realiza la evaluación de los pronósticos con respecto a los datos reales en función al puntaje de las métricas establecidas previamente mostrado en la Tabla 4.10, donde se puede observar una notable reducción del Error absoluto medio y el Error cuadrático medio en relación a la línea base, además el parentesco visual entre las dos series se puede justificar mediante el coeficiente de determinación 0.985 muy cercano a la unidad indicando un ajuste aceptable que puede ser empleado para la puesta en producción.

Tabla 4.10. Evaluación de modelo LSTM (one-step) para pronóstico de irradiancia.

Método	MAE	MSE	RMSE	R2
Holt-Winters	39,31	5790,39	76,10	0,836
LSTM (one-step)	10,082	155,081	12,453	0,985

4.2.5 Pronóstico de irradiancia solar por áreas

Previamente se ha revisado métodos para pronosticar series univariadas enfocadas en puntos geográficos en concreto, sin embargo, los datos de irradiancia disponibles abarcan un conjunto de puntos distribuidos en una matriz de 40 longitudes y 27 latitudes que representan un área, con la finalidad de

pronosticar el valor de irradiancia varios pasos en el futuro y considerar su interacción simultánea entre los puntos que conforman el espacio de estudio se implementará un modelo ConvLSTM para trabajar con imágenes generadas a partir de registros de irradiación siguiendo la metodología propuesta en la sección 3.4.3.

El objeto tipo array resultante de los datos disponibles es (328, 18, 27, 40, 1) que se puede interpretar como un conjunto de 328 secuencias, compuestas por imágenes en escala de grises de tamaño 27 x 40 pixeles en grupos de 18 frames por secuencia, en la Figura 4.31 se muestra una secuencia seleccionada de forma aleatoria dentro del conjunto de datos de entrenamiento para comprender su estructura.

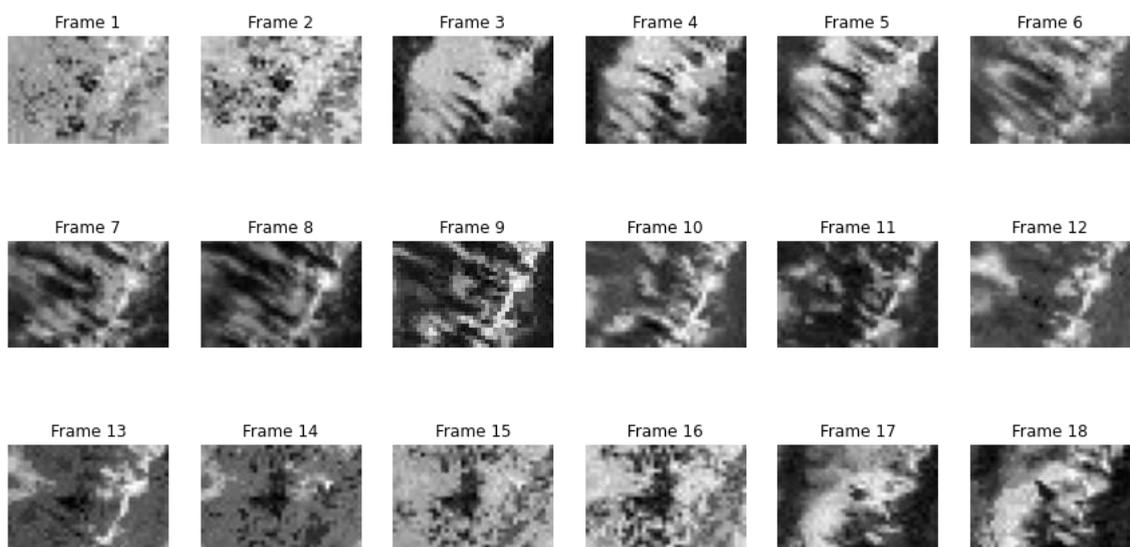


Figura 4.31. Comportamiento de irradiancia solar en secuencia Nro. 220.

Posteriormente, estos datos fueron preprocesados y almacenados en variables para entrenamiento y validación cuyas dimensiones son (262, 17, 27, 40, 1) y (66, 17, 27, 40, 1), respectivamente. Se puede notar la disminución en la cantidad de frames a 17, esto se debe al desplazamiento aplicado a los datos para simular el pronóstico en una unidad de tiempo en el futuro a partir de los datos iniciales.

Para construir el modelo LSTM convolucional, se empleará un conjunto apilado de capas ConvLSTM2D que admite entradas siguiendo la estructura (batch_size,

cantidad de fotogramas, ancho de imagen, alto de imagen, canales de color) y esta red devolverá una predicción (matriz) con las mismas dimensiones de las imágenes que conformar en conjunto de entrenamiento, la configuración del modelo final se muestra en la Tabla 4.11, cabe destacar que el modelo ha sido configurado de tal manera que pueda admitir secuencias de longitud variable.

Tabla 4.11. Configuración de modelo ConvLSTM.

Capas (Tipo)	Dimensiones de salida	Número de parámetros
conv_lstm2d (ConvLSTM2D)	(None, None, 27, 40, 40)	59200
batch_normalization (BatchNormalization)	(None, None, 27, 40, 40)	160
conv_lstm2d (ConvLSTM2D)	(None, None, 27, 40, 40)	115360
batch_normalization (BatchNormalization)	(None, None, 27, 40, 40)	160
conv_lstm2d (ConvLSTM2D)	(None, None, 27, 40, 40)	115360
batch_normalization (BatchNormalization)	(None, None, 27, 40, 40)	160
conv_lstm2d (ConvLSTM2D)	(None, None, 27, 40, 40)	115360
batch_normalization (BatchNormalization)	(None, None, 27, 40, 40)	160
conv3d (Conv3D)	(None, None, 27, 40, 1)	1081
Total		407.001

En la Figura 4.32 se muestra la estructura del modelo ConvLSTM aplicado, a la entrada se especifica las dimensiones del dataset de entrenamiento, estos datos deben ser centrados y normalizados para mejorar el entrenamiento de la red para ello se introduce una capa *BatchNormalization*, en las siguientes etapas se añade un capa de tipo *ConvLSTM* con un kernel de (3x3) empleado para determinar características espaciotemporales en cada secuencia, dentro del diagrama se muestra una línea punteada negra que indica la cantidad de frames procesados en cada bloque (17 frames), al conjunto se adiciona una capa *BatchNormalization* y se van apilando consecutivamente en un total de 4 capas *ConvLSTM* y *BatchNormalization*, posteriormente se implementa una capa de tipo Conv3D para la extracción de características y la salida muestra la predicción generada por el modelo con las mismas dimensiones que las imágenes del conjunto de entrenamiento (27x40x1).

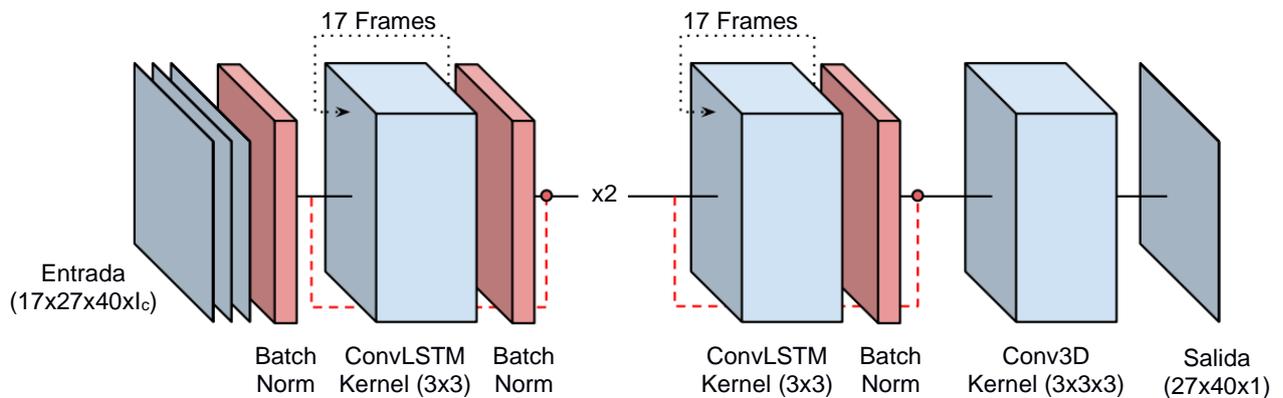


Figura 4.32. Estructura de modelo ConvLSTM para pronóstico de irradiación solar.

