



UNIVERSIDAD DE MARGARITA  
SUBSISTEMA DE DOCENCIA  
DECANATO DE INGENIERÍA Y AFINES  
COORDINACIÓN DE INVESTIGACIÓN Y PASANTÍA

**DESARROLLO DE MODELO DE MACHINE LEARNING PARA EL PRONÓSTICO DE  
LAS CONDICIONES METEOROLÓGICAS EN EL ESTADO NUEVA ESPARTA,  
VENEZUELA.**

Elaborado por: Isaác Figuera

Tutor: Valentina Martínez

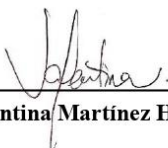
El Valle del Espíritu Santo, octubre de 2023.



**SUBSISTEMA DE DOCENCIA  
DECANATO DE INGENIERÍA Y AFINES  
COORDINACIÓN DE INVESTIGACIÓN Y PASANTÍA**

**CARTA DE APROBACIÓN DEL TUTOR**

Quien suscribe, **Ing. VALENTINA MARTÍNEZ HERNÁNDEZ**, cedula con el número V.- 24.765.943, previo cumplimiento de los requisitos exigidos en el artículo 16° de la *Normativa para el Trabajo Investigación de los Estudiantes de Pregrado de la Universidad de Margarita*, **apruebo para ser remitido al jurado**, el Trabajo de Investigación, cuyo título tentativo es ***DESARROLLO DE MODELO DE MACHINE LEARNING PARA EL PRONÓSTICO DE LAS CONDICIONES METEOROLÓGICAS EN EL ESTADO NUEVA ESPARTA, VENEZUELA***, el cual fue realizado por el estudiante de la carrera de Ingeniería de Sistemas: **ISAAC MAURICIO FIGUERA VELASCO**, cedula con el número: V.- 27.356.099.

  
\_\_\_\_\_  
**Ing. Valentina Martínez Hernández**

## DEDICATORIA

A mi familia; mi mamá, mi papá, mi hermana menor y mi hermano mayor. Gracias a ellos, a pesar de no tenerlo todo, jamás me hizo falta nada y con quienes estaré eternamente agradecido por su indispensable e incondicional apoyo a lo largo de todos mis esfuerzos, tanto académicos como personales. Ustedes son y siempre serán mi mayor fuente de motivación para seguir adelante y hacerlos sentir orgullosos.

A mi novia, quien depositó su confianza en mí y me apoyo en todo momento durante el desarrollo de esta investigación, sin su comprensión, atención y oído no habría podido alcanzar lo mejor de mí, como confidente y primera opinión a considerar en momentos de duda o confusión.

A todo aquel, que, de alguna manera u otra, aportó al desarrollo de este proyecto investigativo, cuyo apoyo y confianza proporcionó la motivación necesaria para llevar a cabo esta densa y compleja investigación.

## AGRADECIMIENTOS

En la culminación de mis estudios de pregrado, representados por esta investigación, quería comenzar por agradecer a mis padres Horlig Velasco y Kerwin Figuera por volver mi deseo de ser ingeniero una realidad, es gracias a ustedes, su apoyo, entendimiento y comprensión que me he convertido en alguien que me enorgullece ser.

A mis hermanos, María José y Samuel, quienes de manera constante e incondicional me motivaban a ser mejor, a seguir adelante y culminar mis esfuerzos. Mis primeros y mejores amigos, con los cuales podía rebotar ideas, pedir opiniones y consultar dudas de manera incesante y recibir respuesta, sin falta alguna. Motivado por los logros de mi hermano e inspirado por las capacidades de mi hermana, logré mi mayor meta hasta el momento.

A mi novia, Valentina Sánchez, por su paciencia en los momentos de duda presentados durante el mayor y más complicado reto de mis estudios. Sin cuyo afecto, aporte y motivación no habría logrado superar los obstáculos y desafíos a lo largo de todo el camino. Su tolerancia, capacidad de observación e inteligencia emocional me permitió ver más allá de mi propia perspectiva en problemas que ofuscaban mi perspectiva y limitaban mi progreso, por ello, estoy en profundo agradecimiento.

A todos aquellos compañeros con los cuales compartí mi tiempo a lo largo de todos nuestros estudios universitarios, de los que recibí y compartí conocimiento, así como motivación, risas y las fantasías de cómo el final de la carrera se vería. Sus aportes, colaboración y compañerismo fueron cruciales para el éxito de todos en este punto de nuestras vidas.

A mi tutora, Ing. Valentina Martínez, cuya paciencia, conocimiento y capacidad de observación permitió focalizar mejor mi atención y esfuerzos para el desarrollo de un proyecto detallado y completo.

A todos los profesores, comprensivos, empáticos, agradables y pacientes, quienes generosamente me compartieron sus experiencias y conocimientos que ayudaron a formarme como persona y estudiante, inspirándome a compartir mis conocimientos de la misma manera que ellos.

## LISTA DE TABLAS

Tabla N°1. Listado de variables meteorológicas.....	47
Tabla N°2. Listado de correlaciones de variables. ....	51
Tabla N°3. Listado de variables meteorológicas seleccionadas para el modelo ML. ....	52
Tabla N°4. Descripción de la variable precipitación.....	53
Tabla N°5. Cantidad de valores nulos por variable.....	54
Tabla N°6. Descripción de distribución de variables. ....	55
Tabla N°7. Cantidad de valores atípicos por variable meteorológica. ....	57
Tabla N°8. Cantidad de valores atípicos por variable meteorológica en múltiples periodos.....	60
Tabla N°9. Diferencia de rangos intercuartílicos entre periodos estudiados. ....	61
Tabla N°10. Tendencia estadística de variables meteorológicas. ....	65
Tabla N°11. Datos meteorológicos crudos para prueba de modelo. ....	73
Tabla N°12. Datos meteorológicos para prueba de modelo con características de tiempo. ....	74
Tabla N°13. Variables exógenas de datos para prueba de modelo. ....	75
Tabla N°14. Descripción Modelo_1 LSTM.....	83
Tabla N°15. Descripción Modelo_2 LSTM.....	86
Tabla N°16. Descripción Modelo_3 LSTM.....	89
Tabla N°17. Descripción Modelo_4 LSTM.....	92
Tabla N°18. Comparación de modelos LSTM.....	94
Tabla N°19. Comparación de rendimiento de modelos LSTM y ARIMA. ....	95
Tabla N°20. Comparación de métricas de rendimiento. ....	96
Tabla N°21. Rendimiento de modelo.....	98
Tabla N°22. Requerimientos técnicos.....	106
Tabla N°23. Dependencias de la propuesta y versiones.....	107
Tabla N°24. Requerimientos operativos. ....	108
Tabla N°25. Inversión inicial del proyecto. ....	109
Tabla N°26. Costos mensuales del proyecto. ....	109
Tabla N°27. Dependencias de la propuesta.....	111
Tabla N°28. Cualidades exógenas.....	113
Tabla N°29. División train y test.....	115
Tabla N°30. Descripción modelo LSTM. ....	117

## LISTA DE GRÁFICOS

Gráfica N°1. Mapa de correlación de variables. ....	50
Gráfica N°2. Diagrama de caja de variables meteorológicas normalizadas. ....	56
Gráfica N°3. Serie temporal normalizada. Fuente: Elaboración propia. (2024) .....	59
Gráfica N°4. Estacionalidad de variables meteorológicas. ....	66
Gráfica N°5. Frecuencia Estacional de variable “Temperatura”.....	67
Gráfica N°6. Residuo de variables meteorológicas.....	68
Gráfica N°7. ACF de variables meteorológicas.....	70
Gráfica N°8. PACF de variables meteorológicas.....	72
Gráfica N°9. Pronóstico de Temperatura ARIMA (1,1,1). ....	79
Gráfica N°10. Entrenamiento de Modelo_1 LSTM. ....	84
Gráfica N°11. Pronóstico de temperatura de Modelo_1 LSTM. ....	85
Gráfica N°12. Entrenamiento de Modelo_2 LSTM. ....	87
Gráfica N°13. Pronostico de temperatura de Modelo_2 LSTM. ....	88
Gráfica N°14. Entrenamiento de Modelo_3 LSTM. ....	90
Gráfica N°15. Pronóstico de temperatura de Modelo_3 LSTM. ....	91
Gráfica N°16. Entrenamiento de Modelo_4 LSTM. ....	93
Gráfica N°17. Pronóstico de temperatura de Modelo_4 LSTM. ....	94
Gráfica N°18. Entrenamiento del modelo de humedad relativa LSTM.....	118
Gráfica N°19. Entrenamiento del modelo de temperatura LSTM. ....	119
Gráfica N°20. Pronóstico de humedad relativa del modelo LSTM. ....	120
Gráfica N°21. Pronóstico de temperatura del modelo LSTM.....	121

## LISTA DE ANEXOS

Anexo N ° 1. Tablero Kanban de Trello durante el desarrollo de la propuesta. ....	124
Anexo N ° 2. Archivos del repositorio PronosticoTiempoLSTM.....	124

UNIVERSIDAD DE MARGARITA  
SUBSISTEMA DE DOCENCIA  
COORDINACIÓN DE INVESTIGACIÓN

**PROPUESTA DE MODELO DE MACHINE LEARNING PARA EL PRONÓSTICO DE  
LAS CONDICIONES METEOROLÓGICAS EN EL ESTADO NUEVA ESPARTA,  
VENEZUELA.**

Autor: Isaác Figuera  
Tutor: Valentina Martínez  
Abril de 2024

**RESUMEN**

El proyecto consiste en la evaluación de los requerimientos y las limitaciones involucradas en el desarrollo de un modelo de Machine Learning para el pronóstico de las cualidades meteorológicas del Estado Nueva Esparta, Venezuela, empleando un registro histórico de múltiples variables meteorológicas que data desde 1960 hasta 2024 localizado en una API pública y gratuita que proporciona un conjunto de datos sobre los cuales se desarrolló una serie de pruebas estadísticas y matemáticas para entender las cualidades de toda la serie temporal, que dictan los criterios que el modelo predictivo debe cumplir. A partir de lo cual se desarrollaron pruebas a menor escala que permitieron realizar una comparación directa en el desempeño de los modelos ARIMA y LSTM, mediante la observación visual y el empleo de métricas de error. Sobre los resultados obtenidos, se desarrolló una propuesta de modelo capaz de pronosticar las variables meteorológicas indicadas del Estado Nueva Esparta en el futuro. Enfocado en la línea de investigación N °3, Sistemas Inteligentes, en el área temática de Ciencia de datos.

Descriptores: Meteorología, Machine Learning, LSTM, Pronóstico, Redes Neuronales, ARIMA.



## INTRODUCCIÓN

La civilización humana guarda una estrecha relación con el estado del tiempo y la evolución del clima, debido a que son tales factores los que dictan las condiciones del entorno del cual los humanos son dependientes, es por ello que, en su incesante búsqueda de conocimientos, el estudio de las condiciones meteorológicas siempre ha estado presente, con métodos cada vez más evolutivos, avanzados y eficaces que el anterior.

Sin embargo, en el presente, la capacidad para pronosticar las condiciones meteorológicas ya existe y son globalmente empleadas por una infinidad de organizaciones y entidades, primordialmente en la forma de servicios por parte de un tercero. De manera que, el criterio involucrado en la determinación de las condiciones meteorológicas evolucionó de cómo hacerlo a cuán económico es posible hacerlo, en respuesta a la alta demanda de tal información por parte de organizaciones cuyas operaciones son dependiente de tales factores.

En este contexto, se presentan casos particulares con la necesidad de un sistema inteligente capaz de emplear la ciencia de datos para el desarrollo de pronósticos fundamentados en un registro históricos del estado del clima y todos los factores involucrados en el sistema meteorológico, que permita desarrollar un modelo predictivo capaz estimar los valores futuros de variables con utilidad práctica (como la temperatura) y de variables indicativas (como la evapotranspiración). Presentándose como una solución práctica, local y económica para entidades gubernamentales, científicas y académicas, así como para individuos o pequeños grupos de investigación científica independiente con limitados recursos y necesidades altamente específicas. La cual; en este caso, será aplicada específicamente en el estado Nueva Esparta, Venezuela.

Es así, que el presente proyecto investigativo está constituido por 6 partes primarias, que representan la manera más eficaz de abordar la situación:

Parte I: la descripción general del problema, que describe la problemática a tratar e investigar, abarcando distintas escalas con el fin de plantear interrogantes que definan el rumbo de la investigación, sus objetivos y valor académico.

Parte II: la descripción teórica, donde se encuentran las bases teóricas y legales que describen los conceptos cruciales para la contextualización y el entendimiento de lo estipulado en el proyecto,

así como la definición de términos y los antecedentes que presiden a la investigación en la misma área.

Parte III: la descripción metodológica, la cual está comprendida por las cualidades metodológicas que definen la estructura y el proceso que se lleva a cabo en la investigación de los fenómenos planteados; abarcando el tipo de investigación a emplear y su naturaleza, así como el objeto de estudio, sobre el cual se emplean las técnicas de recolección de datos y las técnicas para su correspondiente análisis.

Parte IV: el procesamiento y análisis de los resultados, que hace referencia al empleo de las técnicas de análisis definidas sobre los datos obtenidos mediante las técnicas de recolección de datos empleadas.

Parte V: las conclusiones y recomendaciones, que sirven de síntesis y guía para el lector, a partir de lo observado durante el proceso investigativo.

Parte VI: la propuesta, donde se emplea toda la información obtenida durante el proceso investigativo para construir una propuesta funcional y factible, evaluando su viabilidad, descrita desde un nivel práctico, operativo y técnico; explicando su correspondiente estructura y funcionamiento, así como los objetivos planteados para su desarrollo.

## ÍNDICE

DEDICATORIA.....	iii
AGRADECIMIENTOS .....	iv
LISTA DE TABLAS.....	v
LISTA DE GRÁFICOS .....	vi
LISTA DE ANEXOS .....	vii
RESUMEN.....	viii
<b>INTRODUCCIÓN.....</b>	<b>ix</b>
<b>PARTE I.....</b>	<b>13</b>
<b>DESCRIPCIÓN GENERAL DEL PROBLEMA.....</b>	<b>13</b>
1.1 Formulación del problema .....	13
1.2 Interrogantes.....	18
1.3 Objetivo General .....	19
1.3 Objetivo Específico .....	20
1.5 Valor académico de la investigación.....	20
<b>PARTE II.....</b>	<b>22</b>
<b>DESCRIPCIÓN TEÓRICA .....</b>	<b>22</b>
2.1 Antecedentes de investigación .....	22
2.2 Bases teóricas .....	23
2.2.1 Machine Learning (ML).....	23
2.2.2 Aprendizaje supervisado. ....	24
2.2.3 Técnicas de Machine Learning .....	26
2.2.4 Clima. ....	30
2.2.5 Variables Meteorológicas.....	31
2.2.6 Predicción Meteorológica. ....	35
2.3 Bases Legales .....	36
2.3.1 Constitución de la República Bolivariana de Venezuela .....	36
2.3.2 Ley de Meteorología e Hidrología Nacional.....	37
2.4 Definición de términos .....	38
<b>PARTE III.....</b>	<b>42</b>
<b>DESCRIPCIÓN METODOLÓGICA .....</b>	<b>42</b>
3.1. Naturaleza de la investigación.....	42
3.2. Tipo de investigación .....	42
3.3. Diseño de la investigación.....	43
3.4. Acopio y selección de la información .....	43
3.5. Técnicas e instrumentos de recolección de datos.....	44

3.6. Objeto de estudio.....	44
3.7. Técnica de análisis de datos .....	45
<b>PARTE IV .....</b>	<b>46</b>
<b>ANÁLISIS DE DATOS Y RESULTADOS.....</b>	<b>47</b>
4.1 Identificación de las variables meteorológicas más relevantes para el pronóstico del clima en el Estado Nueva Esparta. ....	47
4.2 Selección del periodo de datos meteorológicos históricos del Estado Nueva Esparta a utilizar para el entrenamiento del modelo predictivo. ....	55
4.3 Determinación de la técnica de machine learning más adecuada para el pronóstico meteorológico en el Estado Nueva Esparta. ....	62
4.4 Evaluación del rendimiento del modelo de machine learning en términos de precisión y capacidad de predicción meteorológica en el Estado Nueva Esparta. ....	96
<b>PARTE V .....</b>	<b>101</b>
<b>CONCLUSIONES Y RECOMENDACIONES.....</b>	<b>101</b>
5.1 Conclusiones .....	101
5.2 Recomendaciones.....	104
<b>PARTE VI.....</b>	<b>106</b>
<b>PROPUESTA.....</b>	<b>106</b>
6.1 Importancia de la propuesta .....	106
6.2 Viabilidad de la propuesta.....	106
6.2.1 Factibilidad técnica .....	107
6.2.2 Factibilidad operativa.....	108
6.2.3 Factibilidad económica .....	109
6.2.3.1 Costo de inicio del proyecto.....	110
6.2.3.1 Costo mensual .....	110
6.3 Metodología .....	111
6.4 Objetivos de propuesta.....	111
6.4.1 Objetivo general .....	111
6.4.2 Objetivos específicos.....	111
6.5 Estructura de programación de la propuesta .....	112
6.5.1 Definición de dependencias, llamado a API y estructuración de datos. ....	112
6.5.2 Desarrollo de las cualidades exógenas de soporte a las variables meteorológicas. ....	114
6.5.3 Preprocesamiento de los datos meteorológicos históricos para su empleo como datos de entrenamiento y evaluación de desempeño de un modelo LSTM.....	115
6.5.4 Construcción del modelo LSTM mediante definición de sus capas, dimensiones de entrada y salida y celdas por capa. ....	117
6.5.5 Entrenamiento de los modelos para cada pronóstico en base a las variables meteorológicas, las cualidades exógenas, los parámetros de paciencia e iteraciones. ....	118
6.5.6 Evaluación del rendimiento y exactitud de los valores pronosticados y los valores reales.....	121

6.5.7 Construcción del repositorio público de Github para el almacenamiento del proyecto. ....	123
<b>ANEXOS .....</b>	<b>124</b>
<b>REFERENCIAS .....</b>	<b>125</b>

## **PARTE I**

### **DESCRIPCIÓN GENERAL DEL PROBLEMA**

Esta parte está enfocada en la definición, descripción y contextualización del objeto de estudio para entender su origen, relaciones e incógnitas. Su composición consta de la formulación del problema, la cual es definida por Arias, F. (2012) como “la concreción del planteamiento en una pregunta precisa y delimitada en cuanto a espacio, tiempo y población”, estableciendo la problemática de estudio y el enfoque aplicado a su resolución. Del mismo, se derivan las interrogantes que estipulan las incógnitas a resolver para el cumplimiento de los objetivos específicos y general, los cuales definen las metas a cumplir mediante el proceso investigativo. Finalmente, se aclara el valor académico de la investigación, proveyendo los aspectos innovadores, los aportes a la sociedad y la pertinencia del tema desarrollado.

#### **1.1 Formulación del problema**

Si bien el concepto de la inteligencia artificial (IA) nació con la publicación del artículo de “A Logical Calculus of Ideas Immanent in Nervous Activity” por Warren McCulloch y Walter Pitts en el Boletín de biología matemática en 1943 (Kaplan, A., 2022), es en la última década que se ha sido testigo del auge de dicha disciplina a nivel comercial, industrial y personal, lo cual ha causado una reforma en la perspectiva de muchos con respecto a la manera de llevar a cabo actividades, operaciones y procesos. En efecto, se observa la implementación de distintas técnicas de IA, como el machine learning (ML), para la automatización de tareas de diversos grados de complejidad en múltiples áreas. Esto ha sido posible, según la organización Grand View Research (2023), por el aumento de la cantidad de plataformas que permiten el empleo de las ventajas del ML en una mayor cantidad de áreas profesionales, sin la necesidad de ser expertos en el desarrollo de tales técnicas. Subsecuentemente, causando que el mercado global de machine learning proyecte un crecimiento anual del 34,8% durante los próximos 7 años, de acuerdo con lo indicado por la misma organización.

La naturaleza del machine learning es englobada apropiadamente por la definición del pionero de esta área, quién lo expone como un "campo de estudio que brinda a las computadoras la capacidad de aprender sin ser programadas explícitamente" (Arthur, S., 1959). En tal sentido, surge la posibilidad de instruir a una computadora sin tener que desarrollar de manera explícita cada

requerimiento, regla y parámetro a tomar en cuenta, ofreciendo la capacidad de desarrollar algoritmos predictivos sin la necesidad de entender a cabalidad la manera en que las variables del fenómeno en cuestión interactúan entre sí.

En la actualidad, según Russel, S. y Norving, P. (2008), el concepto de machine learning es implementado mediante el desarrollo de algoritmos de aprendizaje supervisado o no supervisado, a partir de los cuales se derivan diferentes tipos de modelos adaptados a la predicción, clasificación o agrupación de grandes cantidades de datos numéricos, dependiendo de la necesidad del caso.

De esta manera, es posible implementar tal tecnología en cualquier entorno donde se manejen múltiples variables interdependientes y con relaciones complejas e intrincadas, evitando la necesidad de llevar a cabo un proceso laborioso de abstracción y creación de un modelo matemático. Es por tales ventajas, que las áreas de implementación del ML son tan diversas, amplias y diferentes, considerando que, de acuerdo con lo expresado por Orellana, J. (2019), estas “se encuentran actualmente en el día a día de las personas, incluyendo las recomendaciones de artículos de compra que vemos al navegar en internet, identificación de correo spam, y las sugerencias de amigos en redes sociales”, lo cual refleja su capacidad de procesamiento de datos masivos en tiempo real.