La métrica empleada para evaluación de la pérdida en el modelo es el error absoluto medio en conjunto con un optimizador de tipo Adam, en la Figura 4.33 se muestra las pérdidas de entrenamiento y validación de acuerdo a las diferentes épocas de entrenamiento, se puede visualizar que los datos de validación presentan una mejora en su ajuste a partir de la iteración 12 hasta equipararse con la pérdida en el entrenamiento en la época 40, a partir de este punto las respectivas métricas se van reduciendo simultáneamente hasta que deja de presentarse cambios durante el entrenamiento, culminando en la iteración 75 con una pérdida en entrenamiento de 0.0170 y la pérdida de validación de 0.0174.

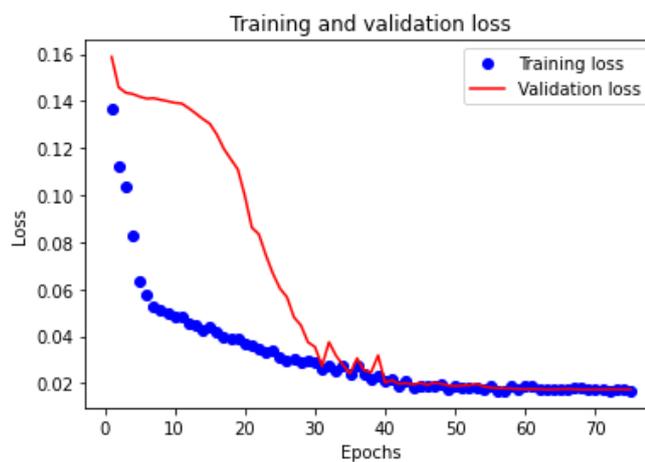


Figura 4.33. Gráfica de pérdida de entrenamiento y validación de modelo ConvLSTM.

Se realizaron pruebas de diferentes configuraciones siguiendo la estructura presentada en la Figura 4.32, donde se variaron las dimensiones de los kernel

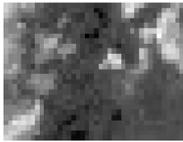
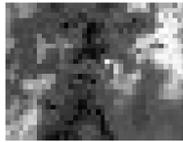
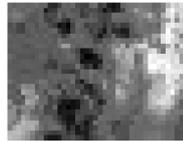
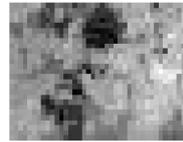
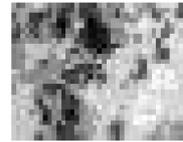
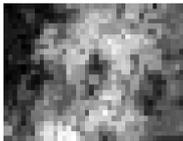
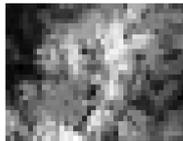
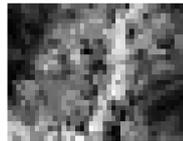
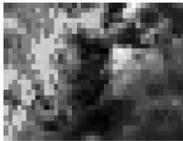
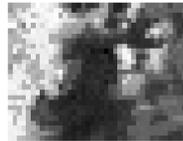
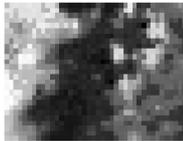
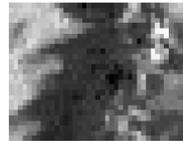
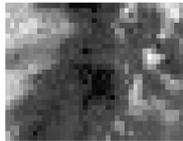
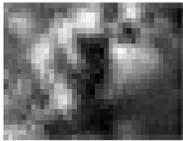
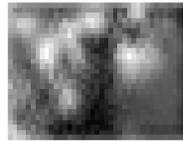
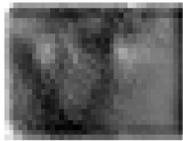
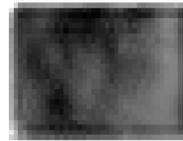
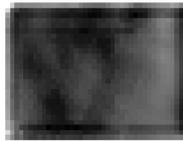
(*input-to-state* y *state-to-state*) para mejorar la extracción de características de las imágenes de entrada y la cantidad de capas Conv-LSTM apiladas consecutivamente, los pronósticos generados por los modelos fueron evaluados a partir del nivel de parentesco con respecto a los datos reales dentro del conjunto de testeo (imágenes), para ello la librería *sewar* permite comparar dos imágenes y calcular una métrica de similitud para este caso se empleará el error cuadrático medio (RMSE).

De todas las secuencias disponibles del conjunto de datos para testeo, se introducen en el modelo los 10 primeros fotogramas para generar pronósticos hasta 8 pasos en el futuro (cada paso equivale a 30 minutos), para este caso se realiza una predicción hasta 5 fotogramas en el futuro, es decir, desde las 12h30 hasta las 14h30, donde cada predicción es evaluada individualmente con respecto a los datos reales, la librería realiza el cálculo de similitud y almacenado en una lista en blanco, que posteriormente será promediada para dar una valoración del pronóstico.

En la Tabla 4.12 se puede visualizar los datos ingresados al modelo, la salida esperada y el pronóstico generado de forma artificial, cada fotograma de salida tiene su respectiva evaluación de acuerdo a la métrica RMSE, donde se puede notar que los primeros pasos mantienen una puntuación baja y su representación es cercana a la real, no obstante, en los siguientes pronósticos se puede evidenciar un desvanecimiento progresivo en la imagen de salida y aunque mantiene un puntaje bajo la representación espacial no es adecuada.

Posteriormente, las evaluaciones efectuadas a cada fotograma son promediados para calcular el rendimiento del modelo, este procedimiento debe ser replicado para las 36 secuencias disponibles.

Tabla 4.12. Evaluación de pronósticos generado por modelo ConvLSTM.

Datos de entrada (Frame 1 – Frame 10)				
Frame 1	Frame 2	Frame 3	Frame 4	Frame 5
				
Frame 6	Frame 7	Frame 8	Frame 9	Frame 10
				
Salida (Frame 11 – Frame 15)				
Real:				
Frame 11	Frame 12	Frame 13	Frame 14	Frame 15
				
Pronóstico:				
Frame 11	Frame 12	Frame 13	Frame 14	Frame 15
				
RMSE Frame 11	RMSE Frame 12	RMSE Frame 13	RMSE Frame 14	RMSE Frame 15
0.11786	0.13706	0.13684	0.10082	0.08104

Se realizaron pruebas con una gran variedad de configuraciones de modelos y los que presentaron los mejores resultados en la etapa de entrenamiento, validación y testeo, se encuentran resumidos en la Tabla 4.13 donde se muestra el detalle de kernels, capas LSTM, número de parámetros entrenados y su evaluación de pronóstico, donde se puede ver que el modelo con el mejor rendimiento tiene un puntaje medio RMSE de 0.12443.

Tabla 4.13. Comparación de pronóstico entre modelos ConvLSTM.