Esta utilidad permite que el ML sea empleado en el manejo de datos meteorológicos, los cuales son obtenidos en tiempo real debido a la constante abstracción de datos de la atmósfera por parte de sensores y mecanismos de distinta índole, produciendo un constante flujo de datos del cual los expertos dependen para realizar estudios y análisis de la atmósfera, la cual está bajo constante cambio, movimiento y mutaciones. Tales comportamientos son transformados en data utilizable, manipulable y más fácilmente observable mediante el empleo de modelos matemáticos y estadísticos, como lo establece Gnoza, N. y Barberena, M. (2018). Sin embargo, el clima no se detiene en ningún momento, por lo que la cantidad de información que puede ser obtenida del entorno es masiva y es empleada para su pronóstico mediante modelos matemáticos altamente complejos y costosos. Por tal motivo, el ML puede ser utilizado como un reemplazo eficaz, simplificando la implementación de procesos, aumentando precisión de pronósticos y disminuyendo costos.

De tal manera, es posible llevar a cabo con mayor facilidad y menor costo las operaciones relacionadas con el pronóstico del clima, cuya importancia es establecida, por la Organización Meteorológica Mundial, como la necesidad del individuo de saber qué ropa utilizar y en base a qué condiciones planificar el día. Sin embargo, este también puede proveer información crucial para la toma de decisiones en áreas de gran relevancia, como el transporte marítimo, aéreo y terrestre, la correcta administración del agua dulce, la planificación de eventos, la agricultura, construcción, manejo de energía y la preparación ante eventos climáticos catastróficos, debido a que las condiciones climáticas rigen los ciclos de los recursos como el agua y las energías renovables, así como el estado de las condiciones en el que una actividad debe desempeñarse. Por tal motivo, saber con antelación el estado del clima, ayuda a planificar con menor posibilidad de error y de circunstancias inesperadas.

Además, en comparación a otras doctrinas, la meteorología es relativamente joven. De acuerdo con la Enciclopedia Británica (2023), sus principales bases teóricas se establecieron a inicios del siglo XX, de la mano de un equipo de meteorólogos suecos y noruegos, bajo la dirección de Vilhelm Bjerknes, quien propuso el concepto de la predicción del clima mediante la utilización de modelos matemáticos basados en principios científicos que forman parte de su propia contribución en las ciencias meteorológicas.

La composición general de tales modelos matemáticos predictivos consiste en una serie de ecuaciones matemáticas que buscan replicar los fenómenos que se dan en la atmósfera, como el movimiento de las masas de aire, la energía y termodinámica involucrada. Es por ello que Gnoza, N. y Barberena, M. (2018) indican que “el trabajo de los meteorólogos consiste en analizar el resultado de modelos matemáticos ejecutados en esas supercomputadoras ...” (p. 34). Su resolución implica cálculos muy complejos y laboriosos que son llevados a cabo por computadores de alta capacidad de procesamiento. Por tal razón, aquellos involucrados en la predicción meteorológica están en la constante búsqueda e implementación de nuevas herramientas tecnológicas para el procesamiento de datos.

Con la llegada del machine learning a inicios del siglo XXI, mediante la implementación de nuevas tecnologías de análisis de datos por parte de Google, se da inicio a una nueva época de procesamiento, donde los analistas se fundamentan en la cantidad de información disponible, más que en abstracción y representación matemática de fenómenos. Es por ello que existen grandes



esfuerzos organizacionales para el desarrollo de modelos de predicción de clima mediante la utilización de ML, como lo han hecho los autores Weyn, J., Durran, D., Caruana, R., & Cresswell-Clay, N. (2021), en un nuevo estudio titulado “Sub-Seasonal Forecasting With a Large Ensemble of Deep-Learning Weather Prediction Models”, donde establecen que:

Presentamos un sistema de ensamble predictivo utilizando un modelo Deep learning Weather Prediction (DLWP) que puede predecir recursivamente seis valores atmosféricos dentro de un espacio de 6 horas (...) Aunque nuestro modelo DLWP no predice precipitaciones, si predice el total de columnas de vapor de agua y da una predicción razonable de 4.5 días del huracán Irma.

De tal manera, se expone la capacidad del ML de procesar grandes cantidades de datos climáticos para el desarrollo de pronósticos altamente complejos, no solo abarcando las condiciones meteorológicas a nivel global, sino también para el manejo de fenómenos climáticos excepcionales, como lo sería el Huracán Irma, el más fuerte observado en el Atlántico según el Centro Nacional de Huracanes de Estados Unidos (2017), el cual sucedió en el 2017 y afectó a algunas partes de los Cayos de la Florida y a las islas de Sotaviento en el Caribe, las cuales están repartidas entre los Países Bajos y Venezuela. Dentro de tal grupo se puede resaltar al Archipiélago los Monjes, Los Roques, Los Frailes y el Estado Nueva Esparta.

En el caso de Venezuela, “el Instituto Nacional de Meteorología e Hidrología (INAMEH), tendrá a su cargo la ejecución de las políticas que en las áreas meteorológica e hidrológica dicte el ministerio con competencia en materia ambiental” (Ley de Meteorología e Hidrología Nacional, 2006, Artículo 12). Esto quiere decir que los asuntos referentes a la meteorología en este país son jurisdicción del INAMEH, el cual establece, en su portal web, que utiliza equipos de vanguardia, como estaciones meteorológicas de superficie, radares con tecnología Doppler, estaciones de lanzamiento de radiosonda, receptores de imágenes satelitales y estaciones de aeropuerto para el monitoreo de las condiciones climáticas en todo el territorio nacional. Sin embargo, en la plataforma también se evidencia que este emplea una API del modelo “Global Forecast System” (GFS) de “National Oceanic and Atmospheric Administration” (NOAA), el cual es una división del Centro Nacional de Huracanes de Estados Unidos, encargada de la predicción climática.

En este contexto, el enfoque global de tal sistema puede representar una desventaja, debido a que “el GFS se ejecuta cuatro veces al día y produce pronósticos con hasta 16 días de anticipación. El componente de pronóstico utiliza el modelo de volumen finito al cubo (FV3) con una resolución

de ~13 km.” (National Oceanic and Atmospheric Administration [NOAA], 2019); lo cual da a entender que el modelo mencionado es, esencialmente, un portal especializado de acceso público donde se pueden observar los comportamientos meteorológicos a lo largo de toda la superficie terrestre con una resolución mínima de 0.25°, esto quiere decir que hay un límite de precisión y que los registros climáticos históricos no son utilizados para la realización de una predicción local precisa y específica, sino que se implementan métodos globales externos para una predicción general de las condiciones sobre una superficie más amplia y ambigua.

Específicamente, en Venezuela, la Ley de Aeronáutica Civil (2001), indica en su artículo 61 que, dentro de los servicios empleados en la navegación aérea, se encuentra el de información meteorológica, estableciendo que su uso es obligatorio para todas las aeronaves que operen en el territorio de la República. Es por ello que en el Estado Nueva Esparta existe una estación meteorológica en el Aeropuerto Internacional del Caribe General en Jefe Santiago Mariño, en el Municipio Díaz, corroborado por el mapa interactivo de las estaciones meteorológicas en el portal web del INAMEH. De acuerdo con Medina, J. y Durante, C. (2007, p.48), la estación meteorológica neoespartana está bajo la administración directa de la Fuerza Aérea de Venezuela y ha sido empleada para la observación de las condiciones climáticas para la realización efectiva de las operaciones aeroportuarias

De manera que, para el cumplimiento de dicha legislación, fue necesaria la implementación de un sistema ASOS (Sistema automático de observación de superficie); el cual, de acuerdo con la NOAA (2019), se encarga de detectar cambios meteorológicos significativos y difunde observaciones horarias y especiales a medida que se cumplen los umbrales de los criterios meteorológicos, proveyendo datos climatológicos básicos como el viento, temperatura, precipitación, presión, entre otros. Así como también son capaces de observar, formatear, archivar y transmitir esta información automáticamente a sistemas globales aeronáuticos.

De esta forma, los datos históricos climatológicos del Estado Nueva Esparta son accesibles públicamente mediante plataformas como Open-Meteo, la cual se define como:

“API meteorológica (interfaz de programación de aplicaciones) de código abierto y ofrece acceso gratuito para uso no comercial. (...) Open-Meteo utiliza una amplia gama de modelos meteorológicos locales con actualizaciones rápidas, lo que garantiza que se genere el pronóstico más preciso para cualquier ubicación a nivel mundial.” (Zippenfenig, P, 2023).

En tal sentido, proporcionan datos de alta calidad y precisión obtenidos localmente, lo que posibilita el aprovechamiento de la data registrada durante los años para proveer un pronóstico más detallado, inmediato y personalizado a las necesidades del contexto, que permita tener un menor margen de error y probabilidad de falla al momento de organizar, planificar y ejecutar eventos o actividades recurrentes, en sectores industriales, públicos, privados y del día a día de un individuo común.

En efecto, la utilización de machine learning para la predicción de las condiciones meteorológicas en el Estado Nueva Esparta se considera un reemplazo efectivo, eficiente y general de los métodos tradicionales, donde se aprovecharán los registros meteorológicos de cada una de las regiones del país, adquiridos de bancos internacionales de datos meteorológicos (como el ECMWF), para alimentar un modelo de ML entrenado en el entendimiento de fenómenos meteorológicos. La implementación del tal sistema proveería una herramienta tangente, propia y precisa para la activa predicción de las variables meteorológicas, empleando información real de las propias estaciones locales, familiarizadas con el comportamiento del entorno. Además, facilitaría una planificación más realista a la hora de realizar operativos que se vean afectados por las condiciones climáticas, como el mantenimiento de infraestructura pública, la realización de eventos turísticos o la prevención de daños materiales y humanos debido a imprevistos climáticos.

## **1.2 Interrogantes**

Es a partir de tal problemática que, tomando en cuenta las variables, limitaciones y el contexto, se desarrolla una interrogante general que engloba el propósito general que la investigación deberá cumplir:

¿Cuál sería el modelo que permita un óptimo pronóstico de las condiciones meteorológicas en el Estado Nueva Esparta?

De tal manera, se derivan interrogantes específicas que buscan ilustrar los pasos secuenciales que la investigación deberá cumplir para dar respuesta a la interrogante general de la misma.

1.- ¿Cuáles son las variables meteorológicas relevantes en el pronóstico del clima en el Estado Nueva Esparta?

- 2.- ¿Cuán largo debe ser el periodo de datos meteorológicos a utilizar para desarrollar un modelo predictivo preciso del clima en el Estado Nueva Esparta?
- 3.- ¿Qué modelo de machine learning es más apto en la predicción del clima en el Estado Nueva Esparta?
- 4.- ¿Cuál es el rendimiento del modelo de machine learning en términos de precisión y capacidad de predicción meteorológica en el Estado Nueva Esparta?

### **1.3 Objetivo General**

Desarrollar un modelo de machine learning para el pronóstico de las condiciones meteorológicas en el Estado Nueva Esparta, Venezuela.

### **1.4 Objetivos Específicos**

- 1.- Identificar las variables meteorológicas más relevantes para el pronóstico del clima en el Estado Nueva Esparta.
- 2.- Seleccionar el periodo de datos meteorológicos históricos del Estado Nueva Esparta a utilizar para el entrenamiento del modelo predictivo.
- 3.- Determinar la técnica de machine learning más adecuada para el pronóstico meteorológico en el Estado Nueva Esparta.
- 4.- Evaluar el rendimiento del modelo de machine learning en términos de precisión y capacidad de predicción meteorológica en el Estado Nueva Esparta.

### **1.5 Valor académico de la investigación**

El presente trabajo de investigación busca desarrollar una herramienta empleada en el campo del pronóstico meteorológico, fundamentada en los registros históricos de las condiciones climáticas del Estado Nueva Esparta; de manera que se pueda aprovechar la información recolectada por los equipos de monitoreo climático, con el fin de emplearla para entrenar y alimentar un modelo predictivo climático preciso de las condiciones meteorológicas de la región, el cual esté en la capacidad de producir predicciones meteorológicas claras y precisas, sin la necesidad de resolver complejas ecuaciones meteorológicas basadas en fenómenos físicos, de forma tal, que consuma pocos recursos computacionales.

Asimismo, se busca establecer una base metodológica y técnica que sirva de referencia práctica al momento de implementar un modelo de machine learning en una problemática de predicción, además de proveer una guía específica sobre el proceso de análisis del contexto, el estudio de las variables involucradas, las decisiones envueltas en la construcción del entorno de desarrollo, de la arquitectura y de las herramientas utilizadas en la construcción de los datasets empleados en el entrenamiento, testeo y validación de la predicción o pronosticación. Se incluyen también los procesos preliminares de formateo, estructuración y limpieza de los datos empleados durante el desarrollo.

De esta manera, la investigación puede ser empleada como sustento en futuros proyectos de temáticas similares, funcionando como base para que otras organizaciones, sean regionales, nacionales o internacionales, públicas o privadas, implementen las mismas mecánicas en sistemas ya existentes, para mejorar el control de riesgo sobre sus procesos y disminuir costos, o para ser implementados en contextos nuevos donde el machine learning desarrolle una solución con una mejor relación de costo-beneficio y una mayor adaptabilidad a condiciones impredecibles.

## **PARTE II**

### **DESCRIPCIÓN TEÓRICA**

Esta parte está centrada en la fundamentación teórica de la problemática y el enfoque desarrollado a lo largo de la investigación. Según Arias, F. (2012), se define como “el producto de la revisión documental-bibliográfica y consiste en una recopilación de autores conceptos y definiciones, que sirven de base a la investigación por realizar (...)”, lo que incluye los antecedentes de la investigación, los cuales son previos trabajos que representan los avances y el estado actual del conocimiento del área. Asimismo, en esta parte se encuentran las bases teóricas, donde se exponen los conceptos más importantes del área. Seguidamente, se complementa la teoría planteada mediante el desarrollo de un glosario de términos, disponibles para resolver dudas de definición en el lector.

#### **2.1 Antecedentes de investigación**

Marín y Pineda (2019) efectuaron un estudio titulado: *Modelo predictivo machine learning aplicado a análisis de datos hidrometeorológicos para un SAT en represas. Arequipa, Perú*, con un enfoque cuantitativo de nivel descriptivo en el desarrollo de una metodología para la implementación de una red neuronal capaz de funcionar como herramienta en la toma de decisiones en el sistema de alerta temprana (SAT). Para ello, emplearon datos meteorológicos e hidrológicos históricos y de monitoreo en tiempo real con el fin de alimentar una red de memoria a corto y largo plazo, resultando en un error promedio de 1.3%, convirtiéndolo en un método viable y factible en la predicción de un fenómeno parcialmente meteorológico, estableciendo un precedente en la funcionalidad y la aplicabilidad del machine learning en la pronosticación de variables meteorológicas.

Por otro lado, la investigación también concluyó que, la mejor técnica de machine learning para la solución de problemas basados en series de tiempos es la LSTM (Long-Short Term Memory, por sus siglas en inglés), debido a su mayor capacidad de adaptación a datos en orden cronológico, el cual es un aspecto que también está presente en la data histórica meteorológica que será empleada en el pronóstico de las variables meteorológicas del Estado Nueva Esparta. De tal manera, el autor abarca una situación con una estructura similar a la del presente trabajo, resultando así en una perspectiva y aproximación de gran valor.

Gnoza y Barberena (2018) realizaron un trabajo titulado: *Estudio de factibilidad del uso de machine learning con múltiples fuentes de datos en el pronóstico del tiempo. Montevideo, Uruguay*; regidos por una metodología de enfoque incremental e iterativa para la minimización de la incertidumbre y la maximización de la retroalimentación durante la conglomeración de datos meteorológicos históricos a partir de distintas fuentes, incluyendo las proporcionadas por el Instituto Meteorológico Uruguayo, los datos generados por una mini-estación meteorológica controlada por arduinos y la obtenida mediante la utilización de API's climatológicas. En conjunto, dichos recursos fueron empleados en la alimentación de un sistema predictivo de machine learning desarrollado en Python, localizado en una arquitectura que implementa módulos de obtención de datos, persistencia de datos (almacenamiento), procesamiento predictivo y presentación al usuario mediante una aplicación web, convirtiendo los datos históricos en información sobre las futuras precipitaciones, exclusivamente, indicando de manera discreta si en un instante específico lloverá o no y, en caso de que sí, en qué medida lo hará.

Para su elaboración, se utilizó una amplia cantidad de plataformas y herramientas que permitieron la construcción del sistema y el cumplimiento de los requerimientos, sin embargo, su correlación principal con el proyecto actual es su enfoque en el desarrollo de una solución de bajo costo a la necesidad de predicción climática mediante recursos locales, tal y como se plantea en el presente caso. Simultáneamente, se presenta como una fuente documental de alto valor para la respuesta de las incógnitas presentadas en la formulación de la problemática y la metodología a seguir para el desarrollo de un análisis inicial a la data meteorológica para el entendimiento de las relaciones entre las variables que abarca y el efecto de tal información en la construcción de datasets de entrenamiento, prueba y validación, la verificación de la precisión y los métodos tomados para satisfacer los requerimientos.

Shah (2021), realizó un proyecto de investigación titulado: *Pronóstico de temperatura a corto plazo utilizando LSTMS y CNN*, desarrollado en un contexto experimental y enfocado en la construcción de un modelo de ML learning adaptado a las cualidades y atributos de la variable histórica de temperatura, fundamentándose en el trabajo de diversos autores para entender la arquitectura y los requerimientos de las técnicas de ML, para contrastarlo y poder concluir qué técnica utilizar y por qué. Seguidamente, el autor expone las técnicas de preprocesamiento a aplicar para mejorar la capacidad predictiva y de precisión de un modelo basado en series temporales, indignado en los beneficios de los algoritmos de convolución para la suavización de la serie de

datos para la eliminación de valores atípicos y la eliminación de diferencias en preparación a la normalización. Igualmente, el autor desarrolla la arquitectura del proceso iterativo de entrenamiento, presentando la secuencia de pasos a llevar a cabo para el mejoramiento recursivo del modelo, utilizando diversos indicadores numéricos estadísticos organizados en una matriz de desempeño, que sirve de referencia para el mejoramiento de los pesos del modelo. Finalmente, el autor concluye en la importancia del uso de tales indicadores para la optimización del proceso y en los beneficios de la convolución de los datos para el mejoramiento de la precisión y la eficiencia del modelo en sí. De tal manera, el proyecto en cuestión servirá como una fuente metodológica para la construcción de la arquitectura del proceso de entrenamiento, y para la aplicación de técnicas de preprocesamiento como la convolución unidimensional.

## **2.2 Bases teóricas**

### **2.2.1 Machine Learning (ML).**

El concepto de machine learning o aprendizaje automático es tan complicado de abarcar completamente en una definición como el concepto de aprendizaje en sí, el cual es definido por el diccionario Merriam-Webster (2023) como: “la modificación de la tendencia del comportamiento a partir de la experiencia”, lo que permite inferir que el aprendizaje es el cambio realizado en las líneas de comportamiento de una entidad, ya sea como producto de la experimentación y/o exposición a la realidad estudiada. De manera que, al aplicar el mismo enfoque en el contexto de las máquinas, el autor Nilsson, N. (1998) indica que: “una máquina aprende cada vez que cambia su estructura, programa o datos (en función de sus entradas o en respuesta a información externa) de tal manera que su futuro rendimiento mejore.”, dando a entender que es mediante la interacción con datos externos que una máquina es capaz de alterar sus procesos internos con el propósito de cumplir ciertos requerimientos. Asimismo, el mismo autor indica que el machine learning usualmente se refiere a los cambios realizados en una máquina que realiza actividades relacionadas con IA, como lo podría ser el reconocimiento de patrones, diagnóstico, robótica y la predicción, indicando que los cambios pueden ser mejoras a una serie de instrucciones preexistentes o un inicio desde cero de nuevos sistemas.

Su funcionamiento se basa en la exposición e interacción con data o inputs, donde la IA o el “agente”, se dedica a percibir y modelar su entorno, computando acciones que pueden incluir la anticipación de sus efectos, de acuerdo con lo establecido por Nilsson, N. (1998), por lo cual,



cualquier cambio realizado a tales variables a partir de lo observado puede ser considerado como aprendizaje. Sin embargo, puede llegar a ser difícil entender los motivos por los cuales es necesario que una máquina realice instrucciones que no fueron indicadas explícitamente. No obstante, los casos de aplicación son generalmente aquellos procesos donde las instrucciones son poco específicas o ambiguas pero se tiene ejemplos de cómo la salida de ser en relación a la entrada, en cuyo caso se vuelve conveniente que sea capaz de ajustar su estructura interna para que, a partir de los ejemplos y sus relaciones, pueda brindar los resultados requeridos y su entendimiento sobre la relación entre ellas.

Subsecuentemente, los autores Giuseppe, M. y Augusto, I. (2013, pág. 24), indican que:

En los últimos años, el uso de la enseñanza máquina se ha desplegado con mucha ligereza, gracias al poder computacional, ya que se pueden ver aplicaciones en dominios como el hallazgo de fraudes, sistemas de recomendación, hallazgo de spam, predicciones financieras, comercio y mercadeo, entre otros.

De este modo, se puede observar que, en las áreas de aplicación previamente mencionadas, los casos individuales de estudio son objetos de alta complejidad, con múltiples dimensiones y variables interdependientes a considerar para la realización de tareas como la categorización, para diferenciar un fraude de tarjeta de crédito de un consumo legítimo, la predicción, para saber qué recomendar en base a las búsquedas recientes, y la abstracción de relaciones entre variables, para el entendimiento de la bolsa, por ejemplo.

Por tales motivos, las cualidades y capacidades del ML encajan con los requerimientos y las limitaciones de la predicción del clima, debido a que este es “el estado de la atmósfera con respecto al calor o al frío, a la humedad o a la sequedad, a la calma o a la tormenta, a la claridad o a la nubosidad”, según el diccionario Merriam-Webster (2023), aclarando que el clima es el producto final de la interacción entre variables interdependientes entre sí, cuyas magnitudes de dependencia pueden variar a lo largo del tiempo y relativas unas a las otras, por lo que la constitución manual de una serie de instrucciones que puedan simular el comportamiento de todas las variables involucradas es eclipsado por la conveniencia de utilizar el ML para la abstracción y predicción de sus relaciones a partir de data registrada del fenómeno en sí.

### **2.2.2 Aprendizaje supervisado.**

Al momento de desarrollar un modelo de ML, la técnica empleada para la comprensión e interpretación de los datos es crucial y está intrínsecamente ligada a las cualidades de los datos en

sí y al resultado final que se busca obtener. Es por ello que existen técnicas como el aprendizaje supervisado, la cual, según Valenzuela, G. (2022), consiste en “construir, a partir de los datos de entrenamiento, un modelo de predicción. (...) Usando este modelo, podremos predecir la variable respuesta de un nuevo conjunto de datos no vistos (test sample), únicamente conociendo sus variables explicativas.”. En esencia, explica que un modelo de ML puede ser entrenado en base a una porción de la data que contendrá la variable independiente y la variable dependiente, a partir de los cuales desarrollará la capacidad de encontrar patrones en las relaciones. Subsecuentemente, el modelo debe ser puesto a prueba con una muestra de la misma data a la cual no fue expuesto previamente y no contendrá la variable dependiente, para así comparar sus estimaciones con los datos reales. Es a partir de allí que se realizan manipulaciones tomando en cuenta el error que el modelo tuvo en sus predicciones.

Por ello, este es un método empleado para la realización de tareas con requerimientos específicos, donde es necesario la identificación de patrones a partir del contraste de datos y la tendencia.

Según Dalúa, J. (2021), existen dos tipos de problemas para los cuales se emplea el aprendizaje supervisado. Comenzando con la clasificación, el autor indica que “se emplea un algoritmo para asignar con precisión datos de prueba en categorías específicas, como separar manzanas de naranjas.”, por lo cual el modelo es entrenado utilizando datos ya clasificados para encontrar patrones entre las cualidades de cada elemento y su categoría asignada, de manera que, al ser probado con un set de datos que no estén clasificados, sea capaz de diferenciar entre categorías. De la misma manera, el modelo de ML puede ser entrenado para diferenciar entre días donde llovió y días donde no llovió a partir de variables independientes como la humedad, temperatura y presión.

Seguidamente, el autor indica que “Los modelos de regresión son útiles para predecir valores numéricos basados en diferentes puntos de datos, como proyecciones de ingresos por ventas para un negocio determinado”. En este caso, busca establecer una relación de magnitudes entre las variables independientes y dependientes, donde el resultado no es la clasificación en distintas categorías, sino un valor continuo o discreto a partir de las relaciones que observa en el set de entrenamiento, es por ello que son perfectos para predecir el comportamiento de una variable a lo largo del tiempo. Este tipo de modelos puede ser empleado para saber cuánto va a llover un día

determinado, dependiendo de la tendencia histórica de la humedad, presión y temperatura, por ejemplo.

### **2.2.3 Técnicas de Machine Learning**

#### **2.2.3.1 Regresión Lineal**

Es una técnica de Machine learning supervisado en la que el modelo encuentra la relación lineal existente entre la variable dependiente y la variable independiente, según Dhingra, D (2023). De manera que el modelo de ML entiende la relación entre variables comparando las magnitudes que la diferencian o separan a lo largo de un mismo eje, el cual podría ser un línea cronológica. Asimismo, tal técnica puede ser aplicada tanto de manera simple como múltiple, en cuyo caso existe más de una variable independiente que afecta a una única variable dependiente.

Aclarando que busca encontrar la mejor función lineal que intercepte ambos valores, de manera que el error sea minimizado. Por lo tanto, tomando en cuenta que en el presente caso, el fenómeno a estudiar está compuesto por diversas variables, sería necesario aplicar la regresión lineal múltiple para la pronosticación de una sola variable. Esto sería una limitante en cuanto a la funcionalidad del modelo, pero se presenta como la solución más simple al problema dado.

Sin embargo, para la aplicación de la regresión lineal, es necesario verificar la naturaleza de la relación entre las variables. Lo cual puede ser realizado mediante la estructuración de diversas gráficas que ilustran el comportamiento y la correlación entre variables. Es allí donde es necesario verificar que la variable dependiente debe estar linealmente relacionada con la variable independiente, que su distribución sea normal, que la varianza del error sea constante para todo valor dependiente (homocedasticidad), que no existan correlaciones entre las variables independientes, que los términos de error estén distribuidos normalmente y que no exista autocorrelación entre ellos, según Dhingra, D (2023).

#### **2.2.3.2 Redes Neuronales Artificiales (RNA)**

El concepto de las RNA nace con el ML, en la mente del creador del primer computador de propósito general, Alan Turing, quien en 1936, vió el cerebro como una forma de ver el mundo de la computación, según Matich, D. (2001). Asimismo, el autor indica que existen diversas maneras en las que este concepto es definido, como lo podría ser:

Redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico.

Haciendo énfasis en el origen biológico de su arquitectura, de manera que busca replicar la estructura en la que las neuronas biológicas de un ser vivo interactúan entre sí para el procesamiento de información, con gran interconexión, formación jerarquizada y alta capacidad adaptativa. Por consiguiente, sus cualidades le proveen capacidad y ventajas diferenciables y significativas a la hora de considerarlo como una herramienta para la predicción de una línea cronológica de variables, como lo indica Matich, D. (2001):

- Capaces de realizar tareas en base a un entrenamiento inicial.
- Desarrollan su propia organización interna dependiendo de la información presentada.
- Capaces de reconstruirse en caso de falla.
- Empleables en operación en tiempo real.
- Fácil implementación en tecnología existente.

Así pues, sus atributos principales presentan sus fortalezas frente a la problemática desarrollada en el presente proyecto, donde su alta capacidad de adaptación y auto-organización lo convierten en una opción viable para la predicción meteorológica. Así como lo indica Marín D. y Pineda, I. (2019):

Una RNA aprende, memoriza y divulga las diversas relaciones encontradas en los datos. Es capaz de modelar complejas relaciones no lineales encontradas en los datos de una cuenca hidrográfica, sin un conocimiento previo y explícito de las características físicas del proceso

Teniendo en cuenta que las variables meteorológicas empleadas en la predicción del comportamiento de una cuenca hidrográficas son esencialmente las mismas que serían empleadas para un pronóstico meteorológico, debido al alto nivel de dependencia entre ambos fenómenos. Entiendo que, a diferencia de la regresión lineal múltiple, las redes neuronales están en la capacidad de identificar patrones y relación no lineales entre los factores involucrados, en caso de ser necesario. Asimismo, las RNA, de acuerdo a su arquitectura, son capaces de producir una mayor cantidad de salidas que los modelos regresivos, así como se puede observar en la siguiente figura:

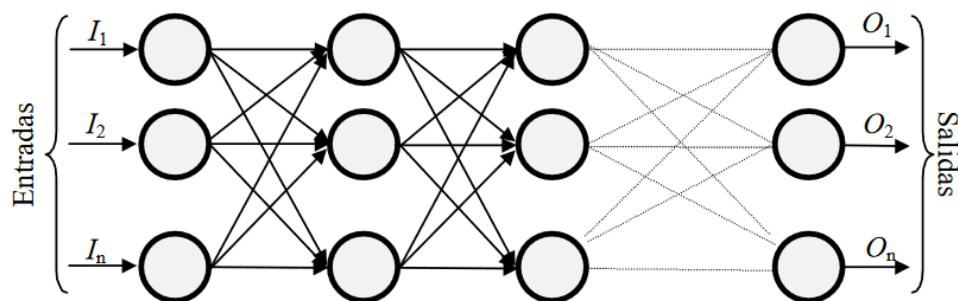


Figura 2.2.3.2: Ejemplo de una red neuronal simple

En la cual se puede observar la arquitectura de una RNA simple, incluyendo la capa de entrada, la capa de salida y todas aquellas entre ambas, que son la capa oculta, la cual puede estar compuesta por más capas. Dentro de la misma, el proceso de los datos es descrito por Matich, D. (2001), como “Los datos ingresan por medio de la ‘capa de entrada’, pasan a través de la ‘capa oculta’ y salen por la ‘capa de salida’.” Entendiéndose que a partir de la comparación numérica dentro de las capas ocultas y su alta conectividad, que la red será capaz de cotejar magnitudes y producir múltiples salidas predictivas, sobre la única que la regresión lineal es capaz de desarrollar.

### 2.2.3.3 Redes Neuronales recurrentes (RNR)

Es un tipo de RNA, el cual es empleado por el sistema Siri de Apple y la búsqueda por voz de Google, debido a su capacidad de recordar su entrada o *input*, dado que posee una memoria interna, por ello Whitfield, B (2023), indica que “Por eso son el algoritmo preferido para datos secuenciales como series de tiempo (...)”. Esencialmente, su principal cualidad es la recursividad, cuando en una RNA los datos son trasladados en una sola dirección, hacia la capa de salida, en las RNR, la información atraviesa un ciclo. Tomando el ejemplo del mismo autor, cuando una RNA lineal intenta procesar la palabra “neuron”, Whitfield indica que “Cuando llega al carácter ‘r’, ya se ha olvidado de ‘n’, ‘e’ y ‘u’.”. Sin embargo, una RNR, “es capaz de recordar esos caracteres gracias a su memoria interna. Produce resultados, los copia y los reincorpora a la red.” De manera que, logra producir un resultado más preciso y apegado a la realidad en casos donde existen largas líneas temporales, que pueden provocar irregularidades de memoria en otras estructuras de redes, proporcionando ventajas significativas en la pronosticación de las condiciones meteorológicas.

Su estructura está calificada en relación a la cantidad de entradas y salidas que tenga, de acuerdo con Whitfield, B (2023):

- Uno a uno: Posee una entrada y una salida.
- Muchos a uno (y viceversa): Posee múltiples entradas y una salida, o al revés.

- Muchos a muchos: Posee múltiples y salidas, no necesariamente la misma cantidad en ambos.

Por ello, es factible la implementación de una RNR para la pronosticación de las condiciones meteorológicas, ya que permite la obtención de múltiples salidas de gran precisión y utilidad directa, a partir de todos los indicadores relacionados con tales fenómenos. Aprovechando la mayor cantidad de recursos para la mejor calidad de resultados.

Sin embargo, las RNR poseen complicaciones inherentes a su diseño. Tomando en cuenta que estas arquitecturas emplean gradientes para el entendimiento del progreso de los datos y su evolución sobre una línea temporal, uno de los problemas más comunes era el de los gradientes explosivos, los cuales, según Whitfield, B (2023), consisten en “cuando el algoritmo, sin mucha razón, asigna una importancia estúpidamente alta a los pesos”. Haciendo referencia a que la repetitividad de la recursividad provoca una alteración exponencial de los criterios (o pesos) del modelo de RNR, por fortuna, es un problema resuelto mediante el truncar o aplastar de los gradientes de criterio. Por otro lado, el problema de los gradientes en disminución, sobre los cuales el autor expone que “ocurren cuando los valores de un gradiente son demasiado pequeños y, como resultado, el modelo deja de aprender o tarda demasiado.” Es decir, sucede lo opuesto a la explosión de gradientes, en cuyo caso, los criterios se vuelven ínfimos y no promueven el aprendizaje del modelo. Si bien, en un principio, las RNR parecían ser un método viable para la pronosticación de variables meteorológicas, sus inconvenientes la vuelven inutilizable por motivos de inestabilidad. Por otro lado, y por fortuna, es partir de ellas que nace el LSTM (*Long-Short Term Memory*), como extensión de las RNR, con una capacidad extendida de memoria y sin los inconvenientes de su origen.

#### 2.2.3.4 LTSM (*Long-Short Term Memory*)

Le brindan la capacidad a las RNR de recordar entradas sobre un periodo largo de tiempo, debido a que “los LSTM contienen información en una memoria, muy parecida a la memoria de una computadora. El LSTM puede leer, escribir y eliminar información de su memoria.” Whitfield, B (2023). Cuyo proceso puede ser entendido mediante el uso de una celda cerrada, donde la información por almacenar depende de la importancia que los pesos o el criterio del modelo o la neurona le asigne en un instante dado, de manera que pueda entender con el tiempo, qué información es importante y cuál no lo es tanto. Su arquitectura está compuesta de la siguiente manera:

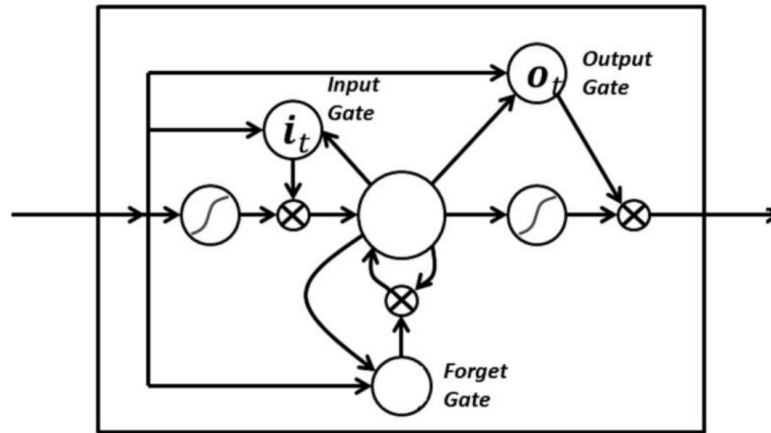


Figura 2.2.3.4: Arquitectura de LSTM

Donde, de las 3 celdas, una maneja la entrada de nueva información por almacenar, el olvido de tal información debido a la pérdida de significancia o si se deja que afecte a la salida en el paso del tiempo. Resolviendo el problema de gradientes de las RNR, manteniendo el proceso de entrenamiento corto y la precisión alta. Es por ello que Marín, D., Pineda, I. (2019), en el desarrollo de un modelo de ML para la predicción del comportamiento de un caudal a partir de las variables meteorológicas del entorno, concluyeron que “la red Long Short Term Memory (LSTM) es el modelo más adecuado para trabajar con datos de orden cronológico, ya que están adaptadas para resolver problemas basados en series de tiempo.”

En definitiva, una de las técnicas de ML previamente mencionadas será empleada en el pronóstico meteorológico del estado Nueva Esparta, a partir de una análisis exploratorio de los datos meteorológicos históricos que brindara información con respecto a la estructura, naturaleza, distribución y constitución de los datos y sus relaciones, los cuales podrán ser contrastados con los requisitos de cada técnica y los requerimientos del producto final deseado, para así decidirse por la mejor técnica de ML sin la necesidad de desarrollar un modelo para cada una y comparar resultados directamente.

#### 2.2.4 Clima.

El clima es un concepto usado libremente de manera coloquial para describir eventos del día a día del entorno, no obstante, este concepto en realidad se refiere a “cuando afirmamos que en nuestra ciudad los inviernos son muy fríos y secos” (Ródriguez, R., et al, 2004), debido a que allí se realizan consideraciones sobre largos periodos de tiempo para concluir que el valor promedio de temperaturas y humedad de ese lugar es, efectivamente, frío y seco. En contraste, cuando llueve

repentinamente y arruina la planificación del día, lo cual es un fenómeno meteorológico que puede durar entre horas y días, se está hablando del tiempo.

De tal manera, se define formalmente al clima como “Conjunto de condiciones atmosféricas que caracterizan una región.” (Real Academia Española, 2022), de manera que, para referirnos al clima, se habla principalmente sobre un atributo o fenómeno meteorológico recurrente de un lugar determinado. Es por ello que el calentamiento global y todas las consecuencias que conlleva en el entorno son denominados como parte del cambio climático, debido a que son variaciones lentas y progresivas de las cualidades meteorológicas de la tierra a lo largo de un periodo extenso.

En el presente trabajo, se decide pronosticar el tiempo y no el clima, debido a que el primero se refiere al valor real de una variable meteorológica como la precipitación o la temperatura, en un instante determinado, proporcionando un servicio práctico para la planificación individual u organizacional en base al tiempo esperado.

#### **2.2.5 Variables Meteorológicas.**

Según Gonzáles, J. (2013), las variables meteorológicas, se definen como “toda propiedad con condición de la atmósfera, cuyo conjunto define el estado del tiempo (a corto plazo) o del clima (a largo plazo).”, lo que refleja que su propósito es la representación de los valores reales de la propiedad del clima en un área determinada o el estado del tiempo en un momento específico, de manera que, a partir de las variables meteorológicas es que se puede abstraer y representar el comportamiento de los distintos atributos que colaboran entre sí, produciendo como resultado el estado del clima o tiempo. Existen muchas variables distintas y se conoce que entre ellas existen relaciones de dependencia o interdependencia, entre las cuales se encuentran la temperatura del aire, la precipitación en mm, la humedad relativa, la presión atmosférica, orientación y velocidad del tiempo, entre otros.

Asimismo, el valor de tales variables es abstraído del área de estudio a partir de las estaciones meteorológicas, según lo indica Gonzáles, J. (2013), quien también añade que “es una instalación destinada a medir y registrar regularmente diversas variables meteorológicas mediante los instrumentos meteorológicos”, los cuales pueden ser pluviómetro, barómetro, termómetro, evaporímetro, entre muchos otros, diseñados para convertir los fenómenos climáticos en valores



numéricos fácilmente registrables y administrables para representar el cambio del clima a lo largo del tiempo.

Más específicamente, las principales variables meteorológicas son definidas individualmente por Gnoza, N. y Barena, M. (2018) de la siguiente manera; comenzando con la presión atmosférica, sobre la cual expone que “La presión disminuye cuanto más alto subamos en la atmósfera, el aire al nivel del suelo soporta más presión debido al peso del aire que está sobre él.” De manera que, se entiende que . Tal fuerza es ejercida sobre la superficie terrestre, las personas, objetos y cualquier otro elemento dentro del gas.

Asimismo, también se intuye que la fuerza ejercida varía dependiendo de la columna de gas debajo de la cual se está, lo que quiere decir que la fuerza de la presión atmosférica disminuye con respecto a la altura, por lo que entre mayor altura habrá menor presión hasta el límite superior de la atmósfera, de igual manera, a menor presión, menor concentración de gases y a mayor presión, mayor concentración.

De igual manera, Gnoza y Barberena (2018, p24) establecen que el gas que conforma la atmósfera no es estático, este también es afectado por la radiación solar que varía dependiendo de la posición del sol y de su ángulo relativo a la superficie, lo cual provoca variaciones en la temperatura de ciertas masas de aire y, subsecuentemente, su densidad y su presión atmosférica. Asimismo, al afectar su densidad con respecto a la densidad de las masas de aire adyacentes, la diferencia provoca que ascienda en la atmósfera, provocando variaciones en la presión y movimientos de masas de aire que siempre viajan de zonas de alta presión a las de baja presión, causando así los vientos. Es por ello que la presión atmosférica como variable meteorológica es crucial para la predicción del clima, debido a que funciona como un indicador práctico para el entendimiento de los vientos y la pronosticación de tormentas en determinados climas.

En el mismo orden de ideas, los autores Gnoza y Barberena (2018, p25) se refieren al viento como “consecuencia del movimiento del aire en una dirección y velocidad específica” y, como se mencionó anteriormente, tal movimiento es causado por diversas causas, entre las cuales se encuentra la diferencia de presión atmosférica, donde Gonzáles, J. (2013), especifica que “cuando entre dos zonas la presión del aire es distinta, éste tiende a moverse desde la zona de alta presión a la zona de baja presión”. La diferencia de presión es causada principalmente por la diferencia en temperatura, la cual provoca que el volumen de la masa de aire aumenta al ser calentada y su densidad disminuye, de manera que, por efecto de flotación, la masa de aire ascienda causando que

otras masa de aire tomen su lugar. Tales interacciones representan una relación entre las variables de temperatura, presión atmosférica y radiación solar, las cuales podrían ser observadas e identificadas como patrones causantes de ciertos comportamientos a través del empleo de machine learning.