Modelo	Input-to-state (Kernel)	State-to-state (Kernel)	Capas apiladas (LSTM + B.N.)	Número de parámetros	RMSE
ConvLSTM(3x3)-3x3	(3x3)	3x3	4	407.001	0.12443
ConvLSTM(2x2)-2x2	(2x2)	2x2	5	233.721	0.14549
ConvLSTM(2x2)-2x2	(2x2)	2x2	4	182.201	0.12462

Una de las ventajas del empleo de la red ConvLSTM es su capacidad para relacionar los patrones espaciotemporales debido a la incorporación de la

estructura no lineal y convolucional de la red. Asimismo, a partir del paso 3 del pronóstico (frame 13) se puede visualizar un efecto de desvanecimiento a medida que las predicciones avanzan en el tiempo como se muestra en la Figura 4.34, de acuerdo a Shi et al. (2015) se hace referencia a este comportamiento de la red LSTM convolucional debido a la propagación de los errores que hacen ineficiente a la predicción a largo plazo, también puede deberse a las incertidumbres propias de los datos de tipo meteorológico que imposibilitan obtener predicciones precisas (nítidas).

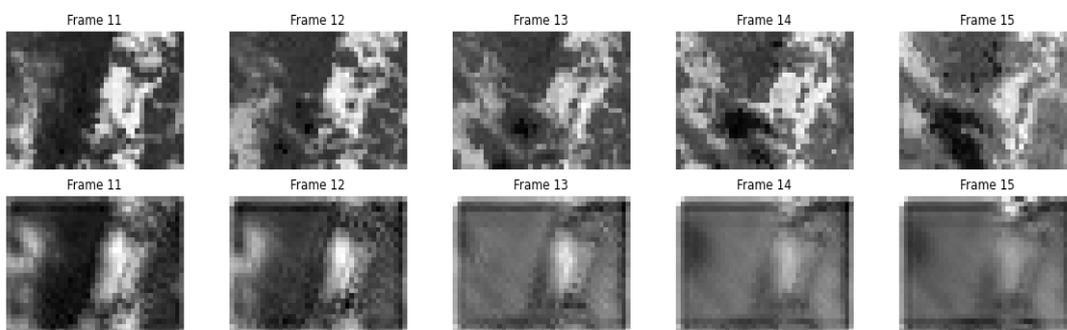


Figura 4.34. Pronóstico de irradiancia solar mediante Conv-LSTM.

El método de pronóstico de irradiancia por áreas presenta un buen rendimiento a corto plazo, de acuerdo a las evaluaciones realizadas con un margen aceptable de 2 pasos en el futuro (60 minutos), mientras que los pronósticos a largo plazo obtenidos por el modelo no son fiables debido a propagación de errores a través de la red, para el desarrollo del presente proyecto se requiere conocer información con un amplio margen de tiempo para la toma de decisiones, por lo que la implementación de una solución basada en deep-learning no es viable, no obstante, puede ser de interés para otros campos de estudio o aplicaciones afines.

4.2.6 Selección de modelo de pronóstico

Con la finalidad de generar pronósticos fiables del nivel irradiancia solar en el área comprendida en la provincia de Pichincha, se han realizado pruebas tanto con métodos estadísticos como modelos de Deep Learning, donde se han podido determinar las fortalezas y falencias de cada alternativa estudiada para su

posterior selección de la solución adecuada para su implementación en una aplicación enfocada al diseño de sistemas fotovoltaicos.

Teniendo en cuenta los resultados obtenidos a partir de las evaluaciones mediante las diferentes métricas consideradas, los mejores puntajes fueron obtenidos del modelo LSTM para series univariadas debido a su capacidad de representación del comportamiento de la irradiancia solar es aceptable, además el costo computacional en conjunto con los tiempos de entrenamiento del modelo son reducidos debido a la rápida convergencia de la red para adaptarse a los datos de entrada, siendo adecuado para los fines del presente proyecto.

4.3 Pruebas de funcionalidad

Una vez que se ha seleccionado la metodología para generar el pronóstico de irradiación solar y tomando como referencia el diseño base presentado en la sección 3.6 se ha desarrollado la aplicación web enfocada a la presentación de resultados al usuario, en la Figura 4.35 se muestran los componentes básicos de la página de inicio de la aplicación donde se pueden identificar los siguientes componentes:

- 1) Encabezado y descripción de la aplicación.
- 2) Panel de navegación.
- 3) Pantalla de ejecución de aplicación.

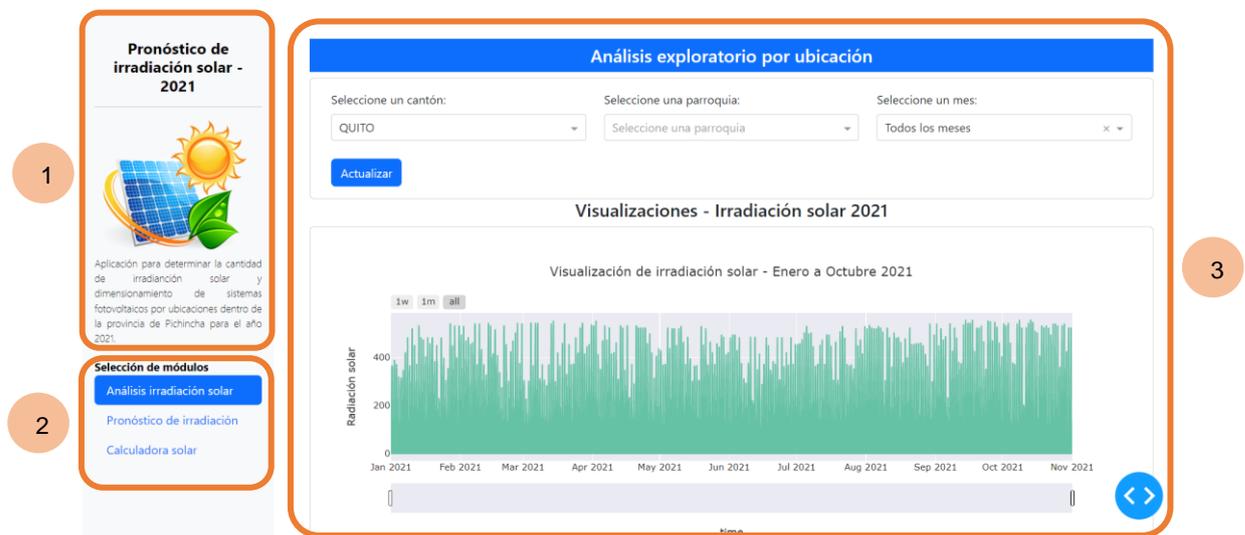


Figura 4.35. Componentes de aplicación de pronóstico de irradiación solar.

Por defecto la aplicación se inicializará mostrando la herramienta “Análisis de irradiación solar”, sin embargo, el usuario puede seleccionar entre las funciones integradas a través del panel de navegación donde se encuentran las siguientes herramientas:

- 1) **Análisis de irradiación solar:** Mediante esta herramienta el usuario podrá conocer el comportamiento de la irradiancia solar en una zona previamente definida, además, se integran elementos visuales que proporcionan información a detalle y resúmenes estadísticos de los registros de luz solar en la zona de interés.
- 2) **Pronóstico de irradiación:** Esta herramienta proporciona la capacidad de generar pronósticos de irradiancia solar a partir de los datos base de una zona específica, el modelo de inteligencia artificial es empleado para generar predicciones hasta 60 días en el futuro cuyos resultados son mostrados al usuario a través de un conjunto de gráficos, estadísticas visuales, tarjetas y tablas que pueden ser tomados de referencia para el dimensionamiento de sistemas fotovoltaicos.
- 3) **Calculadora solar:** Los datos de irradiancia y pronósticos son integrados en esta herramienta permitiendo al usuario detallar el consumo energético requerido, especificaciones técnicas de paneles solares y otros componentes para dimensionar un sistema fotovoltaico adecuado a sus necesidades.

4.3.1 Funcionalidad de módulo “Análisis de irradiación solar”

Por defecto la aplicación se inicializará mostrando el módulo “Análisis de irradiación solar” o a su vez, se puede acceder a la herramienta a través del panel de navegación, donde se puede observar la interfaz mostrada en la Figura 4.36 con los siguientes elementos:

- 1) Panel de búsqueda de zona de interés y temporalidad de análisis.
- 2) Panel de visualizaciones.



Figura 4.36. Componentes de módulo de “Análisis de irradiación solar”.

Para determinar la ubicación se dispone de un conjunto de menús desplegados que contiene la información de los 8 cantones que conforman la provincia de Pichincha, al hacer una selección los datos se actualizan a nivel de parroquias en función del cantón correspondiente, asimismo, se puede establecer el periodo de análisis en función de un mes en concreto o por defecto, visualizar todos los meses que se encuentran disponibles en la base de datos, para este proyecto se considera que los datos disponibles forman parte del conjunto de entrenamiento (enero a octubre de 2021).

Una vez que se tiene definida la zona de interés y el intervalo de tiempo para el análisis, se presentan un conjunto de elementos visuales donde se pueden estudiar diversos aspectos de la irradiancia solar, tales como:

- a) **Serie de tiempo de irradiancia:** Muestra todos los registros disponibles de la base de datos presentados a manera de una serie temporal, se dispone de los registros de irradiancia calculados cada 30 minutos y el usuario puede definir la vista por semanas (1W), mes (1M) o toda la serie (ALL) a través de los botones integrados y desplazar su selección a través de un slider como se muestra en la Figura 4.37 (a).

- b) Estadística de irradiancia solar por horas:** La cantidad de luz solar que recibe un punto geográfico es dependiente de la rotación e inclinación del planeta y es necesario identificar los horarios en la cual la luz incidente es apta para la producción energética, para ello se muestra un resumen estadístico mediante diagramas de cajas de la luz solar por horas que puede ser analizado en la totalidad de la serie de tiempo o de acuerdo a un mes en concreto dentro del dataset, como se muestra en la Figura 4.37 (b).
- c) Distribución de datos:** Otra consideración importante es conocer la distribución de los registros de irradiancia solar de forma mensual, para ello se hace empleo del diagrama de violín que permite visualizar la acumulación de los datos y en conjunto con el diagrama de caja para determinar puntos estadísticos de importancia que se experimentaron durante un mes o en la totalidad de la serie como se muestra en la Figura 4.37 (c).
- d) Irradiación promedio mensual:** Para el diseño de sistemas fotovoltaicos se requiere conocer los meses críticos en la cual la irradiación solar es baja, esta información se obtiene a partir del cálculo del área bajo la curva de la serie de irradiancia solar, estos resultados son presentados al usuario mediante un gráfico de barras correspondiente a cada mes como se muestra en la Figura 4.37 (d).
- e) Heatmap de irradiancia solar:** Mediante un mapa de calor se puede obtener una perspectiva de la irradiancia solar promediada por horas en los diferentes meses del año que se encuentran disponibles en el dataset como se muestra en la Figura 4.37 (e) con la finalidad de identificar los rangos horarios propicios para la generación energética y a su vez, los límites en el día donde la producción puede verse afectada debido a la escasez de luz solar.

Finalmente, cada componente visual se ha integrado en la aplicación web utilizando una estructura conformada por columnas y filas, de tal forma que los gráficos puedan organizarse fácilmente y cuidar las proporciones de las dimensiones para que los datos sean legibles para el usuario, la distribución empleada se muestra en la Figura 4.37.

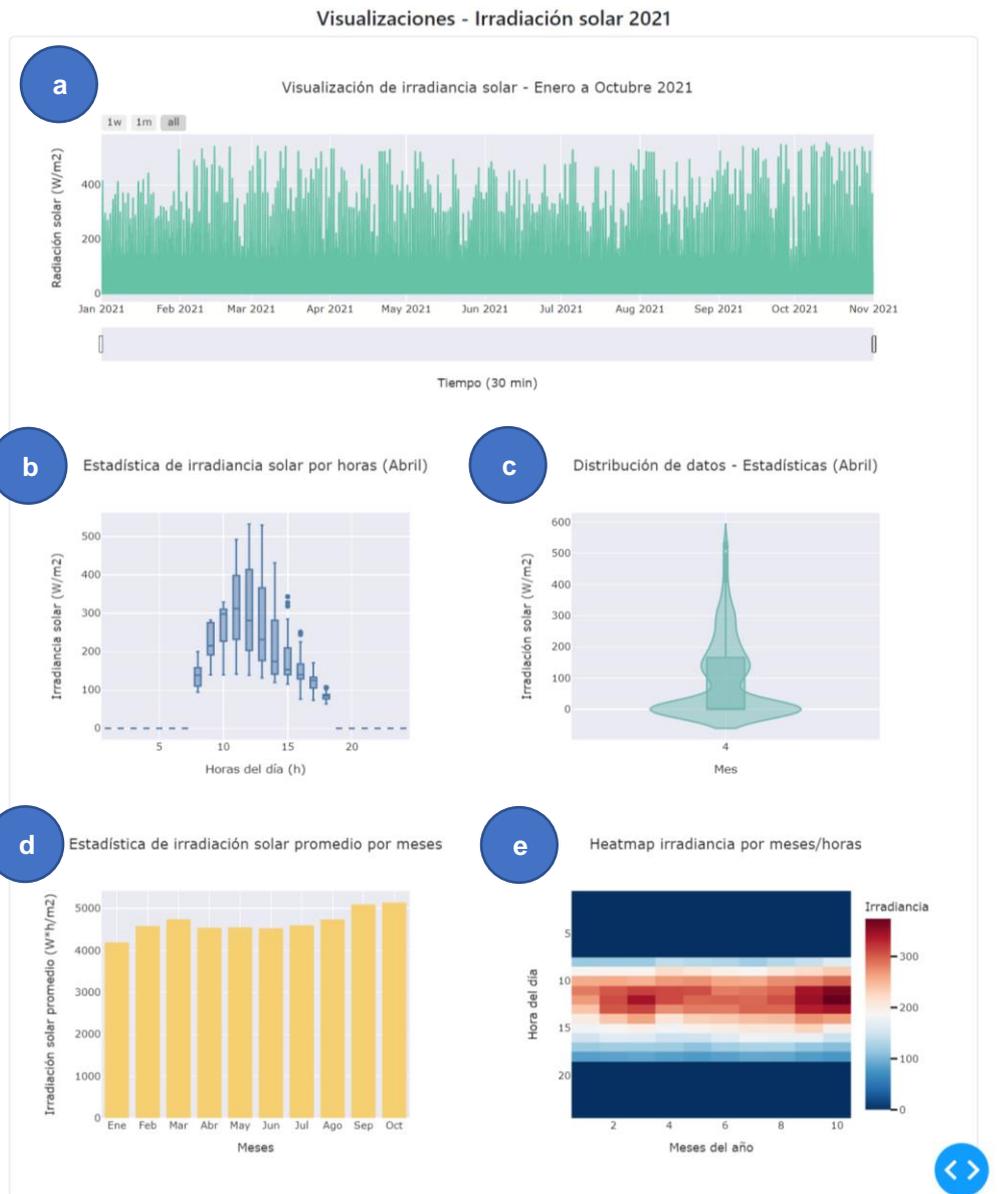


Figura 4.37. Distribución de elementos visuales en herramienta de “Análisis de irradiación solar”.

4.3.2 Funcionalidad de módulo “Pronóstico de irradiación”

Para acceder al módulo de pronóstico es necesario seguir el enlace “Pronóstico de irradiación” ubicado en el panel de navegación, donde se mostrará una nueva pantalla donde el usuario podrá seleccionar la zona de interés siguiendo la misma estructura mostrada en el módulo anterior y un slider donde se puede definir la cantidad de días a pronosticar, por defecto se encuentra marcado 20 días.

Dentro de la interfaz se puede visualizar el botón “Generar pronóstico” que permitirá enlazar los datos de irradiación con el modelo de Deep Learning y posteriormente generar un pronóstico por pasos únicamente en la franja de irradiancia mayor a 50 W/m², hay que tener en cuenta que cada paso es considerado cada 30 minutos, es decir, para un pronóstico de 20 días a futuro se requerirán aproximadamente 480 predicciones consecutivas, por lo que el módulo requerirá un tiempo de espera hasta completar la tarea, mientras el modelo se encuentre en ejecución el botón “Mostrar resultados” se encontrará inhabilitado.