Sin embargo, el viento también puede ser provocado por el efecto de la rotación de la tierra sobre los elementos que están en ella, también denominado como efecto coriolis. Los autores especifican que “debido a la rotación de la Tierra sobre su eje, se produce una desviación inercial en los vientos hacia la izquierda en el hemisferio sur y a la derecha en el norte, lo cual hace que el viento tienda ser paralelo a las isobaras.” Lo cual tiene un efecto formador en el clima y la geología del planeta, al mover las masas de aire frías a territorios más cálidos y viceversa, así como también transporta las nubes y las precipitaciones a lo largo del territorio, de manera que su impacto sobre las condiciones climáticas de un área en un instante específico puede ser determinante. De manera que su comportamiento es otra variable que puede funcionar como indicador de fenómenos meteorológicos, tal relación puede ser abstraída mediante el modelo de machine learning, al ser alimentado con información sobre las velocidades y orientaciones del viento en un determinado momento. Para ello, se utilizarían métricas que representen su comportamiento usualmente representadas mediante valores de velocidad, como kilómetros por hora (km/h), nudos (kt) que equivale a 1,852 km/h, metros por segundo (m/s) y Beauford (Bft), que se basa en el estado del mar, las olas y el viento.

Seguidamente, se encuentra la variable de la humedad relativa, con un gran impacto en la temperatura percibida, probabilidad de precipitaciones, neblinas y otros fenómenos. Es definida por Gonzáles, J. (2013), como “La humedad es la cantidad de vapor de agua que contiene el aire. Esa cantidad no es constante, sino que dependerá de diversos factores, como si ha llovido recientemente, si estamos cerca del mar, si hay plantas, etc.” Por lo que el valor específico de la humedad en el ambiente va a ser relativo a las cualidades del entorno y los fenómenos que han ocurrido o por ocurrir, como las precipitaciones.

De ese modo, se establecen relaciones identificables entre otras variables y factores medibles que pueden ser observados y contrastados por el modelo de ML, para la mejor predicción de potenciales precipitaciones, neblina o temperatura aparente. Este valor es medido en tres distintas maneras, según lo indica el mismo autor, como humedad absoluta; que es la masa de vapor de agua que contiene un volumen determinado de aire, como humedad específica; que es la masa de vapor

de agua contenida en una masa determinada de aire y, finalmente, como razón de mezcla, que es la masa de vapor de agua que contiene una determinada masa de aire seco. En el presente proyecto se empleará una cuarta manera de representar la humedad y es la manera relativa, en forma de porcentaje, esta “representa el contenido de vapor de la masa de aire y  $E$  su máxima capacidad de almacenamiento de éste, llamada presión de vapor saturante.” Indicando la máxima cantidad de vapor de agua que puede contener una masa de aire antes de convertirse en líquida, también conocido como saturación.

Como se mencionó anteriormente, la saturación de vapor en el aire tiene un impacto en la temperatura percibida, pero en la realidad, la humedad relativa depende de factores como la presión atmosférica y la temperatura, definida formalmente por Gonzales, J. (2013, pág. 12) como “una magnitud relacionada con la rapidez del movimiento de las partículas que constituyen la materia.” Entendiéndose que, a mayor agitación, mayor temperatura, siendo es una de las variables más importantes en el entendimiento de fenómenos físicos y químicos, donde la meteorología no es la excepción. Dado que el mismo autor indica que la temperatura es una de las magnitudes más empleadas para la descripción del estado del clima, y es afectada por un sinnúmero de factores, como la hora del día, la altura, ubicación geográfica y época del año, y es un pilar fundamental en la comunicación de pronósticos y el entendimiento de fenómenos.

Por lo tanto, la temperatura funciona como un puente de relaciones entre los fenómenos que la causan y los que causa, por lo que el registro de su valor en distintas condiciones, como lo podrían ser diferentes alturas, será de gran importancia para el establecimiento de relaciones entre variables dentro del aprendizaje del modelo de ML. Para ilustrar, Gonzales, J. (2013, pág. 15), expone un ejemplo;

“(…) la superficie terrestre se calienta el aire durante el día, y lo enfría durante la noche. Si es un día despejado y el suelo se ha calentado mucho, la temperatura del aire será elevada. Si por el contrario está nublado y el suelo apenas ha recibido radiación solar, el aire no alcanzará temperaturas demasiado altas.”

Donde es posible observar y entender el impacto que tiene la temperatura sobre otras variables y la cantidad de información que puede brindar en relación a otros fenómenos meteorológicos, funcionando como un indicador crucial sobre qué está sucediendo y a causa de qué factor, proporcionando una fuente de información crucial para el mejor desarrollo de un modelo de ML capaz de predecir el clima precisamente.

Finalmente, se encuentran las precipitaciones, que son la variable meteorológica más fácilmente observable y es la que más se desea pronosticar, debido a su alto impacto en las operaciones diarias de cualquier organización o individuo, debido a que estas consisten en “cualquier forma de meteoro formado por agua que cae a la superficie terrestre desde la atmósfera”, según Gnoza, N. y Barberena, M. (2018, pág. 33), esencialmente refiriéndose a las lluvias o lloviznas. Tal fenómeno es el producto de la interacción de todas las variables previamente mencionadas y, cuyo origen es explicado por Gonzales, J. (2013, pág. 33), de la siguiente manera:

Una nube puede estar formada por una gran cantidad de gotitas minúsculas y cristalitos de hielo, procedentes del cambio de estado del vapor de agua de una masa de aire que, al ascender en la atmósfera, se enfría hasta llegar a la saturación.

Exponiendo la relación intrínseca entre las nubes, la humedad y las precipitaciones, las cuales forman parte de un ciclo recursivo donde el agua atraviesa todos los estados de la materia, entendiéndose que el comportamiento de las precipitaciones de una determinada zona es dependiente de las cualidades climatológicas inherentes del área y las condiciones de la época. De manera que, a partir de los comportamientos relativos de las variables meteorológicas expresados en registros históricos, es que el el modelo de ML será capaz de establecer relaciones proporcionales, numéricas y cuantificables que podrá emplear más adelante para la predicción del comportamiento de las precipitaciones. Siendo de gran importancia la variable meteorológica de la precipitación para el pronóstico climático.

Asimismo, la precipitación es abstraída del ambiente y cuantificada en las estaciones meteorológicas, las cuales “utilizan pluviómetros generalmente ubicados en las estaciones meteorológicas, que recogen y miden las precipitaciones caídas en el sitio”, de acuerdo con Gnoza, N. y Barberena, M. (2018, pág. 33). Tales mecanismos expresan la altura de la cantidad de agua que precipitó sobre una superficie de un tamaño específico, es por ello que se expresan en milímetros (mm) de altura, valor que será capaz de indicar de manera exacta la cantidad de agua que precipitó dentro de un periodo determinado al modelo de ML, a partir de lo cual establecerá relaciones con otras variables meteorológicas.

#### **2.2.6 Predicción Meteorológica.**

La predicción meteorológica compone el enfoque principal y el producto final de la investigación presente, esta es definida por la Real Academia Española (2022), como:

Conjunto de actuaciones que, siguiendo una metodología determinada y a través de los resultados de los modelos numéricos de predicción, van dirigidas a definir el valor más probable de los parámetros de tiempo (nubosidad, fenómenos significativos, temperatura, etc.), su intensidad o la distribución espacial y temporal.

Entendiéndose que, la tarea consiste en el entendimiento de los factores que dan lugar a los fenómenos meteorológicos más significativos para la correcta definición de sus valores en el futuro, proporcionando una herramienta para la planificación, prevención y preparación de los individuos y organizaciones en sus operaciones. Tal proceso es realizado mediante modelos numéricos de predicción, que buscan cuantificar las relaciones entre los factores y los fenómenos, donde Gnoza, N. y Barberena, M. (2018, pág. 34), expresan que “El uso de modelos para predecir el clima comenzó en 1922 por el matemático inglés Lewis Fry Richardson, que intentó hacer una previsión numérica sin éxito.”, lo cual se debió a los altos requisitos que conlleva la realización de los cálculos involucrados en los modelos predictivos. Sin embargo, sus esfuerzos no fueron en vano, dado que en 1950, un grupo de meteorólogos logró emplear la primera computadora de propósito general (ENIAC), para la predicción exitosa del clima, resultado en el auge del área gracias a la inversión por parte del sector público norteamericano.

En la actualidad, los métodos predictivos más ampliamente implementados y con mejores resultados “está basada en la resolución de las ecuaciones matemáticas correspondientes a las leyes físicas que describen el comportamiento de la atmósfera.”, según Rodríguez, R., Benito, A., y Portela, A. (2004). Restableciendo la necesidad de emplear computadores con alta capacidad de procesamiento, adaptados con lo necesario para resolver complejas ecuaciones basadas en vastas cantidades de datos meteorológicos registrados, los cuales conllevan un alto e intrínseco nivel de complejidad, requiriendo personal especializado y capacitado. Lo cual resulta en un alto costo operativos y de instalación para nuevas organizaciones o entidades que deseen desarrollar sus propios pronósticos meteorológicos.

## **2.3 Bases Legales**

### **2.3.1 Constitución de la República Bolivariana de Venezuela**

**Artículo 127.** Se establecen diversos derechos y deberes relativos al medio ambiente. Abarcando el derecho y deber a mantener y proteger el medio ambiente, así como el derecho de toda persona a disfrutar de una vida y de un ambiente seguro, sano y ecológicamente equilibrado.

Donde se crea la necesidad de entender el medio ambiente, su desarrollo, los fenómenos que componen sus principales procesos y las reglas que delimitan su comportamiento, es por ello que se desarrolla la necesidad individual y organizacional de un servicio capaz de proveer información predilecta sobre las condiciones climáticas en un determinado instante, con altos estándares de precisión y tiempo de respuesta, empleando la cantidad justa de recursos. Para lo cual, si bien ya existen soluciones tradicionales, es posible la minimización de errores mediante la implementación de nuevas tecnologías, como el machine learning.

### **2.3.2 Ley de Meteorología e Hidrología Nacional**

**Artículo 3.** Se declara de interés general y uso público la información básica meteorológica e hidrológica, la cual se considera patrimonio de la República Bolivariana de Venezuela. La información existente para el momento de la entrada en vigencia de la presente Ley, que se encuentre almacenada o archivada, no podrá ser destruida, ocultada u omitida. La misma debería ser notificada a las autoridades del Instituto Nacional de Meteorología e Hidrología (INAMEH) y remitida en los lapsos que, a tal efecto, se establezcan, con el fin de que sea incorporada al Banco Nacional de Datos Meteorológicos e Hidrológicos.

En el artículo mencionado, se aclara que toda la información recolectada sobre las actividades meteorológicas en Venezuela es de acceso público y no puede ser destruida, ocultada u omitida, de tal manera que pueda ser utilizada para el entrenamiento de algoritmos de predicción meteorológicas.

**Artículo 8.** De acuerdo con esta Ley, los fines del Sistema Nacional de Meteorología e Hidrología (SINAMEH), estarán dirigidos a:

1. Integrar y organizar a todas las personas, órganos y entes públicos que forman parte del Sistema.
2. Estimular y promover los programas de formación y capacitación del recurso humano necesario para el desarrollo meteorológico e hidrológico del país.
3. Establecer programas de incentivos a la actividad de investigación y desarrollo en el área meteorológica e hidrológica.
4. Gestionar vías de financiamiento para las actividades del Sistema.
5. Establecer los mecanismos para distribuir efectivamente las funciones del Sistema entre las personas, órganos y entes que lo conforman.

6. Promover mecanismos para la divulgación, difusión e intercambio de la observación de fenómenos meteorológicos e hidrológicos y de las que surjan como producto de su procesamiento e investigación.
7. Promover y alcanzar la modernización de la red a través de la actualización tecnológica, crecimiento y fortalecimiento de la misma.

De acuerdo con la precitada norma, se entiende que dentro de los propósitos del INAMEH se encuentra la implementación y aceptación de nuevas tecnologías referentes a las áreas competentes de la institución, como lo podría ser la predicción meteorológica mediante el machine learning.

## 2.4 Definición de términos

**Abstraer:** “Separar por medio de una operación intelectual un rasgo o una cualidad de algo para analizarlos aisladamente o considerarlos en su pura esencia o noción”. (Real Academia Española, 2023, definición 1)

**Algoritmo:** “Un algoritmo es una serie de pasos organizados, que describe el proceso que se debe seguir, para dar solución a un problema específico.” (Fadul, 2004).

**Altitud:** “Elevación o altura sobre el nivel del mar.” (Real Academia Española, 2023, definición 1)

**API:** “Son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos” (¿Qué es una API? - Explicación de interfaz de programación de aplicaciones - AWS, s. f.)

**Atmósfera:** “La atmósfera es la capa gaseosa que envuelve la Tierra, y que se adhiere a ella gracias a la acción de la gravedad”. Rodríguez, R., Benito, A., y Portela, A. (2004)

**Columna de aire:** “Volumen de aire, que se extiende desde un punto determinado hasta el tope de la atmósfera.” Rodríguez, R., Benito, A., y Portela, A. (2004).

**Conglomeración:** “Unir fragmentos de una o varias sustancias con un conglomerante, con tal coherencia que resulte una masa compacta.” (Real Academia Española, 2023, definición 2)

**Correlación:** “Medida de la tendencia de la evolución de dos variables” (Real Academia Española, 2023, definición 2)

**Cotejo:** “Comparación y examen de dos cosas para apreciar sus semejanzas y diferencias” (Oxford Languages and Google - Spanish, 2022)

**Cronología:** “Serie de personajes o sucesos históricos por orden de fechas.” (Real Academia Española, 2023, definición 2)

**Datasets:** “Conjunto de datos, ordenado bajo un sistema de almacenamiento que otorga los lineamientos principales de búsqueda o directorio de la información que se quiere trabajar.” (Solis, 2023)

**Diagnóstico:** “Recoger y analizar datos para evaluar problemas de diversa naturaleza.” (Real Academia Española, 2023, definición 1)

**ECMWF:** “ECMWF es el Centro Europeo de Previsiones Meteorológicas a Plazo Medio. Somos a la vez un instituto de investigación y un servicio operativo 24 horas al día, 7 días a la semana, que producimos predicciones meteorológicas numéricas globales y otros datos para nuestros Estados miembros y cooperantes y la comunidad en general.” (Setchell, s. f.)

**Hidrología:** “Disciplina que estudia las aguas de la Tierra.” (Real Academia Española, 2023, definición 1)

**Incertidumbre:** “Falta de seguridad, de confianza o de certeza sobre algo, especialmente cuando crea inquietud.” (Oxford Languages and Google - Spanish, 2022)

**Metodología Incremental:** “Presume la mejora continua en relación con el aporte de valor sobre una base preexistente, y conlleva gradualidad, creatividad y heurística.” (Luna & Gallo, 2018)

**Infraestructura:** “Conjunto de medios técnicos, servicios e instalaciones necesarios para el desarrollo de una actividad o para que un lugar pueda ser utilizado.” (Oxford Languages and Google - Spanish, 2022)

**Intrínseco:** “Íntimo, esencial.” (Real Academia Española, 2023, definición 1)



**Iterativo:** “Es la práctica de elaborar, refinar y mejorar un proyecto, producto o iniciativa” (Martins, 2022)

**Jerarquía:** “Principio que, en el seno de un ordenamiento jurídico, impone la subordinación de las normas de grado inferior a las de rango superior.” (Real Academia Española, 2023, definición 1)

**Jurisdicción:** “Autoridad o poder para juzgar y aplicar las leyes.” (Oxford Languages and Google - Spanish, 2022)

**Legislación:** “Conjunto o cuerpo de leyes por las cuales se gobierna un Estado, o una materia determinada.” (Real Academia Española, 2023, definición 1)

**Magnitudes:** “Es una cantidad medible de un sistema físico a la que se le pueden asignar distintos valores como resultado de una medición o una relación de medidas.” (Pozo, 2008)

**Masa de aire:** “es un volumen de aire definido por su temperatura y contenido de vapor de agua” (Rodríguez et al., 2004)

**Minimizar:** “Reducir considerablemente, o al mínimo, una cosa material o inmaterial, especialmente el valor o importancia de algo o alguien.” (Oxford Languages and Google - Spanish, 2022)

**Modelo:** “Arquetipo o punto de referencia para imitarlo o reproducirlo.” (Real Academia Española, 2023, definición 1)

**Neurona:** “Es una célula importante del sistema nervioso, y su función principal es recibir, procesar y transmitir información a través de señales químicas y eléctricas gracias a la excitación eléctrica de la membrana plasmática.” (Solé y Manrubia, 2009)

**Predecir:** “Anunciar por revelación, conocimiento fundado, intuición o conjetura algo que ha de suceder.” (Real Academia Española, 2023, definición 1)

**Pronóstico:** “Señal por la que se conjetura o adivina una cosa futura” (Garcia & Gross, 1994, p. 462)

**Python:** “Python es un lenguaje de programación interpretado, orientado a objetos de alto nivel y con semántica dinámica. ” (Luca, 2020)

**Radiosonda:** “Conjunto de aparatos registradores automáticos que transmiten desde un globo informaciones meteorológicas por medios radioeléctricos” (Garcia & Gross, 1994, p. 476)

**Retroalimentación:** “La retroalimentación indica un método de control de sistemas, a través del cual, los resultados derivados de una actividad se reintroduce de nuevo en el sistema con el objetivo de mantener un control y una optimización de su comportamiento.” (Software DELSOL, 2021)

**Siri:** “Siri es un asistente personal virtual que viene a resolver una necesidad de los usuarios.” (Onretrieval, 2023)

**Tangente:** “Que toca otra línea o plano en algún punto sin llegar a cortarla.” (Oxford Languages and Google - Spanish, 2022)

**Variables:** “Magnitud que puede tener un valor cualquiera de los comprendidos en un conjunto.” (Real Academia Española, 2023, definición 5)

## **PARTE III**

### **DESCRIPCIÓN METODOLÓGICA**

La tercera parte de la presente investigación comprende su descripción metodológica, la cual es definida por Arias, F. (2012, p. 16) como “conjunto de pasos, técnicas y procedimientos que se emplean para formular y resolver problemas”, por lo tanto, es el segmento en el cual el investigador se dedica a definir los procesos mediante los cuales recogerá, procesará y utilizará información referente al objeto de estudio para la solución de la problemática planteada, es decir, es el “cómo” del proyecto. De esta manera, se abordan y definen aspectos de la investigación, tales como su naturaleza y tipo, conceptos que son productos de la naturaleza del problema en sí y del nivel de profundidad, respectivamente. Asimismo, puntos como el diseño de la investigación son desarrollados, haciendo referencia a “estrategia concebida para obtener la información que se desea” (Sampieri et al., 2003b), el cual es complementado con las técnicas de recolección y análisis de tales datos, obtenidos del objeto de estudio designado.

#### **3.1. Naturaleza de la investigación**

Según Tamayo, M. y Tamayo, Y. (2009, p. 62), el tipo de investigación se define a partir de la naturaleza del problema establecido, los objetivos planteados y los recursos dispuestos. De manera que, tomando en cuenta que el elemento central del presente proyecto es la utilización de información numérica sobre los fenómenos meteorológicos para la construcción de un algoritmo predictivo, se puede concluir que es de naturaleza cuantitativa.

La investigación cuantitativa consiste, según Sampieri et al (2003, p. 4), “en el uso de la recolección de datos para probar hipótesis con base a medición numérica y el análisis estadístico, para establecer patrones de comportamiento y probar teorías.” En efecto, se evidencia el empleo de métodos numéricos para el entendimiento del comportamiento de los fenómenos estudiados a partir de las variables que lo representan, analizando los registros numéricos y desarrollando conclusiones sobre las teorías propuestas.

#### **3.2. Tipo de investigación**

Según Arias, F. (2006, p21), el tipo de investigación se refiere al grado de profundidad con que se aborda un fenómeno u objeto de estudio. En el caso presente, tomando en cuenta la naturaleza

de la investigación y el objetivo, el cual consiste en la propuesta de un modelo predictivo, es lógico concluir que el tipo de investigación se clasifica como un proyecto factible.

Este es definido por Sampieri et al. (2003, p. 9), como “aquellos proyectos o investigaciones que proponen la formulación de modelos, sistemas entre otros, que dan soluciones a una realidad o problemática real planteada, la cual fue sometida con anterioridad o estudios de las necesidades a satisfacer”. En este caso, la formulación del modelo hace referencia a la construcción de un modelo predictivo meteorológico para la solución de determinadas problemáticas relacionadas con la predicción del clima en el Estado Nueva Esparta.

### **3.3. Diseño de la investigación**

Según Sampieri et al (2003, p. 128), “se refiere al plan o estrategia concebida para obtener la información que se desea con el fin de responder al planteamiento del problema.” De manera que, se puede comprender que el diseño de investigación de un proyecto hace referencia a la manera en la que los recursos serán empleados para responder a las interrogantes formuladas en el planteamiento del problema, principalmente refiriéndose a los métodos de obtención de información.

Según diversos autores existen varios diseños de investigación que pueden ser empleados para la aproximación de ciertos tipos de problemas. Sin embargo, al considerar que la información a utilizar será obtenida mediante la revisión de las obras de distintos autores relevantes en el área del objeto de estudio, el tipo de investigación a utilizar en el proyecto presentado es el documental, el cual se define, según Tamayo, M. y Tamayo, Y. (2009, p. 130), como “la que se realiza con base en revisión de documentos, manuales, revistas, periódicos, actas científicas, conclusiones y seminarios y /o cualquier tipo de publicación considerado como fuente de información.” De tal manera, se evidencian las distintas fuentes de información a ser empleadas en la realización del presente proyecto, para la obtención de datos históricos climáticos, la definición del tipo de algoritmo predictivo y el método de aprendizaje empleado, entre otros elementos.