Cuando las predicciones se encuentran disponibles se presenta una alerta con el mensaje: “¡Los datos y el modelo se han cargado correctamente!” y posteriormente se podrá habilitar el botón “Mostrar resultados” que actualizará las tarjetas y gráficos de acuerdo a la cantidad de días que se requiera realizar la predicción como se muestra en la Figura 4.38, esta información será tomada desde el slider y su valor puede ser modificado sin requerir de un nuevo entrenamiento para actualizar los elementos visuales.



Figura 4.38. Componentes del panel de pronóstico de irradiación solar.

Para la visualización de los pronósticos se ha segmentado el área de visualización mediante secciones de acuerdo al tipo del objeto visual, métricas simples como promedios, valores mínimos y máximos se encuentran detallados en un formato de tarjetas, el promedio diario de irradiación puede obtenerse a detalle mediante tablas y finalmente, gráficos con información estadística que se encuentra distribuida como se detalla a continuación:

- a) Pronóstico de irradiancia solar:** Los pronósticos generados por el modelo son integrados en un dataframe de tal forma que se pueden visualizar como una serie de tiempo (naranja) y los datos reales que forman parte del conjunto de testeo se muestran en la misma gráfica (azul) donde se puede visualizar el ajuste del modelo empleado como se muestra en la Figura 4.39 (a), sin embargo, para datos actualizados donde no exista un conjunto de pruebas se representaría sólo la serie correspondiente al pronóstico.
- b) Estadística de irradiación pronosticada:** Se realiza el cálculo de irradiación a partir de los datos pronosticados cuyo promedio mensual se representa mediante un gráfico de barras con la finalidad de identificar la existencia de valores de irradiación crítico, el elemento se muestra en la Figura 4.39 (b).
- c) Irradiación solar diaria:** Para tener una percepción detallada del comportamiento de la luz solar, se muestra la irradiación solar de forma diaria en la Figura 4.39 (c) donde se puede tomar con antelación las fechas que pudiesen proporcionar una menor producción energética debida a su baja cantidad de irradiación solar, este tipo de información es relevante para el suministro de electricidad en urbes donde exista una dependencia a partir de energía solar que puede ser equilibrada mediante fuentes de energía alternativas.
- d) Histograma de irradiación:** El aporte solar se encuentra representado mediante un histograma en la Figura 4.39 (d) para visualizar la tendencia de luz solar disponible dentro del periodo de pronóstico, identificar la cantidad de días con baja irradiación solar.
- e) Irradiancia por horas:** Este tipo de visualización ha sido reutilizada del módulo anterior, con la finalidad de analizar el comportamiento de la luz solar mediante un conjunto de diagramas de caja distribuidos por cada hora del día como se muestra en la Figura 4.39 (e), de esta manera se puede analizar parámetros estadísticos base proporcionados a partir de las predicciones del modelo de Deep Learning.

Conjunto de tarjetas de información

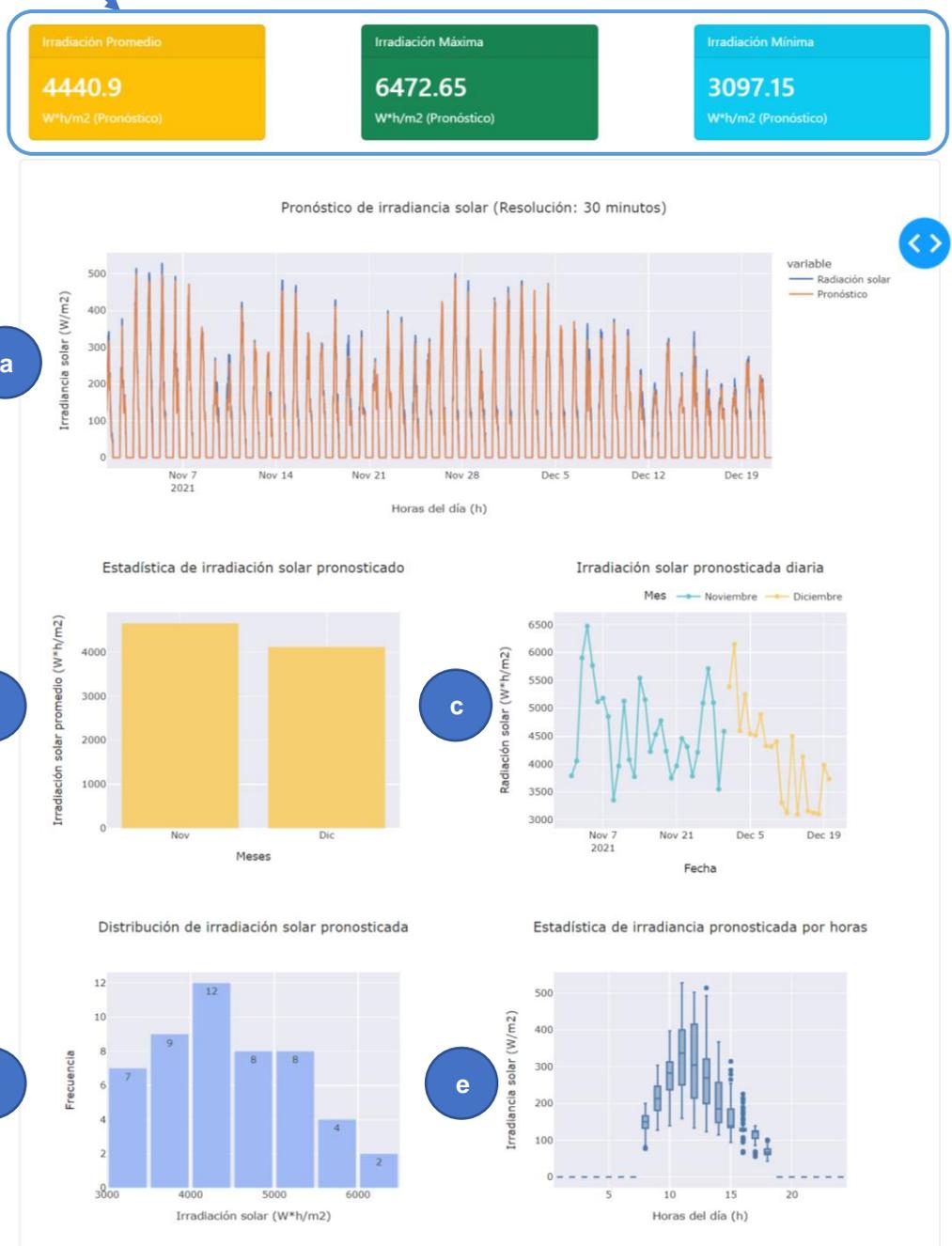


Figura 4.39. Distribución de elementos visuales en herramienta de “Pronóstico de irradiancia”.

4.3.3 Funcionalidad de módulo “Calculadora solar”

Las herramientas revisadas previamente permiten al usuario conocer el comportamiento de la irradiancia e irradiación solar dentro de una zona de interés, generar pronósticos a corto plazo para identificar un bajo aporte solar mediante la

incorporación de elementos visuales con estadísticas de los datos para el respectivo análisis de las series de tiempo de las bases de datos y la información generada artificialmente, sin embargo, el propósito del presente proyecto es integrar ambas funcionalidades para dar soporte al diseño de sistemas fotovoltaicos aislados para ello se ha diseñado un módulo enfocado para este fin, se utilizará el criterio de dimensionamiento del mes crítico.

Para acceder al módulo se debe seleccionar el enlace “Calculadora solar” dentro del panel de navegación, se desplegará la pantalla inicial de la aplicación donde se pueden identificar los siguientes componentes, como se muestra la Figura 4.40.