### **3.4. Acopio y selección de la información**

Es un elemento crucial en la estructura metodológica de todo proyecto cuantitativo cuyo diseño es de tipo documental. El cual, según Cázarez, L., Christen, M., Jaramillo E., Villaseñor L. y Zamudio, L. (1999:22), consiste en “reunir, antes que nada, todo el material publicado o inédito sobre el mismo”, en cualquiera de sus formatos académicos, para el correcto proceso investigativo,

debido a que su beneficio se encuentra en que "sabiendo qué datos o ideas se han expuesto anteriormente sobre el tema (...), podrá el investigador partir de bases sólidas para perfeccionar su propio pensamiento y evitar la repetición de ideas". De tal manera, es posible contextualizar la investigación y entender qué en qué áreas indagar y qué ideas proponer.

En el caso de la presente investigación, la recopilación está enfocada en papeles académicos como tesis de grado, artículos científicos y publicaciones en revistas científicas referentes al área principal de interés, el pronóstico del tiempo con machine learning, para el acopio de información referente a las ventajas y desventajas de las distintas técnicas de ML en variados contextos de aplicación dentro de la meteorología. Para la búsqueda y el almacenamiento organizado de dicha información, se hará uso, principalmente, del servicio de internet. Donde se tomará en cuenta factores como la pertinencia de la información en relación a los objetivos, la relevancia referente a la investigación y que no sea mayor a 8 años, debido a la velocidad con la que las tecnologías involucradas evolucionan.

### **3.5. Técnicas e instrumentos de recolección de datos**

La definición de las técnicas de recolección de datos a emplear dependerá de la naturaleza del objeto estudiado. En relación a ello, Tamayo, M. y Tamayo, Y. (2006, p.114) afirman que "La técnica de recolección de datos es la parte operativa del diseño investigativo. Hace relación al procedimiento, condiciones y lugar de la recolección de datos." Por lo tanto, tomando en cuenta el diseño de investigación previamente descrito, se puede afirmar que la técnica que se utilizará en este caso será la revisión documental.

De acuerdo con Sampieri et al (2014, p. 128), tal técnica consiste en "detectar, obtener y consultar la biografía y otros materiales que parten de otros conocimientos y/o informaciones recogidas moderadamente de cualquier realidad, de manera selectiva, de modo que puedan ser útiles para los propósitos del estudio."; cuya definición implica la recolección de documentación a partir de la cual se recaban conocimientos y hechos que fundamentan la teoría y los procesos empleados en la ejecución de la investigación y la constitución del modelo propuesto.

### **3.6. Objeto de estudio**

Como resultado del diseño documental, el proyecto de investigación no posee una población y muestra que proporcionen la información referente a la problemática. En cambio, la información es obtenida a partir del objeto de estudio, el cual es definido por Caraballo, C., Iglesias, L. y

García, F. (2010, p.1) como “la parte más restringida de la realidad sobre la cual recae el problema de investigación y sobre la que actúa el investigador”, dando a entender que el campo, la problemática y los conceptos involucrados en el entendimiento del contexto y la resolución del inconveniente presentado, son los elementos que componen el objeto de estudio, y alrededor de los cuales se basa todo el proyecto investigativo.

De esta manera, el objeto de estudio del proyecto presente engloba la conexión o relación existente entre el pronóstico meteorológico y el machine learning, buscando establecer un fundamento teórico y documental que sustente el desarrollo de un modelo de machine learning capaz de pronosticar de manera precisa las condiciones meteorológicas del Estado Nueva Esparta. Para ello, el objeto de estudio está compuesto por las variables meteorológicas involucradas, las posibles técnicas de machine learning a emplear, la disponibilidad y estado de los registros históricos meteorológicos de la zona y el desarrollo del modelo en sí.

### **3.7. Técnica de análisis de datos**

Según Arias, F. (2012, p. 101), "en este punto se describen las distintas operaciones a las que serán sometidos los datos que se obtengan (...) En lo referente al análisis, se definirán las técnicas lógicas (inducción, deducción, análisis-síntesis), (...)". En tal sentido, la técnica de análisis de datos a emplear debe ser seleccionada en base a las cualidades de los datos obtenidos mediante las técnicas de recolección de datos, por lo tanto, se tomaron en cuenta los objetivos del proyecto presente, el diseño de la investigación y la naturaleza de los datos que se manejarán. En efecto, la técnica de análisis de datos a utilizar en el proyecto presente sería el análisis y síntesis de los datos obtenidos para la construcción de resúmenes.

En un principio, se empleará la síntesis y el análisis para la definición de las variables meteorológicas que serán empleadas en la predicción del clima, definiendo importancia y relaciones entre las mismas. Seguidamente, a partir de métodos estadísticos y analíticos, se obtendrán datos sobre la integridad de los registros históricos, para a partir de ello definir qué periodo será utilizado en el modelo de ML. Después de ello, tales datos atravesarán inspecciones estadísticas que desarrollarán una imagen clara de las cualidades, características y relaciones que guardan los datos entre sí, información que será crucial para la definición de la técnica que será usada en la conformación del modelo de ML, la cual debe estar ajustada a las necesidades de los datos. Finalmente, después del desarrollo, durante el testeo, la verificación de la precisión proveerá

datos cuantitativos que representarán qué tan cerca estuvo el modelo de los datos reales, los cuales proveerán información substancial en cuanto a la utilidad del mismo.

## PARTE IV

### ANÁLISIS DE DATOS Y RESULTADOS

En el presente capítulo, se indaga a profundidad en la información recopilada durante el momento investigativo, sobre los cuales se aplicarán técnicas de análisis para el descubrimiento de perspectivas que puedan producir resultados tangibles que sirvan de fundamento para las conclusiones y recomendaciones a por venir.

#### **4.1 Identificación de las variables meteorológicas más relevantes para el pronóstico del clima en el Estado Nueva Esparta.**

En primer lugar, es fundamental definir las variables meteorológicas que pueden ser empleadas en el desarrollo del modelo predictivo de Machine Learning. Para ello, ha sido necesario tomar en cuenta la fuente de la cual provienen los datos a utilizar, debido a la individualidad con la que distintos sistemas pueden llegar a organizar, categorizar y almacenar datos meteorológicos, así como los formatos y las unidades utilizadas para representar los valores y la estructura de la serie de tiempo en sí, es decir, si los valores medidos de cada variable están estructurados en promedios dentro del lapso de una hora, un día, una semana o cada mes del año, lo que corresponde a la resolución de los datos en el tiempo.

En este caso, para el cumplimiento de los objetivos descritos en este proyecto, es necesario tener acceso a una fuente de datos meteorológicos históricos del Estado Nueva Esparta, Venezuela; de manera que se tengan suficientes registros para entrenar al modelo de ML en el desarrollo de pronósticos de las futuras condiciones del clima en la región. Los registros deben abarcar múltiples variables meteorológicas organizados por un índice que indica el momento en el tiempo donde los valores fueron obtenidos, por lo que para cada instante se tiene un registro de cada variable.

Las opciones disponibles se presentan en forma de API's, las cuales permiten realizar `petición o *request* de datos históricos, tales como: las fechas de inicio y fin, la latitud y longitud de la zona en cuestión, la zona horaria, y la resolución deseada de los datos en el tiempo; cuyas limitaciones, condiciones y costos varían de servicio en servicio. Por tal motivo, se compararon distintas opciones, hasta encontrar la que mejor se adapta a las necesidades del proyecto.



De esta forma, se tomó en cuenta a OpenWeather que, además de ofrecer una API para pronósticos del tiempo, proporciona una alternativa de donde se pueden obtener datos históricos, cuya principal limitante es que solo ofrecen valores de una semana por solicitud, sobre la cual existe una tarifa de costo fija. Por otro lado, también se consideró WeatherBit, quienes ofrecen una API de datos históricos de cualquier región que le fuese estipulada en múltiples tipos de parámetros como latitud y longitud, código postal, código del aeropuerto, nombre de ciudad, entre otros. Sin embargo, posee poca documentación, es necesario establecer un perfil para el uso de una clave identificativa y sólo tiene registros históricos desde 1991.

Finalmente, se decidió por un tercer candidato, la API de Open-Meteo, debido a que es una de las pocas opciones gratuitas, con gran disponibilidad de datos e integración con el lenguaje de desarrollo de la propuesta. Esta ofrece un extenso registro histórico de un diverso repertorio de variables meteorológicas en formato diario o por hora, sin embargo, se decidió por esta opción debido a que dispone de variables más primitivas y específicas, fáciles de correlacionar y comparar, así como más útiles en un sentido práctico. Asimismo, sus registros datan de hace más de 60 años y poseen una extensa documentación para el desarrollo de sus llamadas de API en múltiples lenguajes.

De esa manera, se desarrolló la solicitud a la API histórica de Open-Meteo, donde se incluían los datos de latitud y longitud, las variables que se desean obtener en el lapso del tiempo requerido, en este caso, desde el 01/01/1960 hasta 01/01/2024, abarcando las siguientes variables meteorológicas:

**Tabla N°1.** Listado de variables meteorológicas.

Variable Meteorológica	Definición	Identificador de variable
Temperatura	Magnitud referida a la noción de calor medible mediante un termómetro.	temperatura
Humedad Relativa	La humedad es la cantidad de vapor de agua que contiene el aire.	Humedad_relativa

Punto Dew	Es la temperatura a la que se debe enfriar un cuerpo de aire para saturarse de vapor de agua. Es decir, indica la temperatura mínima para que el vapor de agua se condense y precipite al suelo.	Punto_dew
Temperatura aparente	Medida de cómo realmente se siente cuando se combina la humedad relativa con la temperatura real del aire.	Temperatura_aparente
Precipitación	Cualquier forma de hidrometeoro que cae de la atmósfera y llega a la superficie terrestre. Medido en mm sobre un área determinada.	Precipitación
Presión a nivel del mar	Es la fuerza que ejerce las distintas capas de gas de la atmósfera sobre los objetos dentro de ella, la cual es causada por el efecto de la gravedad sobre la masa de gas. Tomada a nivel del mar.	Presion_mar
Presión Superficial	Es la presión de la atmósfera a la altura de la superficie, tomando en cuenta su altura.	Presion_sup
Covertura de nubes	Porcentaje de la superficie de la región que fue cubierta por nubes en un determinado periodo.	Nubosidad
Factor de Evotranspiración	La pérdida de humedad de una superficie por evaporación directa junto con la pérdida de agua por transpiración de la vegetación	Et0_evot
Déficit de Vapor de Presión	Es la diferencia entre la cantidad de vapor de agua que puede retener la atmósfera y la que contiene en ese momento.	Deficit_VP
Velocidad del viento a 10m	Velocidad del movimiento del aire en una dirección y velocidad específica.	Velocidad_viento

Temperatura del suelo superficial	Temperatura del suelo superficial. (5m)	Temp_t_superficie
-----------------------------------	---	-------------------

*Fuente: Elaboración propia. (2024)*

Seguidamente, es importante diferenciar que no todas las variables empleadas tendrán el mismo propósito en el desarrollo de la propuesta, debido a que su valor agregado se fundamenta alrededor de los factores meteorológicos que los habitantes en general podrían considerar útiles para su rutina diaria. Por ejemplo, no existe una utilidad práctica en predecir variables como la presión atmosférica o la humedad relativa, puesto que no representan nada tangible al usuario común. Sin embargo, tales variables podrían guardar relación con otras que sí sean de utilidad real, como la temperatura en un determinado día, la humedad relativa o cuán nublado estará, relaciones que servirían como indicadores de futuros fenómenos en el desarrollo del modelo predictivo. Por ejemplo, el porcentaje de humedad relativa es una variable con poca utilidad directa, pero podría indicar una mayor probabilidad de precipitaciones o una mayor temperatura aparente, por lo que complementa la habilidad de pronosticar otra que sí sea de utilidad real. En tal sentido, todas las variables están en la capacidad de ser la entrada del modelo predictivo, pero solo un grupo específico estará constituido por las variables de salida, es decir, los valores a pronosticar.

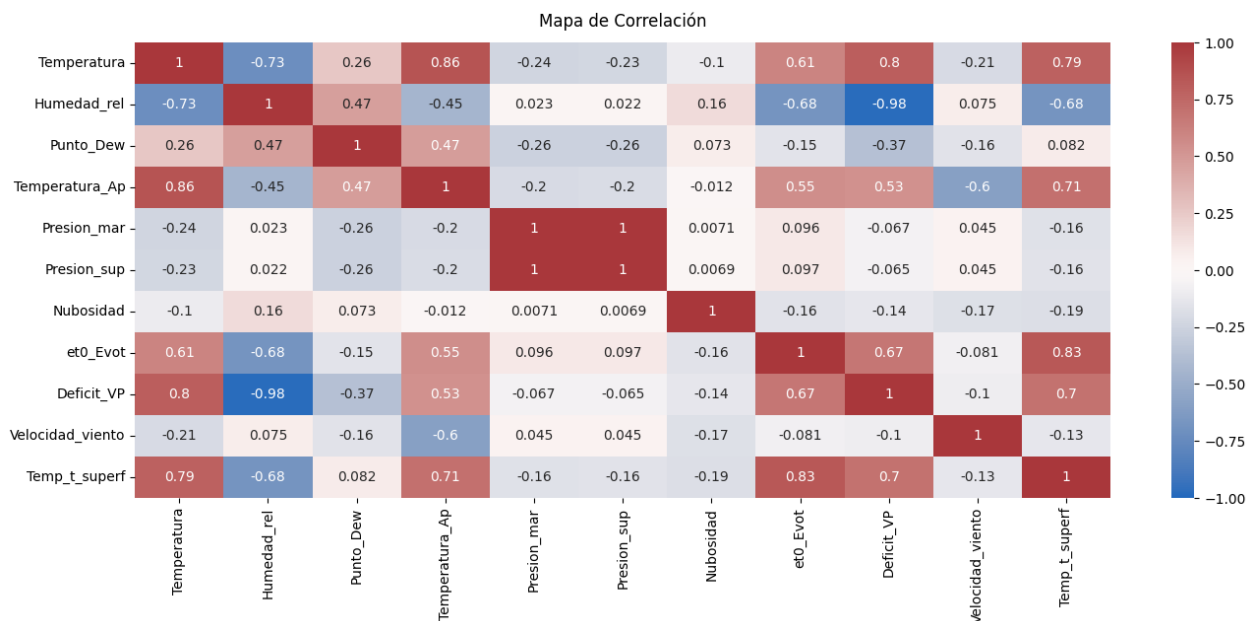
Para el proceso de designación de variables meteorológicas, se decide emplear una técnica de comparación de coeficientes de correlación de Pearson (CCP), buscando cuantificar la relación lineal que pueda existir entre cada una, lo que permite observar si la relación es positiva, es decir, ambas tienden a aumentar juntas, o negativa, lo cual indica que cuando una aumenta, la otra disminuye. En ambos casos, la relación es tomada en cuenta para el modelo de pronóstico, sin embargo, en caso de que la relación se aproxime al valor cero (0), significa que no existe una relación lineal obvia y, por ende, no hay una interdependencia que pueda ser de utilidad.

Esta técnica de CCP es utilizada frecuentemente para un análisis de causalidad donde se desea encontrar los factores más influyentes en el comportamiento de una variable dependiente, usualmente, afectada por múltiples variables independientes simultáneamente. Como lo hace Chairani, S. (2022), quién utilizó la técnica de CCP para encontrar la correlación entre la lluvia, la temperatura, la humedad relativa y la productividad de los campos de arroz en Aceh Besar, concluyendo que “Las precipitaciones y la productividad de los campos de arroz tuvieron una

correlación negativa en Aceh Besar, lo que indica que las lluvias no afectaron la productividad de los campos de arroz en Aceh Besar.” De este modo, desarrolló un esquema de correlación, siguiendo los valores de relación lineal más altos entre todas las variables involucradas en el proceso de producción de arroz en los campos, como la temperatura, humedad, radiación solar, vientos, fertilizantes, entre otros.

Para aplicar esta técnica, se debe calcular la covarianza entre el par de variables que se desea comparar, y esta es dividida por el producto de la desviación estándar de las mismas variables, obteniendo una medida normalizada de la covarianza, es decir, entre -1 y 1. Al aplicar tal técnica para cada par posible de las variables meteorológicas, se obtienen los siguientes resultados:

**Gráfica N°1.** Mapa de correlación de variables.



**Fuente:** Elaboración propia. (2024)

En la gráfica N°1, se pueden observar los coeficientes de correlación de Pearson entre cada par posible de variables meteorológicas, es decir, todas con todas. Por ello existen 144 valores distintos. Sin embargo, a lo largo de la diagonal de la matriz, se encuentran las relaciones entre la misma variable, por ejemplo, temperatura con temperatura, cuyos coeficientes de correlación siempre serán 1 y por ello no es tomado en cuenta. Seguidamente, a ambos lados de esa diagonal se encuentran los mismos grupos de valores, pero sus variables invertidas, cuando en un lado una

relación es humedad con temperatura, al otro lado está temperatura con humedad, lo cual resulta el mismo coeficiente, es por ello que solo se toma en cuenta uno de los dos.

Tomando todo ello en cuenta, se puede empezar a entender lo que cada celda y su valor representa para el análisis de las relaciones. Como se mencionó previamente, los valores rondan entre 1 y -1, representando la normalización de la covarianza entre las dos variables que interceptan en cada celda, de manera que se puede observar la interdependencia, o la ausencia de ella, entre cada par posible de variables. Tomando como ejemplo la celda en la que la temperatura y la humedad relativa interceptan, se observa un valor de -0.66, donde se puede llegar a entender que ambas variables se encuentran medianamente correlacionadas negativamente una con la otra, lo que quiere decir que, cuando el valor de una aumenta, el de la otra disminuye, generalmente.

Paso seguido, se requiere enlistar las relaciones lineales entre las variables, tomando en cuenta que es necesario eliminar las relaciones entre la misma variable, los duplicados (donde las variables están invertidas) y, finalmente, las relaciones cuyo coeficiente de correlación sea menor a 0.5 o mayor a -0.5, lo que refleja que no es suficiente para que exista alguna relación lineal proporcional o inversa. En efecto, a continuación se exponen los resultados:

**Tabla N°2.** Listado de correlaciones de variables.

Variables		Coeficiente Correlacional
Temperatura	Deficit_Vapor_Presion	0.76428166
Temperatura	et0_Evotranspiracion	0.59113984
Temperatura	Humedad_relativa	-0.6592877
Humedad_relativa	Temperatura_tierra_superficial	-0.5695188
Humedad_relativa	et0_Evotranspiracion	-0.6939663
Humedad_relativa	Deficit_Vapor_Presion	-0.9771219

Variables		Coeficiente Correlacional
Temperatura_Apar ente	Deficit_Vapor_Presion	0.58462456
Temperatura_Apar ente	et0_Evotranspiracion	0.55422762
Temperatura_Apar ente	Velocidad_viento	-0.6471305
Presion_mar	Presion_superficie	0.96945521
et0_Evotranspiraci on	Temperatura_tierra_superficial	0.71506821
et0_Evotranspiraci on	Deficit_Vapor_Presion	0.69376578
Deficit_Vapor_Pres ion	Temperatura_tierra_superficial	0.63896963

***Fuente:*** Elaboración propia. (2024)

De esta forma, se logra identificar la cualidad principal que sirve de base para seleccionar las variables meteorológicas a emplear en el desarrollo del modelo de pronóstico, ya que permite determinar cuáles son las que tienen mayor influencia en su comportamiento, en función de los resultados esperados, sentando los cimientos sobre los cuales se realizarán las siguientes tareas competentes a la preparación de datos y la construcción del modelo de ML.

Al observar las relaciones, se realiza un conteo de elementos distintos, para identificar todas las variables involucradas en las relaciones con mayor correlación, resultando en una lista de variables expuestas a continuación, las cuales serán utilizadas en el desarrollo del modelo final a partir de la técnica seleccionada en función de los requerimientos determinados:

**Tabla N°3.** Listado de variables meteorológicas seleccionadas para el modelo ML.

Variable meteorológica
Temperatura
Humedad_relativa

Presion_mar
et0_Evotranspiracion
Deficit_Vapor_Presion
Velocidad_viento
Presion_superficie

***Fuente:*** Elaboración propia. (2024)

Es importante aclarar que se descartan las variables de temperatura\_superficial y temperatura\_aparente, debido a la alta similitud que poseen con la variable temperatura, convirtiéndose en un exceso innecesario de valores que provocaría una carga innecesaria de procesamiento al momento de preprocesar datos y entrenar el modelo.