- Panel de selección de ubicación.
- Selección del tipo de consumo.
- Definición de autonomía de sistema expresado en días.
- Panel desplegable para definición de consumo en corriente alterna (110V).
- Panel desplegable para definición de consumo en corriente continua (12 o 24V).
- Panel de especificaciones técnicas de voltajes y baterías.
- Panel de especificaciones técnicas de panel solar.

Figura 4.40. Componentes de panel de “Calculadora solar”

Para poder dimensionar adecuadamente un sistema fotovoltaico se debe especificar la ubicación geográfica, el consumo previsto, características eléctricas de la instalación y las especificaciones técnicas de los módulos solares y baterías que el usuario debe detallar de acuerdo a su necesidad, para el caso de datos técnicos (paneles y acumuladores) se han considerado por defecto características de equipos estándar para facilitar el cálculo del sistema.

Los paneles de consumos AC y DC pueden ser desplegados para mostrar el tipo de consumo de acuerdo a la cantidad de componentes, su potencia nominal y tiempo de uso previsto que deben ser especificados a detalle para que el cálculo del sistema fotovoltaico pueda satisfacer la demanda del usuario, cabe destacar que los consumos DC corresponden únicamente a luminarias debido a su compatibilidad con la tensión de salida del sistema solar y evitar pérdidas al usar luminarias convencionales AC, en la Figura 4.41 se muestra los campos correspondientes a consumos energéticos que se encuentran disponibles en la aplicación.

Detalle de consumos en Corriente Alterna
^

Ingrese los consumos energéticos AC de acuerdo a su necesidad:

Detalle de carga	Cantidad (#)	Potencia (W)	Tiempo de uso (h)
Ordenador	<input type="text" value="1"/>	<input type="text" value="90"/>	<input type="text" value="2"/>
Lavadora	<input type="text" value="1"/>	<input type="text" value="1720"/>	<input type="text" value="0,5"/>
Televisor	<input type="text" value="1"/>	<input type="text" value="70"/>	<input type="text" value="4"/>
Refrigerador	<input type="text" value="1"/>	<input type="text" value="20"/>	<input type="text" value="24"/>
Otros dispositivos	<input type="text" value="1"/>	<input type="text" value="300"/>	<input type="text" value="1"/>

Detalle de consumos en Corriente Continua
^

Ingrese los consumos energéticos DC de acuerdo a su necesidad:

Detalle de carga	Cantidad (#)	Potencia (W)	Tiempo de uso (h)
Iluminación de salón	<input type="text" value="1"/>	<input type="text" value="60"/>	<input type="text" value="5"/>
Iluminación de cocina	<input type="text" value="1"/>	<input type="text" value="50"/>	<input type="text" value="3"/>
Iluminación de baños	<input type="text" value="1"/>	<input type="text" value="60"/>	<input type="text" value="2"/>
Iluminación de pasillos	<input type="text" value="1"/>	<input type="text" value="30"/>	<input type="text" value="1"/>
Iluminación de dormitorios	<input type="text" value="1"/>	<input type="text" value="50"/>	<input type="text" value="4"/>

Figura 4.41. Panel de detalle de consumos AC y DC.

El cálculo del sistema fotovoltaico mostrará al usuario información de diseño como características de irradiación, generación eléctrica, rendimiento esperado, etc. de acuerdo a los componentes visuales empleados se detalla su respectiva interpretación:

- a) Ángulo de inclinación:** Para optimizar la generación eléctrica a partir de módulos solares es necesario identificar su inclinación con respecto a la superficie, esta información dependerá de la ubicación geográfica de la instalación y el uso previsto de acuerdo a la época del año, la aplicación recomendará la configuración en función de los datos establecidos previamente.
- b) Selección de inversor:** En función de la cantidad de consumos de corriente alterna se aplicará un factor de demanda para asegurar el correcto funcionamiento de los dispositivos conectados a la red, se recomienda emplear un inversor con una potencia nominal aproximada (valor comercial) mayor o igual al recomendado por la aplicación.
- c) Selección de regulador:** Para controlar los ciclos de carga y descarga del sistema de baterías este componente debe ser seleccionado en función de la corriente máxima que puede proporcionar un generador fotovoltaico, se recomienda el empleo de un regulador con una corriente nominal mayor a la calculada.
- d) Detalle de consumo:** En función de la cantidad de cargas AC y DC se presenta al usuario la proporción con respecto al consumo total, estos datos son considerados durante el dimensionamiento del sistema fotovoltaico.
- e) Dimensionamiento de generador solar:** La aplicación tomará como referencia la irradiación del mes crítico de la ubicación de interés, en función a los consumos se calculará la cantidad de paneles solares requeridos para satisfacer la demanda energética definida por el usuario, asimismo, se muestra la configuración recomendada a ser implementada bajo la nomenclatura #S, #P (módulos conectados en serie, módulos conectados en paralelo), de forma gráfica se presenta la máxima producción a partir del sistema fotovoltaico con respecto al consumo energético diario.
- f) Dimensionamiento de baterías:** Esta característica dependerá de el tiempo de uso de las cargas y la autonomía esperada del sistema (expresada en días), la

aplicación calculará la capacidad en Ah requerida y tomando las especificaciones técnicas para baterías se calculará la configuración del sistema de acumulación (nro. baterías conectadas en serie, nro. baterías conectadas en paralelo), gráficamente se representa la capacidad máxima del sistema con respecto a la capacidad requerida.

- g) Panel de gráficos de irradiación y generación eléctrica:** Este panel consta de un conjunto de cuatro visualizaciones donde se presentan la irradiación solar promedio mensual calculada a partir de la base de datos y pronósticos, potencia eléctrica generada por meses con respecto al mes crítico y el detalle del consumo y generación eléctrica.
- h) Producción energética:** Se muestra el total de energía eléctrica anual que se puede generar en la zona de interés.
- i) Ahorro estimado:** A partir de la tarifa del servicio eléctrico dispuesta por la Agencia de Regulación y Control de Energía y Recursos Naturales no Renovables (ARCERNNR) en abril de 2022 y ejecutada por Empresa Eléctrica Quito (EEQ) en la provincia de Pichincha de 9.2 centavos por cada Kilovatio-hora (ARCERNNR, 2022), se calcula el ahorro monetario que representa el uso de energía solar.
- j) Reducción de emisión de CO₂:** Mediante el uso de energías renovables cada kilovatio-hora producido evita la emisión de cierta cantidad de contaminantes liberados al ambiente (CO₂) provenientes de fuentes de energía convencionales, este factor puede ser traducido su equivalente en árboles plantados que indican la cantidad de ejemplares necesarios para absorber dicha cantidad de CO₂ utilizando como referencia la vida útil estimada de un sistema fotovoltaico de aproximadamente 25 años (La corriente, 2022).



Figura 4.42. Componentes de módulo “Calculadora solar”.

4.4 Análisis costo/beneficio

4.4.1 Costos

Para la implementación y puesta en marcha de la aplicación web desarrollada se requerirá de la inversión en una serie de servicios necesarios para que la aplicación

pueda presentarse al usuario final, además de requerir servicios adicionales para el almacenamiento de la base de datos y disponer de herramientas necesarias para mantener el modelo de inteligencia artificial actualizado, los costos estimados de implementación se detallan en la Tabla 4.14.

Tabla 4.14. Costos de implementación.

Servicio	Costo por año
Dominio web	\$20,00 / año
Hospedaje web	\$120,00 / año
Diseño web	\$ 0,00
Almacenamiento AWS S3-standard	\$ 60,00 / año
Amazon SageMaker Instancia: ml.g4dn.4xlarge CPU: 16 Memoria: 64 GB GPU: 1 Almacenamiento: 225 GB	\$ 1.084,80 / año
Total	\$ 1.284,80 / año

4.4.2 Beneficios

Mediante la implementación de una herramienta para el análisis de irradiación solar y cálculo de sistemas fotovoltaicos, se pretende impulsar el empleo de energías renovables a pequeña y gran escala permitiendo al usuario dimensionar de forma rápida y sencilla un sistema fotovoltaico a partir de su demanda energética, utilizando como referencia datos actualizados y pronósticos generados a través de modelos de inteligencia artificial.