Asimismo, se decide observar la variable precipitación debido a la naturaleza de los datos que almacena, los cuales consisten en la cantidad de mm que precipitó sobre la superficie terrestre en el instante que indica el índice. Sin embargo, debido a que en la región no llueve de manera frecuente, los valores son principalmente “0mm”, como se observa en la siguiente tabla:

**Tabla N°4.** Descripción de la variable *precipitación*.

Variable	Precipitación
Cantidad de valores	561048
Promedio	0.036245
Desviación Estándar	0.168080
Valor mínimo	0.000
Valor máximo	19.1

***Fuente:*** Elaboración propia. (2024)

Donde es posible observar que el valor máximo son 19mm de agua que precipitaron en un determinado instante, y el valor mínimo es 0mm ya que no puede llover una cantidad negativa de milímetros de agua. Sin embargo, es el muy bajo promedio de los valores de la precipitación que indica que la mayoría de los valores son cero o muy bajos, por lo que se intuye que la mayoría de los valores son ceros representando a los días que no llovió, acompañados de esporádicos valores decimales que representan los días en los que sí llovió y en qué medida, se entiende que la

naturaleza de los datos de precipitación requieren un modelo capaz desarrollado para clasificar los días que si llovió y los que no llovió, no un modelo de regresión lineal o de red neuronal que buscar continuar la frecuencia y la escala de los valores proporcionados.

Es así, que se logra obtener las variables sobre las cuales se desarrollan las técnicas de análisis necesarias para el entendimiento de la naturaleza de los datos y poder determinar un criterio sobre el cual definir qué modelo de pronóstico emplear que pueda proporcionar un resultado más preciso y eficaz durante las siguientes etapas del presente trabajo.

#### **4.2 Selección del periodo de datos meteorológicos históricos del Estado Nueva Esparta a utilizar para el entrenamiento del modelo predictivo.**

A la hora de seleccionar el periodo de datos meteorológicos históricos a emplear en el desarrollo del modelo de ML, se desea que estos estén organizados en una serie cronológica continua, sin interrupciones ni ausencia de datos. Para ello, es importante el empleo de técnicas de exploración de datos, como la descomposición de una serie temporal y los rangos intercuartílicos, que permitan ilustrar su naturaleza a lo largo de la secuencia temporal, permitiendo visualizar valores atípicos que podrían desviar promedios o tendencias, los valores vacíos, incorrectos o de distinto formato, los datos irrelevantes y los duplicados; de forma tal que se puedan aplicar procesos para la corrección de tales faltas o, directamente, la remoción de las mismas como último caso, debido a que en el desarrollo de un modelo predictivo es importante mantener la continuidad de los datos históricos.

Es por ello, que a este punto se realiza la solicitud de datos a la API de Open-Meteo, indicando la latitud y longitud de la región correspondiente al estado Nueva Esparta, y la variables previamente descritas. Para así poder realizar las correspondientes evaluaciones que determinarían el segmento de datos sobre los datos reales que serán utilizados en el modelo de pronóstico más adelante.

De tal manera, se decide evaluar cada una de las series temporales de cada variable meteorológica para verificar que estas no contengan valores nulos o vacíos que puedan intervenir en el entrenamiento de un modelo de pronóstico, debido a que una laguna o salto en la secuencia temporal podría afectar la manera en la que el modelo entiende los fenómenos meteorológicos, mediante métodos integrados de la plataforma de desarrollo, resultando en:



**Tabla N°5.** Cantidad de valores nulos por variable.

Variable	Suma de Valores Null
date	0
Temperatura	0
Humedad_rel	0
Presion_mar	0
Presion_superficie	0
et0_Evot	0
Deficit_VP	0
Velocidad_viento	0

***Fuente:** Elaboración propia. (2024)*

Al verificar que no hay ausencia de datos o discontinuidad de los valores de cada variable, nace la necesidad de entender cuáles son los valores máximos, mínimos y promedios, ya que este procedimiento permite observar la realidad de los valores atípicos, lo cual se logra calculando el valor promedio de cada una de las variables a lo largo de toda la serie de tiempo. Además, esto facilitará la obtención de la desviación estándar, al promediar la diferencia entre el promedio y cada uno de los valores, de manera que, en conjunto con los valores máximos y mínimos que cada variable alcanza, se puede obtener una idea clara de la distribución de los valores a lo largo de la media. En tal sentido, si existe una considerable diferencia entre el promedio y los valores máximos y mínimos, es razonable asumir la existencia de valores atípicos. En la siguiente tabla es posible observar los atributos calculados para cada variable:

**Tabla N°6.** Descripción de distribución de variables.

Variable	Promedio	Desviación std.	Valor mínimo	Valor máximo
Temperatura	26.35	1.46	21.23	35.98
Humedad_rel	81.99	6.90	22.72	97.63
Punto_Dew	22.95	1.13	8.38	26.63
Temperatura_Ap	28.33	2.42	20.89	40.72
Presion_mar	1012.33	1.89	1003.50	1019.60
Presion_sup	1011.63	1.89	1002.82	1018.90

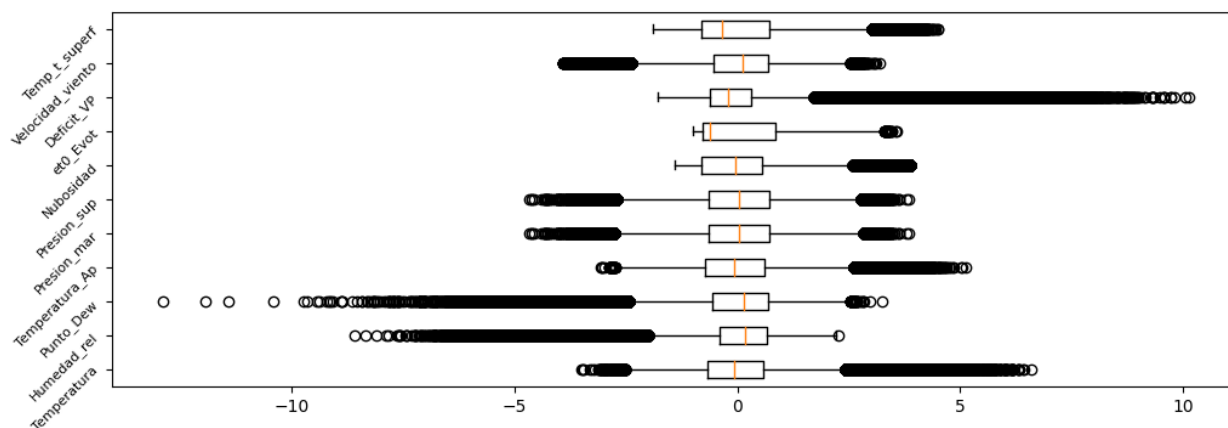
Nubosidad	26.44	18.95	0.00	100.00
et0_Evot	0.19	0.19	0.00	0.89
Deficit_VP	0.64	0.32	0.07	3.85
Velocidad_viento	25.86	6.65	0.00	47.09
Temp_t_superf	30.97	5.15	21.18	54.18

***Fuente:*** Elaboración propia. (2024)

Es así que se pueden observar irregularidades con múltiples variables, mediante la comparación de su promedio en contraste con su valor máximo o mínimo alcanzado, como por ejemplo: la variable temperatura, cuyo promedio indica un total de 26.35 C° de toda la historia de la región, valor que es consistente con la naturaleza tropical de la zona. Sin embargo, al comparar tal valor promedio con el máximo alcanzado en un instante, siendo de 35.98 C°, se concluye que la posibilidad de valores atípicos es alta, en especial tomando en cuenta que los datos son promedios por hora, indicando que ese punto máximo ocurrió en el transcurso de una hora.

Es por ello que se continúa el análisis de la distribución de las variables con el desarrollo de un diagrama de caja, el cual consiste en “un gráfico que resume un conjunto de datos. La forma del diagrama de caja muestra cómo se distribuyen los datos y también muestra los valores atípicos” (Latam, A. 2023); por lo que es capaz de ilustrar la distribución normal de las variables y, por ende, aquellos valores que se encuentran fuera del margen de la misma, como se muestra a continuación:

**Gráfica N°2.** Diagrama de caja de variables meteorológicas normalizadas.



***Fuente:*** Elaboración propia. (2024)

La línea naranja dentro de cada recuadro es la media estándar de cada variable meteorológica, es decir, su valor promedio, alrededor de la cual las desviaciones típicas se ubican, siendo representados por el resto del recuadro blanco, el cual abarca el 25% de la distribución a la izquierda de la media y el 75% a la derecha de la media, es decir, el primer cuartil y el tercer cuartil. Por otro lado, las líneas con limitantes a ambos lados de la distribución central representan los valores mínimos y máximos, es por ello que todos aquellos puntos (o círculos negros en este caso), representan valores atípicos o que no parecen seguir la distribución de la mayoría de los valores.

Tomando en cuenta la clara existencia de un gran número de valores atípicos, se decide contabilizarlos para cada una de las variables, sabiendo así exactamente cuántos valores, en proporción a la cantidad total, están fuera del margen típico. Para ello, se emplea la misma técnica que desarrollan las gráficas de caja, que es denominada IQR o rango intercuartílico, la cual identifica todos los valores que se encuentren bajo el punto  $Q1 - 1.5 \text{ IQR}$  o mayor al punto  $Q3 + 1.5 \text{ IQR}$ , es decir, todos los valores atípicos que se pudieron visualizar en las gráficas de caja, para ser contabilizados e identificados; resultando en:

**Tabla N°7.** Cantidad de valores atípicos por variable meteorológica.

Variable	Cantidad
Temperatura	12777
Humedad_rel	22374
Punto_Dew	11903
Temperatura_Ap	6291
Presion_mar	2726
Presion_sup	3117
Nubosidad	11114
et0_Evot	41
Deficit_VP	23723
Velocidad_viento	15592
Temp_t_superf	2187

**Fuente:** Elaboración propia. (2024)

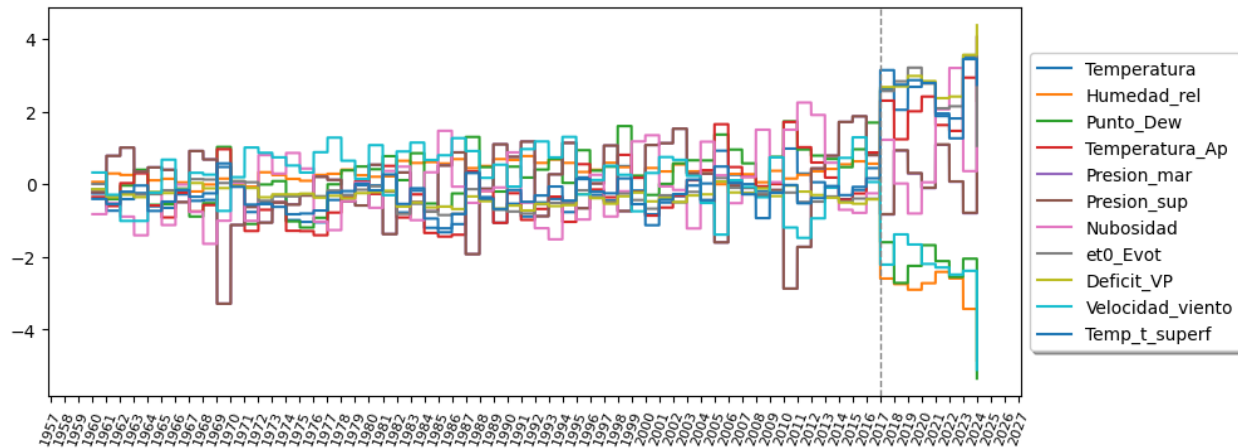
Es así que se puede verificar la existencia de valores atípicos en la distribución de cada una de las variables, entendiéndose que en ellas existe una cantidad considerable demasiado alta como para ser ignorada.

Por tal motivo, se desea observar el comportamiento de cada una de las variables a lo largo de la serie temporal en relación a su media, de manera similar en la que se observó en la Tabla N°3. Sin embargo, si los valores de cada una son representados gráficamente en conjunto, no se podrían comparar directamente, ya que los rangos dentro de los que cada variable existe son diferentes, como se pudo observar con el promedio de la variable temperatura y la variable Humedad. Es por ello que se emplea la normalización de variables, la cual consiste en, “una forma de escalar y transformar los datos para que estén en un rango común, independientemente de la escala original de los datos” (Blanco, J. 2023). Esencialmente, reemplazando el valor promedio de cada secuencia de variable por 0 y la varianza relativa a la medida es representada proporcionalmente entre 0 y 1, tanto positivo como negativo. Para lograrlo, se utilizan los atributos estadísticos calculados en la Tabla N°3, para aplicar la siguiente fórmula a cada dato de la serie temporal:

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

Donde  $X_{normalizada}$ , es el dato individual ya normalizado,  $X$  representa al dato individual en sí y  $X$  con los subíndices max y min representan los valores máximos y mínimos que tal variable alcanzó; resultando en otro conjunto de datos normalizados que pueden ser graficados y superpuestos en una serie temporal organizada cronológicamente por cada año individual. De tal manera, se puede observar si el fenómeno causante de los valores atípicos está atado a un instante en el tiempo, es decir, los valores de las variables cambian inesperadamente y fuera de rango a partir o durante un periodo. La respuesta a ello se puede observar en la siguiente gráfica:

**Gráfica N°3. Serie temporal normalizada.**



***Fuente: Elaboración propia. (2024)***

En la gráfica previa se pueden observar las variaciones de cada variable por año, con respecto a su media, y es directamente comparable al mismo aspecto de las demás variables. Si bien es una gráfica poco práctica para entender el comportamiento de cada variable individualmente, permite ilustrar el comportamiento de todas como un conjunto interdependiente, siendo posible observar que a lo largo de la mayoría de la serie temporal se han mantenido en un rango de variación constante, es decir, no existen picos o valles que puedan resaltar como valores atípicos a excepción de algunos casos.

Sin embargo, desde el año 2017, es posible observar un cambio drástico en la tendencia de algunas variables, lo cual ha provocado que el rango dentro del cual las variables existen cambie drásticamente. Por ejemplo: en ese caso, el valor promedio de la velocidad de viento disminuye significativamente con respecto a su media normalizada de 0, así como el déficit de presión de vapor aumenta considerablemente en relación a su media, comportamiento que se ve reflejado de una u otra manera en el resto de las variables. Es a tal comportamiento errático que se le atribuyen las irregularidades observadas en la Tabla N°3 y Tabla N°4, con respecto a los extremos valores mínimos y máximos de algunas variables, así como la cantidad de valores atípicos detectados mediante el uso de la técnica IQR.

Por tal razón, los datos meteorológicos planteados después del año 2017 no serán tomados en cuenta para el desarrollo del modelo de ML, ya que el inexplicable cambio de tendencia influiría

negativamente en la capacidad real de pronosticación, considerando que esta se ve moldeada por la evaluación de patrones de los hechos previamente documentados.

Para verificar la efectividad de este cambio, se realizó la prueba de IQR en los datos meteorológicos, excluyendo los valores del año 2017 en adelante, para así observar la cantidad de valores atípicos que existen durante el periodo restante. Resultando en:

**Tabla N°8.** Cantidad de valores atípicos por variable meteorológica en múltiples periodos.

Variable	1960 - 2016	1960 - 2024	Diferencia
	Cantidad	Cantidad	
Temperatura	1307	12777	11470
Humedad_rel	2852	22374	19522
Punto_Dew	5862	11903	6041
Temperatura_Ap	4552	6291	1739
Presion_mar	3248	2726	-522
Presion_sup	3248	3117	-131
Nubosidad	9116	11114	1998
et0_Evot	0	41	41
Deficit_VP	3466	23723	20257
Velocidad_viento	16235	15592	-643
Temp_t_superf	1279	2187	908

***Fuente:** Elaboración propia. (2024)*

Es así que se puede observar como en la mayoría de las variables la cantidad de valores atípicos disminuyeron significativamente, así como en otros casos se observa que aumentaron en menor medida. Si bien esto último puede ser contraintuitivo, es importante tomar en cuenta que el criterio que decide cuáles valores son o no atípicos depende íntegramente de la distribución de los valores de las variables, a partir del cual las ecuaciones del IQR generan los rangos intercuartiles que abarcan los valores típicos. En efecto, al eliminar el periodo del año 2017 en adelante, donde se observan tendencias en todas las variables, es posible que la media de los valores de esas variables haya cambiado y, por ende, sus rangos intercuartiles incluyan menos valores, tomando a una mayor cantidad de valores como atípicos. Tal teoría puede ser confirmada mediante la comparación de los

límites de los rangos intercuartílicos antes y después de la remoción de los datos, tal y como se observa a continuación:

**Tabla N°9.** Diferencia de rangos intercuartílicos entre periodos estudiados.

	1960 - 2024		1960 - 2016		Diferencia	
Rango	0.25	0.75	0.25	0.75	0.25	0.75
Presion_mar	1011.09998	1013.70001	1011.09998	1013.59998	0.00000	-0.10004
Presion_sup	1010.40558	1012.99912	1010.40582	1012.90759	0.00024	-0.09152
Velocidad_viento	22.26477	30.43012	22.96394	30.73523	0.69917	0.30511

***Fuente:** Elaboración propia. (2024)*

De tal manera, es posible observar el impacto que la remoción de la tendencia final en la serie temporal tiene en los rangos intercuartílicos que dictan o determinan cuáles valores son atípicos o no, lo que explica el porqué en ciertos casos la cantidad de valores atípicos puede aumentar cuando la intención original es disminuirlo. Al final, no existirá técnica estadística o método matemático que logre agrupar o determinar con total falta de error cuáles valores realmente son atípicos y cuales pueden atribuirse a la naturaleza impredecible de las variables meteorológicas que el proyecto presente busca pronosticar. Es por tal razón, que se emplearán los datos desde el punto más antiguo (1960) hasta el año 2016, para evitar la influencia negativa de los valores atípicos observados. Asimismo, no se aplicarán más técnicas que manipulen los datos meteorológicos que serán alimento del modelo de ML, así como por la ausencia de otra causa probable que indique un evidente problema con la veracidad o integridad de los datos.

#### **4.3 Determinación de la técnica de machine learning más adecuada para el pronóstico meteorológico en el Estado Nueva Esparta.**

La técnica de Machine Learning a utilizar para el desarrollo del modelo de pronóstico depende íntegramente de las cualidades de los datos que se busca pronosticar. En el presente caso, se toma en cuenta las cualidades estadísticas de múltiples variables de la información que se está manejando, la cual implica que el modelo debe estar en la capacidad de utilizar las relaciones existentes entre las distintas variables y entre el presente y el pasado, mediante el uso de técnicas de análisis gráficos y métodos estadísticos que puedan dar respuesta a las cualidades de los datos.

Asimismo, se considera la gran extensión de tiempo en el que los datos de las variables meteorológicas están distribuidos.

Para lograr una determinación concluyente y fácilmente comprensible de la técnica de ML a utilizar, se decidió evaluar modelos de prueba de las técnicas de pronóstico más utilizadas en la actualidad para problemas de series temporales, comprendidos por: los modelos ARIMA (Autocorrelación Integrada de Media Móvil) y las redes neuronales recurrentes LSTM (Long-Short Term Memory). En tal sentido, los modelos deberán pronosticar una sola variable seleccionada arbitrariamente para simplificar el proceso de prueba, en este caso: la temperatura. De igual manera, se utilizarán las demás variables meteorológicas como refuerzo para la predicción. Es por ello que todas deben ser evaluadas en conjunto.

Primeramente, para el correcto desarrollo de un modelo ARIMA, es necesario verificar ciertas cualidades en los datos y determinar algunos atributos. ARIMA es una abreviatura de Media móvil integrada autorregresiva, cuyo funcionamiento, de acuerdo con Kutzkov, K. (2023), está definido por tres parámetros,  $p$ ,  $d$  y  $q$ ; representado de la siguiente manera:

$$ARIMA(p, d, q)$$

Como se puede apreciar, cada parámetro guarda relación con un elemento de la abreviatura del modelo. En el caso de las siglas AR se refieren a la Autocorrelación, que según Ahmed, I (2023), “Es un tipo de modelo de serie temporal que utiliza los valores pasados de una serie temporal para predecir valores futuros.”, el cual está representado en la función como el parámetro “ $p$ ” y hace referencia a la correlación que tienen los valores del presente con los del pasado. El valor de este atributo puede ser determinado mediante la función de autocorrelación (ACF).

Seguidamente, el atributo “ $d$ ”, en la función de ARIMA, se refiere a la integración del modelo y hace referencia a la existencia de una tendencia o no, es decir, si el modelo es estacionario o no-estacionario. Para ello, se realizará una prueba de Dick-Fuller que indica la tendencia de las distintas series temporales y su estacionariedad podrá ser comprobada con la aprobación o el rechazo de la hipótesis nula que la prueba establece.

Finalmente, el atributo “ $q$ ”, corresponde a las siglas MA, el cual “es un tipo de modelo de serie temporal que utiliza los errores pasados de una serie temporal para predecir valores futuros.” (Ahmed, I. 2023), este componente compara medias calculadas con valores del pasado para encontrar diferencias y usarlas en la predicción de valores futuros, el cual es un atributo que puede



ser determinado al aplicar la función parcial de autocorrelación (PACF) y observar la cantidad de pasos observados.