Disponer de una vista previa de la cantidad de componentes necesarios para iniciar un proyecto solar (módulos solares, reguladores, inversores, baterías, etc.) facilitará la búsqueda de alternativas en el mercado local o internacional, eliminando el riesgo que el diseño no pueda suministrar la energía requerida debido al uso de fuentes de datos desactualizadas, se elimina la dependencia de paquetes informáticos de pago que en su mayoría no disponen de una fuente de datos fiable dentro del territorio ecuatoriano.

Otro aspecto importante a tener en cuenta es la aplicabilidad de datos de irradiación solar y sus pronósticos que pueden ser empleados en campos diferentes al sector energético como en la agricultura, donde la irradiación solar juega un rol fundamental en el control de crecimiento de vegetales bajo invernaderos, permitiendo suministrar recursos de forma eficiente a los cultivos con la finalidad de optimizar tiempos de producción.

Finalmente, dentro del ámbito académico el desarrollo de sistemas automatizados a pequeña escala que necesitan funcionar de forma ininterrumpida o es difícil tener acceso a una fuente de alimentación convencional, la principal alternativa son los sistemas fotovoltaicos que requieren ser estudiados a detalle para asegurar un funcionamiento adecuado de los dispositivos electrónicos como son los casos de sistemas aéreos no tripulados de alta duración, sistemas de monitoreo y alimentadores automáticos para camarón, plataformas terrestres, señalética, entre otros.

CAPÍTULO 5

5. CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- El empleo de modelos de Deep Learning para el pronóstico de series de tiempo de irradiancia solar proporcionan una aproximación muy cercana a los valores reales experimentados a corto plazo (60 días), los modelos resultantes pueden ser integrados en herramientas web para el desarrollo de aplicativos para el análisis de la luz solar y el dimensionamiento de sistemas fotovoltaicos para diferentes ubicaciones dentro de la provincia de Pichincha.
- La ejecución de mantenimientos preventivos de forma periódica en los equipos de medición de irradiancia solar y otras variables ambientales en las estaciones meteorológicas determinará la calidad, confiabilidad de los datos y delimitará el área donde se pueda desarrollar aplicaciones de ciencia de datos.
- Se construyó una base de datos de irradiancia solar a través de la integración registros captados localmente a través de estaciones meteorológicas del Fondo para la protección del agua (FONAG) y los registros satelitales del Laboratorio Nacional de Energía Renovable (NREL), mediante el empleo del modelo LightGBM para regresión permitiendo obtener un conjunto de datos preciso, continuo en el tiempo, con una resolución espacial de 4 Km para datos de irradiancia solar en la provincia de Pichincha para el año 2021 y de fácil acceso mediante el empleo del formato netCDF especializado para el almacenamiento de variables multidimensionales.
- Con la finalidad de generar pronósticos se implementó el modelo LSTM one-step para generar predicciones a partir de series univariadas tomando como referencia los datos asociados a un punto coordinado dentro del conjunto de datos, obteniendo puntajes sobresalientes en las métricas de evaluación MAE (10.08), MSE (155.08), RMSE (12.45) y R2 (0.98) permitiendo obtener predicciones precisas hasta 60 días en el futuro. Asimismo, se implementó el modelo Conv-LSTM para generar pronósticos utilizando de forma simultánea todos los registros de los puntos coordinados que conforman la provincia de Pichincha mediante el

formato de imágenes, la mejor evaluación obtenida fue 0.124 empleando la métrica RMSE permitiendo obtener predicciones satisfactorias a gran escala hasta 60 minutos en el futuro, para el desarrollo de la aplicación en el presente proyecto se utilizó el modelo generado mediante redes LSTM one-step.

- La interfaz interactiva para visualización, generación de pronósticos y cálculo de sistemas fotovoltaicos se encuentra desarrollado mediante Dash debido a su capacidad de incorporar visualizaciones a partir de dataframes y mostrarlos en un entorno HTML a través de un servidor web que se ejecuta de forma local, de esta manera se puede diseñar cada componente de la aplicación utilizando únicamente Python como lenguaje de programación y en conjunto con Bootstrap permite al científico de datos personalizar cada elemento web mediante la incorporación de plantillas y temas prediseñados.

Recomendaciones

- Verificar la calidad de los datos de las estaciones en tierra previo al ajuste con los datos satelitales, debido a que los sensores pueden encontrarse calibrados, pero presentar un desfase con respecto al tiempo real de las mediciones, lo que puede ocasionar problemas durante el entrenamiento de los modelos de regresión.
- Para el entrenamiento de modelos LSTM empleando datos de irradiancia solar en series de tiempo univariadas y por áreas, se recomienda eliminar los ciclos nocturnos para evitar el sesgo en los datos que disminuyen considerablemente el rendimiento del modelo.
- Para implementación del modelo LSTM desarrollado en el presente proyecto se recomienda el empleo de una tarjeta gráfica integrada para optimizar el proceso de cálculo de las predicciones, debido a que se requieren calcular aproximadamente 2.880 puntos correspondientes a 60 días pronosticados considerando una resolución temporal de 30 minutos.
- La implementación de redes conv-LSTM para pronóstico de irradiancia solar presentaron resultados satisfactorios a corto plazo (60 minutos), pueden ser empleados para prever la producción energética en plantas solares que abarcan grandes superficies de terreno con la finalidad de generar alertas tempranas para

incorporar suministros auxiliares de electricidad para satisfacer la demanda energética en horas pico, pudiéndose desarrollar herramientas a partir de la implementación de este tipo de redes.

- Mediante el almacenamiento de registros de irradiancia solar a través de ficheros netCDF se pueden implementar estrategias de segmentación basados en inteligencia artificial con la finalidad de identificar zonas con un elevado potencial de luz solar, facilitando la identificación de puntos estratégicos para la implementación de proyectos fotovoltaicos de alto rendimiento.

BIBLIOGRAFÍA

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* (arXiv:1907.10902). arXiv. <https://doi.org/10.48550/arXiv.1907.10902>
- Amat, J. (2021, febrero). *Skforecast: Time series forecasting with Python and Scikit-learn*. <https://www.cienciadedatos.net/documentos/py27-time-series-forecasting-python-scikitlearn.html>
- ARC. (2020). *Atlas del sector eléctrico ecuatoriano*. Agencia de Regulación y Control de Energía y Recursos Naturales No Renovables. <https://www.controlrecursosyenergia.gob.ec/wp-content/uploads/downloads/2021/06/Atlas-2020-baja.pdf>
- ARCERNNR. (2022, mayo 10). *Las tarifas de energía eléctrica no se incrementarán en el 2022 – Ministerio de Energía y Minas*. Recursos y energía. <https://www.recursosyenergia.gob.ec/las-tarifas-de-energia-electrica-no-se-incrementaran-en-el-2022/>
- Bajaj, A. (2022, julio 21). *Anomaly Detection in Time Series*. neptune.ai. <https://neptune.ai/blog/anomaly-detection-in-time-series>
- Barrios, J. (2022, junio 14). Light GBM vs XGBoost. ¿Cuál es mejor el algoritmo? *Juan Barrios*. <https://www.juanbarrios.com/light-gbm-vs-xgboost-cual-es-mejor-algoritmo/>
- Brownlee, J. (2017, abril 6). Time Series Forecasting with the Long Short-Term Memory Network in Python. *MachineLearningMastery.com*. <https://machinelearningmastery.com/time-series-forecasting-long-short-term-memory-network-python/>

- Cantos, J. (2016). *Configuración de instalaciones solares fotovoltaicas*. Ediciones Paraninfo, S.A.
- Cebecauer, T., Šúri, M., & Perez, R. (2010). *HIGH PERFORMANCE MSG SATELLITE MODEL FOR OPERATIONAL SOLAR ENERGY APPLICATIONS*. 5.
- Corcobado, T. D., & Rubio, G. C. (2010). *Instalaciones solares fotovoltaicas*. McGraw-Hill Interamericana.
- FONAG. (2022). *Anuario hidrometeorológico 2021*. https://www.fonag.org.ec/web/wp-content/uploads/2022/06/Anuario-2021_Final.pdf
- Fouilloy, A., Voyant, C., Notton, G., Motte, F., Paoli, C., Nivet, M.-L., Guillot, E., & Duchaud, J.-L. (2018). Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. *Energy*, 165, 620-629. <https://doi.org/10.1016/j.energy.2018.09.116>
- GeoSolutions. (2022). *NetCDF-family serving basics*. https://docs.geoserver.geosolutions.it/edu/en/multidim/netcdf/netcdf_basics.html
- Gil, A. (2021, marzo 9). *Los grandes productores de energía solar en el mundo*. El Orden Mundial - EOM. <https://elordenmundial.com/mapas-y-graficos/grandes-productores-energia-solar-mundo/>
- Graves, A. (2014). *Generating Sequences With Recurrent Neural Networks* (arXiv:1308.0850). arXiv. <http://arxiv.org/abs/1308.0850>
- Hanke, J. E., & Wichern, D. W. (2006). *Pronósticos en los negocios*. Pearson Educación.
- Heras, M. (2018, diciembre 28). *Error Cuadrático Medio para Regresión*. <https://www.iartificial.net/error-cuadratico-medio-para-regresion/>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

- Karani, D. (2022, julio 21). *Optuna Guide: How to Monitor Hyper-Parameter Optimization Runs*. neptune.ai. <https://neptune.ai/blog/optuna-guide-how-to-monitor-hyper-parameter-optimization-runs>
- Kayri, M., Kayri, İ., & Gencoglu, M. (2017). *The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data*. 1-4. <https://doi.org/10.1109/EMES.2017.7980368>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
- La corriente. (2022, marzo 9). ¿Cómo de limpia es la electricidad que consumimos? *Sociedad Cooperativa eléctrica La Corriente*. <https://lacorrientecoop.es/como-de-limpia-es-la-electricidad-que-consumimos/>
- Lim, Y. (2022, abril 1). *State-of-the-Art Machine Learning Hyperparameter Optimization with Optuna*. Medium. <https://towardsdatascience.com/state-of-the-art-machine-learning-hyperparameter-optimization-with-optuna-a315d8564de1>
- Liu, J., Wang, F., & Zhen, Z. (2020). Deep Learning Based Visualized Wind Speed Matrix Forecasting Model for Wind Power Forecasting. *2020 IEEE 3rd Student Conference on Electrical Machines and Systems (SCEMS)*, 952-958. <https://doi.org/10.1109/SCEMS48876.2020.9352407>

- López de Benito, J. (2019, agosto 21). PVGIS es una herramienta que permite calcular tu producción fotovoltaica. *Energy News*. <https://www.energynews.es/pvgis-produccion-fotovoltaica/>
- López, J. F. (2017, octubre 2). *Coeficiente de determinación (R cuadrado)*. Economipedia. <https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html>
- Molina, A. (2017). *Modelo de radiación solar*. FCFM Chile. <https://solar.minenergia.cl/downloads/radiacion.pdf>
- Moro, M. (2010). *Instalaciones solares fotovoltaicas*. Editorial Paraninfo.
- Narvaez, G., Giraldo, L. F., Bressan, M., & Pantoja, A. (2021). Machine learning for site-adaptation and solar radiation forecasting. *Renewable Energy*, 167, 333-342. <https://doi.org/10.1016/j.renene.2020.11.089>
- NREL. (2022). *Data Download Usage Guide*. <https://developer.nrel.gov/docs/solar/nsrdb/guide/>
- Olah, C. (2015, agosto 27). *Comprender las redes LSTM -- blog de colah*. <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- Ordoñez, F., Vaca, D., & Lopez, J. (2019). Assessment of the Solar Resource in Andean Regions by Comparison between Satellite Estimation and Ground Measurements: Study Case of Ecuador. *Journal of Sustainable Development*, 12(4), Art. 4. <https://doi.org/10.5539/jsd.v12n4p62>
- Ordoñez Palacios, L. E., León Vargas, D. A., Bucheli Guerrero, V. A., & Ordoñez Eraso, H. A. (2020). Predicción de radiación solar en sistemas fotovoltaicos utilizando técnicas de aprendizaje automático. *Revista Facultad de Ingeniería*, 29(54). <https://doi.org/10.19053/01211129.v29.n54.2020.11751>

- Pasion, C., Wagner, T., Koschnick, C., Schuldt, S., Williams, J., & Hallinan, K. (2020). Machine Learning Modeling of Horizontal Photovoltaics Using Weather and Location Data. *Energies*, 13(10), Art. 10. <https://doi.org/10.3390/en13102570>
- Plena energía. (2022, marzo 23). *Project Sunroof: Qué es, características y alternativas*. <https://www.plena-energia.com//post/project-sunroof>
- Rivera, D. (2020, marzo 29). *Estacionalidad Compleja: ¿Obstáculo para predecir con series de tiempo?* DABIA. <https://www.grupodabia.com/post/2020-03-29-modelos-predictivos/>
- Sengupta, M., Xie, Y., Lopez, A., Habte, A., Maclaurin, G., & Shelby, J. (2018). The National Solar Radiation Data Base (NSRDB). *Renewable and Sustainable Energy Reviews*, 89, 51-60. <https://doi.org/10.1016/j.rser.2018.03.003>
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W., & Woo, W. (2015). *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting* (arXiv:1506.04214). arXiv. <http://arxiv.org/abs/1506.04214>
- Sinc. (2020, junio 12). *Cómo predecir mejor la radiación solar*. Agencia SINC. <https://www.agenciasinc.es/Noticias/Como-predecir-mejor-la-radiacion-solar>
- Taniguchi, H., Otani, K., & Kurokawa, K. (2001). Hourly forecast of global irradiation using GMS satellite images. *Solar Energy Materials and Solar Cells*, 67(1-4), 551-557. Scopus. [https://doi.org/10.1016/S0927-0248\(00\)00327-5](https://doi.org/10.1016/S0927-0248(00)00327-5)
- Tous, M. R. (2010). *Energía solar fotovoltaica*. Grupo Planeta (GBS).
- Urraca, R., Antonanzas, J., Alia-Martinez, M., Martinez-de-Pison, F. J., & Antonanzas-Torres, F. (2016). Smart baseline models for solar irradiation forecasting. *Energy Conversion and Management*, 108, 539-548. <https://doi.org/10.1016/j.enconman.2015.11.033>

- Vaca, D., & Ordóñez, F. (2020). *Mapa solar del Ecuador 2019*. EPN.
<https://www.epn.edu.ec/mapa-solar-del-ecuador/>
- Viscondi, G. F., & Alves-Souza, S. N. (2021). Solar irradiance prediction with machine learning algorithms: A Brazilian case study on photovoltaic electricity generation. *Energies*, 14(18). Scopus. <https://doi.org/10.3390/en14185657>
- Zarco, P., Ariza, F., & López, R. (1996). *Métodos de obtención de la radiación solar mediante teledetección: Órbita polar vs órbita geoestacionaria*. Dialnet.
<https://dialnet.unirioja.es/download/articulo/5339437.pdf>
- Zarzalejo, L., Santigosa, L., Polo, J., Martín Pomares, L., & Espinar, B. (2006). Estimación de la radiación solar a partir de imágenes de satélite: Nuevos mapas de evaluación de la irradiancia solar para la península Ibérica. *Averma*, 10, 11.71-11.78.

APÉNDICES

APÉNDICE A