Es así que, para comenzar con el proceso de selección de técnica y del desarrollo de pruebas, según Korstanje, J. (2023), es necesario realizar un proceso de descomposición, donde se busca extraer múltiples tipos de variaciones del conjunto (Tendencia, estacionalidad y ruido) de datos históricos, las cuales serán utilizadas como parámetros en la construcción de modelos, indicándoles sobre las características estadísticas de los datos que deben comprender. Sus elementos principales son:

- Tendencias: Se refiere a un patrón ascendente o descendente a lo largo de todo el lapso de tiempo.
- Estacionalidad: Se refiere a un movimiento recurrente de la variable a lo largo del tiempo, como lo podría ser la temperatura variando por temporada. Permiten observar el periodo en el que los datos tienden a repetirse a sí mismos, así como las ventas de un negocio podrían repetirse semanalmente o el desempeño de un empleado mensualmente.
- Ruido: Se refiere a la parte de la variabilidad que no puede ser explicada por la tendencia o la estacionalidad. Al momento de desarrollar un modelo matemático que represente el comportamiento de la variable, jamás se conseguirá replicar a la perfección, esa diferencia se atribuye al ruido de los datos.

De tal manera, se aplican técnicas de análisis de datos para ilustrar los tres tipos de variación previamente descritos, a cada una de las variables meteorológicas que se busca pronosticar en el modelo de ML. Primeramente, se procede a observar la existencia de tendencia en las variables. Para ello, se decide aplicar un método estadístico denominado prueba aumentada de Dick-Fuller (ADF), la cual establece una prueba de hipótesis que permite definir sin ambigüedad si un dataset es no-estacionario (posee una tendencia) o no. El cual lo logra mediante la verificación de la existencia de una raíz unitaria en la serie temporal, tal atributo, según Prabhakaran, S. (2022) “es una característica de una serie de tiempo que la hace no estacionaria.” y también explica que una raíz unitaria existe cuando el valor de  $\alpha$  es equivalente a 1 en la siguiente ecuación:

$$Y_t = \alpha Y_{t-1} + \beta X_e + \epsilon$$

Donde, la variable  $Y_t$ , es el valor de la serie temporal en el momento 't' y  $X_e$  hace referencia a una variable explicativa separada, que también es una serie temporal, lo que significa que la

presencia de una raíz unitaria implica que la serie temporal es no-estacionaria. De tal manera, dentro de la prueba Dick-Fuller, se verifica la hipótesis nula de que  $\alpha=1$  en la siguiente ecuación, donde alfa es el coeficiente del primer *lag* (también llamado rezago) en Y, la serie temporal, de acuerdo con Prabhakaran, S. (2022):

$$y_t = c + \beta t + \alpha y_{t-1} + \phi \Delta Y_{t-1} + e_t + \epsilon$$

Es allí donde,  $y_{t-1}$  es equivalente al primer rezago de la serie temporal y  $\Delta Y_{t-1}$  hace referencia a la primera diferencia de la serie en el momento  $t - 1$ .

Así como Prabhakaran, S. (2022) indica, esta prueba tiene una hipótesis nula comparable a la prueba de raíz unitaria previamente mencionada, donde, el coeficiente de  $Y_{t-1}$  es 1, lo que implica la presencia de una raíz unitaria. Si no se rechaza, la serie se considera no estacionaria.

Seguidamente, la prueba aumentada Dick-Fuller, mencionada en un principio, se fundamenta en lo ya descrito con ciertos cambios para una mayor precisión y escrutinio, manteniendo la misma hipótesis nula donde  $\alpha=1$ . Respecto a ello, Brownlee, J. (2020), establece que la prueba ADF posee:

- Hipótesis nula (H0): Si no se rechaza, sugiere que la serie temporal tiene una raíz unitaria, lo que significa que no es estacionaria. Tiene alguna estructura dependiente del tiempo.
- Hipótesis alternativa (H1): Se rechaza la hipótesis nula, sugiere que la serie temporal no tiene raíz unitaria, lo que significa que es estacionaria. No tiene una estructura dependiente del tiempo.

Donde:

- Valor  $p > 0,05$ : No se logra rechazar la hipótesis nula (H0), los datos tienen raíz unitaria y no son estacionarios, es decir, posee tendencia.
- Valor  $p \leq 0.05$ : Rechaza la hipótesis nula (H0), el dato no tiene raíz unitaria y es estacionario, es decir, no posee tendencia.

Para la aplicación del método, se emplea la librería de *Adfuller* en el apartado de herramientas estadísticas del conjunto de librerías matemáticas *StatsModels* para *Python 3.12*, para la obtención de los valores de  $p$ , para cada variable meteorológica. De tal manera, al aplicar el método, se obtuvieron los siguientes resultados:

**Tabla N°10.** Tendencia estadística de variables meteorológicas.

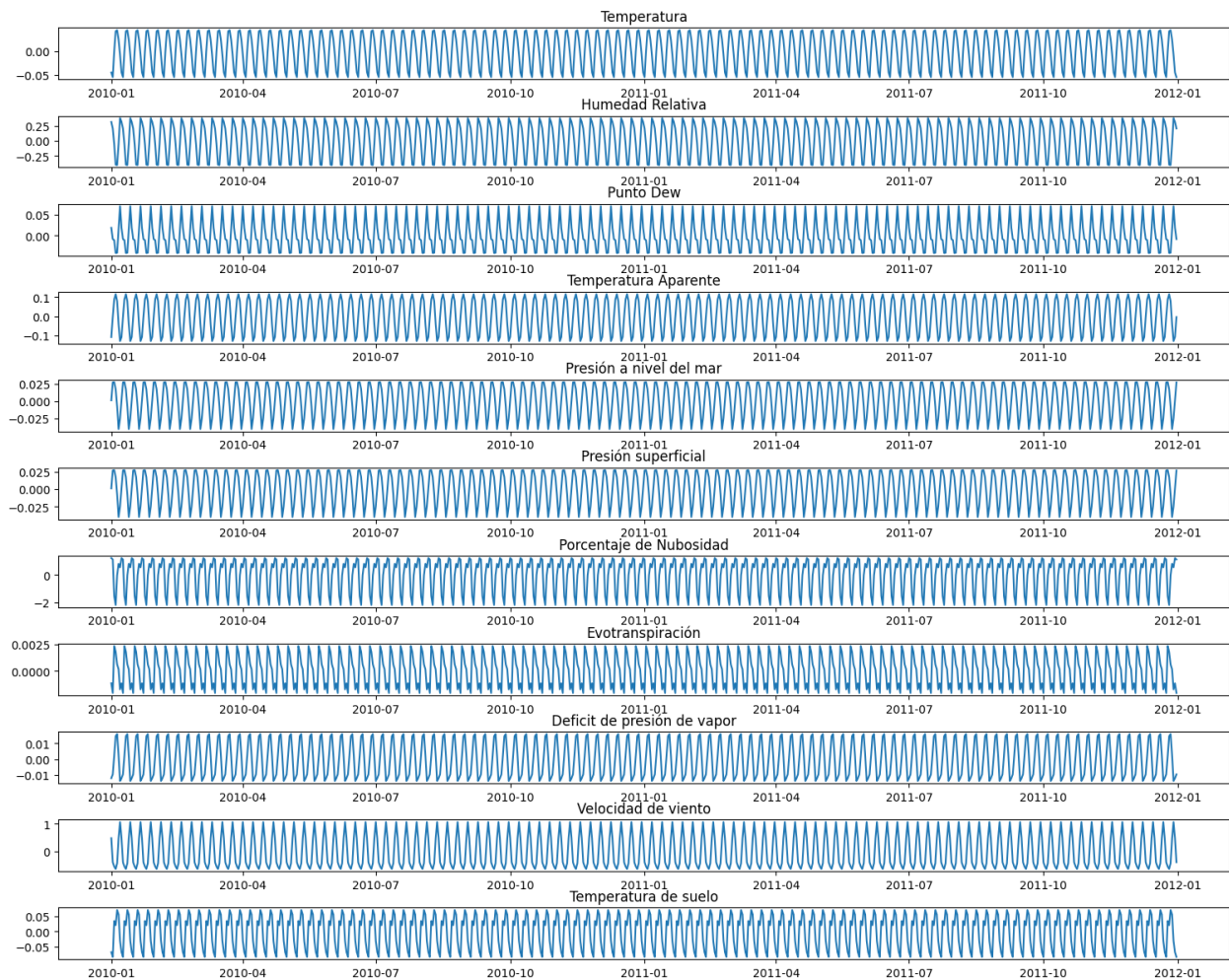
Variable	Valor P
Temperatura	3.09E-21
Humedad_rel	5.59E-06
Punto_Dew	2.03E-22
Temperatura_Ap	6.32E-23
Presion_mar	3.00E-29
Presion_sup	3.00E-29
Nubosidad	2.00E-29
et0_Evot	8.72E-19
Deficit_VP	7.80E-06
Velocidad_viento	5.62E-20
Temp_t_superf	2.36E-20

***Fuente:*** Elaboración propia. (2024)

Como se puede observar, los valores  $p$  que representan la tendencia de cada uno de las variables meteorológicas están expresadas en notación científica debido a lo pequeño que son las magnitudes de tales tendencias, por lo que se puede intuir que las inclinaciones positivas o negativas que las series temporales puedan llegar a presentar están muy por debajo del umbral establecido de  $p \leq 0.05$ . De manera que, se rechaza la hipótesis nula ( $H_0$ ), y estas no tienen raíz unitaria y son estacionarias, es decir, no poseen tendencia alguna. Es por ello, que el valor del atributo “ $d$ ” para la función de ARIMA es igual a 1, indicando una ausencia de tendencia.

Seguidamente, al evaluar la estacionalidad de los datos, se decide observar una pequeña muestra de toda la línea temporal de cada variable, y así apreciar cercanamente el comportamiento a lo largo de un año. Según Malkari, N. (2023) “el componente estacional de una serie de datos de tiempo se refiere a los patrones cíclicos que se observan dentro de un año u otro período de tiempo fijo.”; de manera que, el análisis consiste en comparar los valores entre periodos de tiempo constantes, observando la influencia del momento temporal en el valor de las variables. Para ello, se emplea otra herramienta estadística del conjunto de librerías matemáticas *StatsModels*, denominada *Seasonal Decomposition*, capaz de comparar los valores automáticamente y presentar resultados donde es posible visualizar la presencia de un comportamiento cíclico, resultando en:

**Gráfica N°4. Estacionalidad de variables meteorológicas.**



*Fuente: Elaboración propia. (2024)*

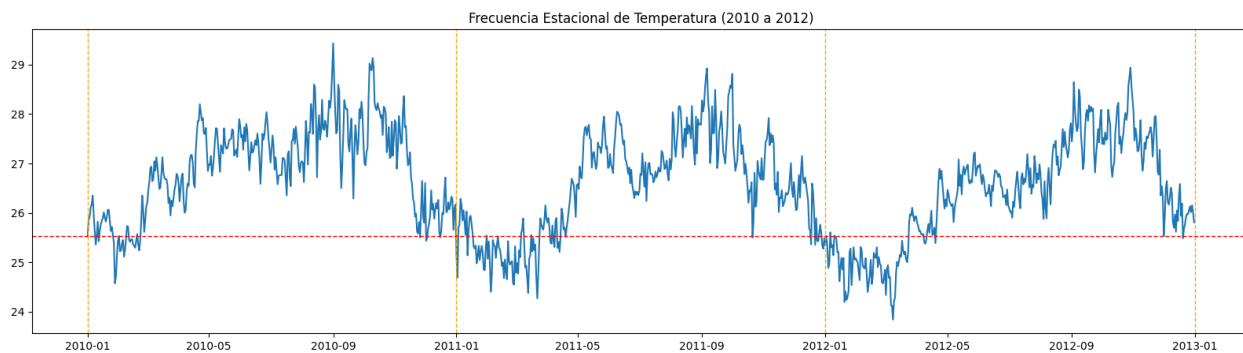
De acuerdo con lo descrito por Malkari, N (2023);

**Si hay una fuerte estacionalidad en los datos, el gráfico estacional mostrará picos y valles claros y consistentes que se repiten a lo largo del tiempo. Si la estacionalidad es débil, el gráfico estacional mostrará patrones más irregulares y dispersos.**

En consecuencia, se puede intuir a partir de la gráfica previa, que el comportamiento estacional de los datos de cada variable meteorológica es correspondiente a una serie temporal de alta estacionalidad, lo que quiere decir que existe una clara e inherente relación entre el comportamiento de las variables y el periodo temporal en cuestión. Este conocimiento es de vital importancia debido a que, al momento de desarrollar un modelo LSTM o ARIMA, es necesario indicar si los datos son estacionales o no, y en caso de que lo sea, dentro de qué periodo, es decir, a lo largo de qué lapso del tiempo el patrón de las variables tiende a repetirse. En este caso, como los datos están

estructurados por día, el valor de frecuencia depende de la cantidad de días que estén dentro del plazo observado, lo que quiere decir que si los valores se tienden a repetir cada semana, entonces la frecuencia sería 7, si se repiten mensualmente, sería de 30. De tal manera, para saber la frecuencia de la serie temporal, basta con observar la variable que se desea predecir, en este caso la temperatura, a lo largo de un par de años, y ver si existe un patrón recurrente allí. Resultando en:

**Gráfica N°5.** Frecuencia Estacional de variable “Temperatura”.



**Fuente:** Elaboración propia. (2024)

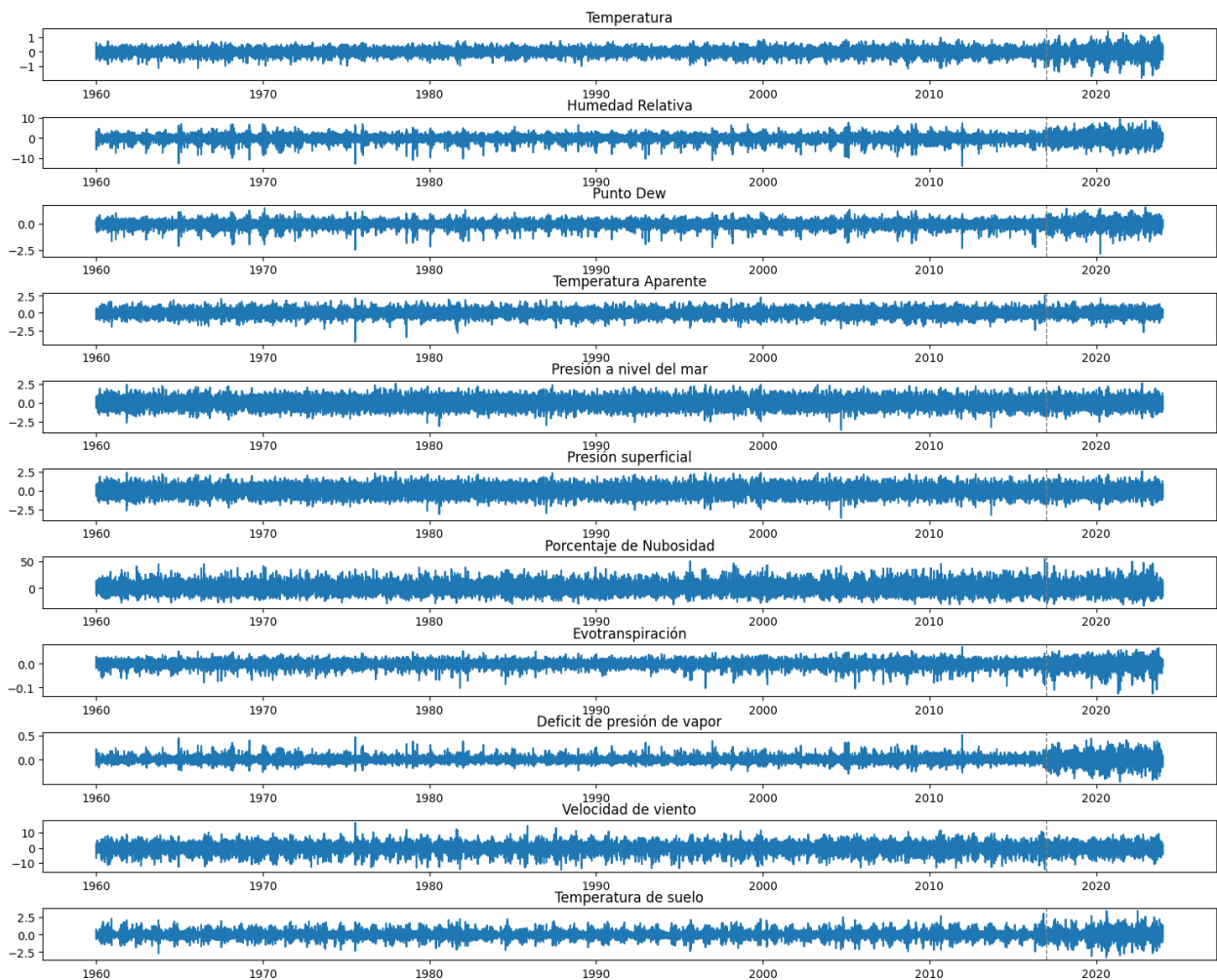
De este modo, se pueden observar lado a lado los valores de temperatura para los años 2010, 2011 y 2012, cuyas líneas de disección verticales representan el comienzo de cada uno de los años y la línea de disección vertical está fijada en el valor de la temperatura al comienzo del año 2010, con el fin de que sea fácil visualizar cómo los valores de la temperatura se repiten a lo largo de cada año en comparación al año anterior, ya que, entre líneas verticales, el comportamiento de los datos es el mismo. Como resultado, se identifica que la frecuencia de la estacionalidad de la variable temperatura es anual y su frecuencia en cantidad de pasos es de 365, ya que los datos están estructurados por día y cada año posee 365 días (aproximadamente).

Posteriormente, después de identificar las variaciones dentro de las categorías de estacionalidad y tendencia, se obtiene como resultado el conjunto de variaciones que no encajan en ninguna de las dos categorías previas, denominadas como resto o ruido. Así como lo indica

Malkari, N. (2023); “El componente residual de los datos de una serie temporal representa la variación aleatoria que queda después de tener en cuenta la tendencia y los componentes estacionales.” Por consiguiente, se le puede atribuir parte del error que se presente al momento de

recrear las variables meteorológicas en un modelo de pronóstico, debido a que estas fluctuaciones no encajan dentro de los patrones más predecibles de estacionalidad y tendencia. Para encontrarlo, se utiliza la misma herramienta del apartado anterior llamada *Seasonal Decompost*, la cual está en la capacidad de aislar las variaciones y presentar el ruido presente en la serie temporal.

**Gráfica N°6.** Residuo de variables meteorológicas.



**Fuente:** Elaboración propia. (2024)

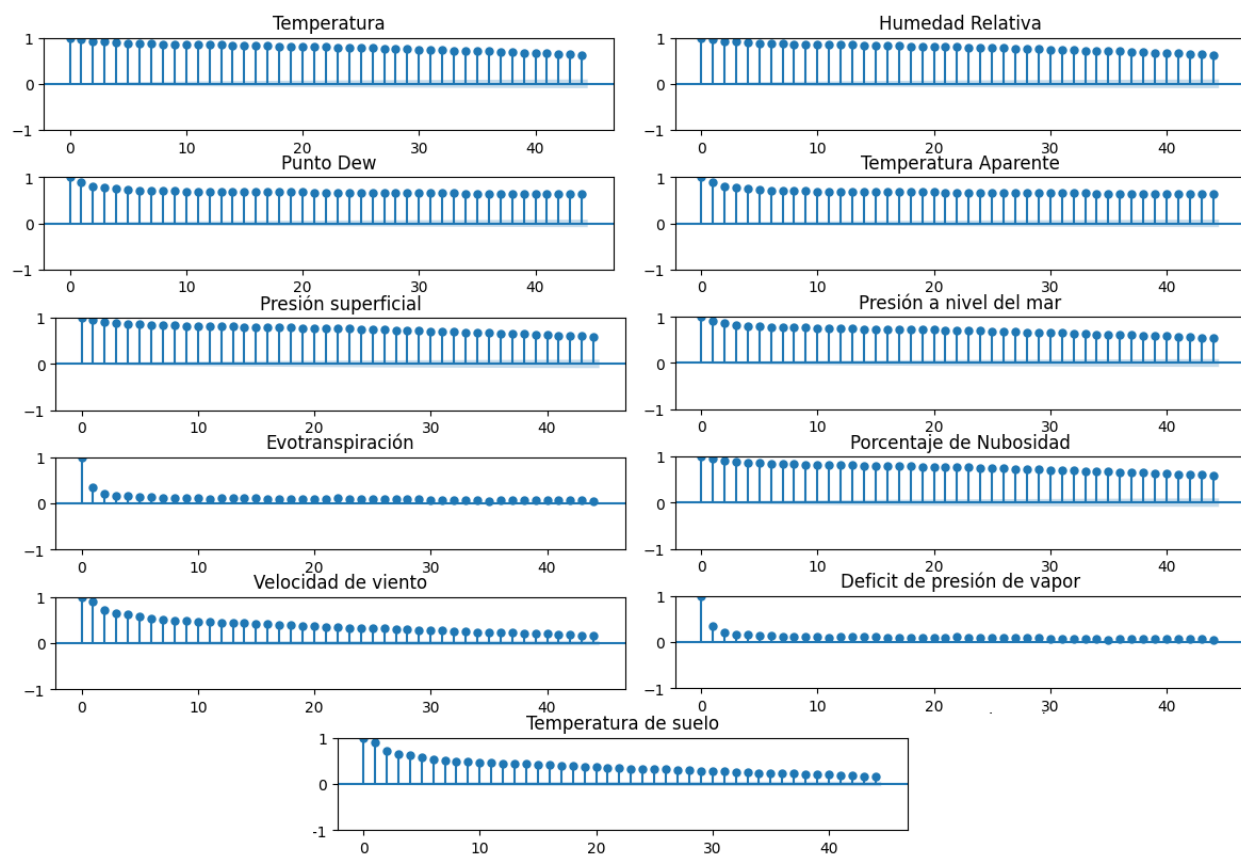
En la gráfica se ilustran las variaciones que abarcan los factores que no son considerados o incluidos en modelos matemáticos que puedan recrearlos fielmente, y también pueden ser consideradas como el error que siempre permanecerá al momento de construir y entrenar el modelo de pronóstico.

Es de tal manera que se puede descomponer los tipos de factores generales que influyen en la variación de los valores de una métrica a lo largo del tiempo, para así comprender su composición, origen y comportamiento, los cuales son factores importante para definir qué técnica de ML utilizar en la predicción de las variables. Sin embargo, es necesario aplicar un paso más de análisis que pueda ilustrar la correlación que los datos meteorológicos del presente tienen con los datos del pasado.

Esta técnica es denominada la función de autocorrelación (ACF), la cual, de acuerdo con Korstanje, J. (2023); “La autocorrelación es la correlación entre el valor actual de una serie temporal con los valores pasados. Si este es el caso, puede utilizar los valores presentes para predecir mejor los valores futuros.” Esto quiere decir que el programa identifica el valor del presente, es decir, el más reciente en los datos, y los compara porcentualmente con cada valor anterior uno por uno, a los cuales se les denomina “paso”. De tal manera, el análisis ilustra la equivalencia de cada valor con el más reciente, en orden inverso al cronológico, para entender la influencia de los datos del pasado sobre los del presente y entender si es posible utilizarlos para pronosticar los del futuro.

Existen dos posibles resultados para la prueba de autocorrelación, positiva o negativa. Al obtener el primer resultado de autocorrelación positiva, se puede inferir que un valor alto en el presente, implica un valor alto en el futuro y viceversa. Sin embargo, en el caso contrario, en el que resulte una correlación negativa, se esperaría lo opuesto, que un valor alto en el presente implique un valor bajo en el futuro y viceversa. Similar al mapa de correlación entre variables meteorológicas de la Gráfica N°1, donde en vez de comparar la correlación entre los datos del pasado y el presente de una misma variable, se compararon los datos entre las variables. Para su ejecución, se empleó nuevamente la librería de *StatsModels*, en cuyo apartado de *TsaPlots*, incluía la herramienta de *Plot\_ACF*, la cual se encarga de correlación los valores dentro de los pasos especificados y presentar los resultados en forma de gráfica, lo cual se efectuó para cada variable meteorológica. Resultando en:

**Gráfica N°7. ACF de variables meteorológicas.**



***Fuente: Elaboración propia. (2024)***

En el eje de las X es donde se puede observar la cantidad de pasos en el pasado, en base a los cuales el método compara, para entender la correlación entre los valores existentes a cierta cantidad de pasos del presente. De manera que, tomando como ejemplo a la gráfica de autocorrelación de temperatura, la primera en el conjunto, se observa que en el paso 0 existe una correlación perfecta de 1, debido a que está comparando un valor con sí mismo. Sin embargo, entre más pasos se va alejando del presente, la correlación va disminuyendo poco a poco. De igual manera, se puede observar una correlación mayormente positiva a lo largo de todos los pasos, indicando que un valor alto en el presente indica un valor alto en el futuro, o viceversa.

Tal información es importante para la definición del modelo de pronóstico a utilizar, debido a que, según Ahmed, I. (2023), “El gráfico ACF se puede utilizar para identificar el orden de un modelo AR.”; por lo que, al entender la autocorrelación, es posible determinar el valor inicial del atributo correspondiente a la Autoregresión en el modelo de ARIMA. Por ello, el valor inicial de



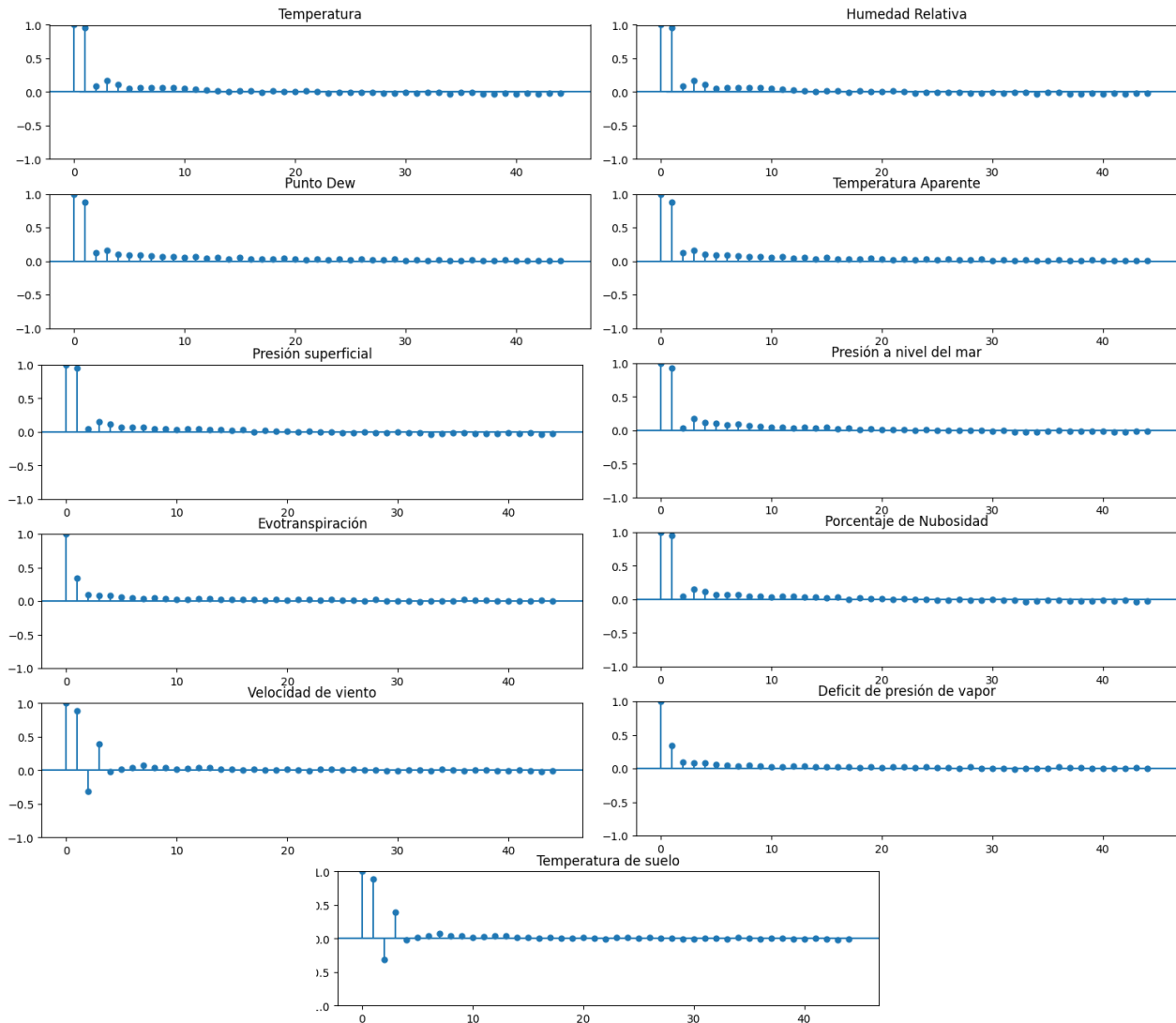
“ $q$ ” que será proporcionado al modelo será de 1, a partir del cual podrá estimar si es correcto o debe ser modificado.

Seguidamente, en contraste con ACF, existe la Función de Autocorrelación Parcial (PACF), las cuales comparten muchas similitudes, cuya principal diferencia es que al comparar los valores de cada paso en el pasado con el presente, la PACF se asegura de eliminar los efectos de los previos pasos en los siguientes pasos, es decir, con cada paso, solo muestra la autocorrelación adicional al anterior. Sobre el cual, Korstanje, J. (2023) indica que “Si el valor de hoy es el mismo que el de ayer, pero también el mismo que el de anteayer, el ACF mostraría dos pasos altamente correlacionados. La PACF solo mostraría ayer y eliminaría anteayer.” Es por ello que, a diferencia de la ACF, la PACF no contiene correlaciones duplicadas cuando la variabilidad puede ser explicada por múltiples puntos en la serie temporal.

Asimismo, su entendimiento tiene un impacto en la definición del modelo a utilizar, ya que la gráfica de la PACF puede ser utilizada para identificar el orden de un modelo MA (Media Móvil, por sus siglas en inglés), el cual “es un tipo de modelo de serie temporal que utiliza los errores pasados de una serie temporal para predecir valores futuros.” (Ahmed, I., 2023). Igual al caso del AR, los valores de la prueba PACF permiten entender los índices de MA, el cual también forma parte del modelo ARMA que se desea probar, como sus últimas dos siglas, donde su valor es representado como el parámetro “ $p$ ”, el cual es equivalente a la cantidad de pasos de autocorrelación observados en la prueba PACF.

Nuevamente, para su ejecución, se empleó la previamente utilizada librería de *StatsModels*, en cuyo apartado de *TsaPlots*, incluía la herramienta de *Plot\_PACF*, que permite correlacionar los datos sin duplicados y observarlos en gráficas intuitivas. Resultando en:

**Gráfica N°8. PACF de variables meteorológicas.**



***Fuente: Elaboración propia. (2024)***

Es así que, mediante el empleo de PACF, se puede confirmar la alta correlación que existe entre cada valor inicial de las series temporales de cada variable y los pasos inmediatamente consiguientes. De acuerdo con lo estipulado por Ahmed, I. (2023) “Si el gráfico PACF muestra picos en los primeros rezagos, entonces un modelo MA puede ser apropiado.”, y así determinar que para el desarrollo de un modelo ARMA, el valor inicial del parámetro “ $q$ ” es equivalente a 1, la cantidad de pasos de autocorrelación parcial observados en la gráfica previa.

Como se mencionó previamente, para simplificar el proceso, se decidió arbitrariamente que en las pruebas de los modelos, la variable a pronosticar sería la temperatura y las características

exógenas (o las que describen a la temperatura) serán la temperatura, humedad, déficit de presión de vapor, punto dew y la evapotranspiración. De tal manera, los datos meteorológicos presenta la siguiente estructura:

**Tabla N°11.** Datos meteorológicos crudos para prueba de modelo.

index	Temperatura	Humedad_rel	Punto_Dew	et0_Evot	Deficit_VP
1960-01-01	25.765831	84.121933	22.865833	0.175073	0.530945
1960-01-02	25.472082	82.250359	22.199165	0.166359	0.584334
1960-01-03	25.305414	75.174706	20.522081	0.199068	0.813798
1960-01-04	25.115831	77.865349	20.957499	0.174493	0.714004
1960-01-05	25.801249	73.842354	20.767916	0.20929	0.872991

***Fuente:*** Elaboración propia. (2024)

Es posible observar la estructura del set de datos o *dataframe*, sobre el cual ambos modelos serán desarrollados para comparar sus capacidades de pronóstico; donde cada fila tiene los valores promedios de cada atributo meteorológico correspondientes a cada día desde 1960 hasta el año 2016, es así cómo puede entender el estado del clima y utilizar las demás variables ajenas a la que se desea pronosticar para descubrir patrones y predecir el comportamiento mediante los procesos internos de cada uno, descritos previamente.

Sin embargo, previo al desarrollo de cualquiera de los modelos, es necesario preparar los datos para lograr explotar más patrones mediante el uso de la ingeniería de características, de acuerdo con Kutskov, K. (2023). En este caso, primeramente, se descompone la compleja estructura de fecha y hora en distintos atributos que, individualmente, engloban cada componente de la fecha, de manera que el modelo pueda observar específicamente la hora, día, mes y año de cada medición de cada variable, en vez de observar la fecha completa, y así se convierte la entrada o *input* al modelo en uno multidimensional, lo que permite dar una mayor perspectiva del comportamiento de las variables. De tal modo, los datos están estructurados de la siguiente manera:

**Tabla N°12.** Datos meteorológicos para prueba de modelo con características de tiempo.

index	Temperatura	Humedad_rel	Punto_Dew	et0_Ev ot	Deficit_V P	year	month	day	day_of _week
1960-01-01	25.765831	84.121933	22.865833	0.1750 73	0.530945	1960	1	1	4
1960-01-02	25.472082	82.250359	22.199165	0.1663 59	0.584334	1960	1	2	5
1960-01-03	25.305414	75.174706	20.522081	0.1990 68	0.813798	1960	1	3	6
1960-01-04	25.115831	77.865349	20.957499	0.1744 93	0.714004	1960	1	4	0
1960-01-05	25.801249	73.842354	20.767916	0.2092 9	0.872991	1960	1	5	1

***Fuente:*** Elaboración propia. (2024)

Es así como la columna *date* es reemplazada por las columnas de *hora*, *día*, *mes*, *año* y *día\_de\_semana*, por lo que se logró descomponer un atributo complejo que habría sido de poca utilidad para los modelos, en sus valores más primarios por separado para que los modelos puedan tener mayor control en el descubrimiento de patrones. Asimismo, la columna de *día\_de\_semana*, representa qué día de la semana fue en esa fecha, donde un *0* equivale a un lunes y un *6* a un domingo. Este atributo es representado numéricamente, debido a que los patrones y la eficiencia permanecen, mientras que hacerlo por nombre no traería ningún beneficio y serían más caracteres que los modelos tendrían que procesar. El proceso de descomposición de la columna *date* se realizó mediante los atributos que este tipo de datos, *datetime*, posee dentro de *Python*, es decir, que cuando el valor de la columna se observaba como una cadena de texto, ésta podía ser descompuesta al acceder directamente a sus atributos de *hora*, *día*, *mes* y *año*, los cuales fueron asignados a nuevas columnas con el mismo nombre, en la misma fila de donde se obtuvieron.

Continuando con la ingeniería de características, otro método para aumentar las dimensiones de los datos de entrada para la mejor explotación de patrones es el desarrollo de características de *lag* o retardo. Al respecto, Gordon, J. (2023) explica:

Al desplazar los valores de una variable hacia atrás o hacia adelante en el tiempo durante un cierto número de períodos de tiempo, las características rezagadas pueden capturar dependencias temporales y tendencias en los datos, proporcionando información valiosa y mejorando la precisión de los modelos predictivos.

Estas características son particularmente prácticas en la identificación de patrones y relaciones entre variables a lo largo del tiempo. Es por ello que son comúnmente usadas en campos de finanzas, economía y pronóstico del clima, así como lo indica Gordon, J. (2023). Para la construcción de los atributos de retardo, se desarrolló una función en *Python* que recibe como parámetros a las columnas del set de datos a partir de las cuales se crearán, es decir, las cuales se retardarán, y la cantidad de períodos distintos que se quieren retardar cada una. En el caso presente, se decide que serían tres periodos: 7 días, 30 días y 365 días; con el fin de que los modelos tengan como referencia los valores de las variables una semana, un mes y un año en el pasado. De tal manera, se obtienen un total de 15 nuevos atributos denominados exógenas, debido a que estos “son predictores que son independientes del modelo que se utiliza para pronosticar.” (Amat, J. y Escobar, J., 2021) Es de tal manera que los parámetros calculados servirán para aumentar la perspectiva y, en consecuencia, la precisión del pronóstico. La cantidad de variables exógenas desarrolladas son las siguientes:

**Tabla N°13.** Variables exógenas de datos para prueba de modelo.

Nombre Variable Exógena	Lapso de retardo	Variable Original
retardado_7_Temperatura	7 días	Temperatura
retardado_7_Humedad_rel	7 días	Humedad_rel
retardado_7_Punto_Dew	7 días	Punto_Dew
retardado_7_et0_Evot	7 días	et0_Evot
retardado_7_Deficit_VP	7 días	Deficit_VP
retardado_30_Temperatura	30 días	Temperatura
retardado_30_Humedad_rel	30 días	Humedad_rel
retardado_30_Punto_Dew	30 días	Punto_Dew
retardado_30_et0_Evot	30 días	et0_Evot
retardado_30_Deficit_VP	30 días	Deficit_VP
retardado_365_Temperatura	365 días	Temperatura
retardado_365_Humedad_rel	365 días	Humedad_rel
retardado_365_Punto_Dew	365 días	Punto_Dew
retardado_365_et0_Evot	365 días	et0_Evot

retardado_365_Deficit_VP	365 días	Deficit_VP
--------------------------	----------	------------

***Fuente:*** Elaboración propia. (2024)

En la anterior tabla, se exponen las variables exógenas desarrolladas mediante el método de retardo, las cuales reflejan el valor de la variable en el lapso especificado en su nombre, es así como, para cada punto en el tiempo, el modelo tendrá a su disposición el valor promedio de cada variable en ese instante y el valor una semana antes, un mes antes y un año antes; lo que facilita el descubrimiento de nuevos patrones mediante la explotación de los datos, mejorando así la capacidad predictiva y su correspondiente precisión.

Como paso final, previo a la construcción de los modelos de pronósticos de prueba de ARIMA y LSTM, se deben incorporar las columnas retardadas en el set de datos meteorológicos que contienen las columnas de tiempo descompuestas. A partir de este punto, se separan los datos en distintos grupos a lo largo del índice o la línea de tiempo, es decir, que todos los grupos de datos tendrán las mismas columnas, pero se referirán a períodos distintos en el tiempo. El objetivo de ello es que estén organizados en grupos de Entrenamiento, Prueba y Validación. Al respecto, Aditya, Y. (2023)., afirma que “conjunto de entrenamiento se usa para ajustar el modelo, mientras que el conjunto de prueba se usa para evaluar su rendimiento.” En tal sentido, se entiende que el propósito del conjunto de entrenamiento es enseñarle al modelo a identificar los patrones y comportamiento dentro de los datos, proporcionando toda la información disponible, tanto la variable que se desea pronosticar, cómo las variables exógenas y todas aquellas que la acompañen. Por otro lado, el conjunto de Prueba nunca es utilizado para entrenar el modelo, es simplemente para contrastar los pronósticos que el modelo desarrolla con lo que en realidad sucedió, para probar su grado de precisión. Finalmente, el conjunto de validación es utilizado como el examen o la prueba, conjunto que es igual al de Prueba, cuya única diferencia es la ausencia de la variable que se busca predecir, en este caso, la temperatura. De manera que el modelo pueda apoyarse en las variables exógenas para rellenar los valores con su estimación de lo que la temperatura debería ser. Es en contraste con esos valores proporcionados por el modelo que se compararon los datos del conjunto de prueba.

En efecto, se toma en cuenta a Aditya, Y. (2023), quién indica que: “Una regla general común es utilizar el 80 % de los datos para el entrenamiento y el 20 % para las pruebas”. Se dividen los datos meteorológicos en tales proporciones, empleando funciones nativas de Python, mediante el

conteo de las filas, se calcula el 80% de ellas y su valor se les asigna a un set de entrenamiento llamado *df\_train* y el resto es asignado a un set de prueba denominado *df\_test*. Finalmente, para crear el conjunto de validación, se realiza una copia de *df\_test*, eliminando la columna de temperatura, la variable que queremos predecir, resultando en un set de datos llamado *df\_valid*; concluyendo así con todos los preparativos necesarios para realizar las pruebas de modelos de pronóstico.

Se inicia con el modelo de tipo ARIMA, como se describió anteriormente, es una abreviatura de Media móvil integrada autorregresiva, la cual, de acuerdo con Kutzkov, K. (2023), el funcionamiento de ARIMA está definido por tres parámetros,  $p$ ,  $d$  y  $q$ . Por lo que, a partir de las pruebas realizadas, la función del modelo con parámetros es ilustrada de la siguiente manera:

$$ARIMA(1,1,1)$$

Sin embargo, para su construcción en Python, se utiliza la librería de *PmdARIMA*, este “es una generalización de una media móvil autorregresiva” (*pmdarima.arima.ARIMA — Pmdarima 2.0.4 Documentation*, s. f.), es decir, permite el desarrollo de las técnicas de AR y MA mencionadas previamente dentro de un entorno de desarrollo de Python. Asimismo, el mismo autor indica que “se ajusta a datos de series de tiempo en un esfuerzo por pronosticar puntos futuros. Los modelos ARIMA pueden ser especialmente eficaces en los casos en que los datos muestran evidencia de no estacionariedad”, verificando nuevamente la compatibilidad del modelo en uso con las necesidades del proyecto y las cualidades evidenciadas de los datos en los pasos previos. Dentro de la librería mencionada, se encuentra la función *auto\_arima*, capaz de ejecutar el modelo a partir de ciertos parámetros ya preparados, como los set de datos de entrenamiento y las cualidades exógenas. A partir de ello, se muestra la búsqueda de cuadrícula sobre diferentes valores de los parámetros  $p$ ,  $d$  y  $q$ , con el objetivo de que concluya en los valores determinados previamente, para verificar la veracidad de las pruebas realizadas. Al final, se devuelve el modelo con el valor AIC más pequeño, el cual, según Kutzkov, K. (2023) “es una medida de la complejidad del modelo que optimiza simultáneamente la precisión y la complejidad de un modelo de predicción”.

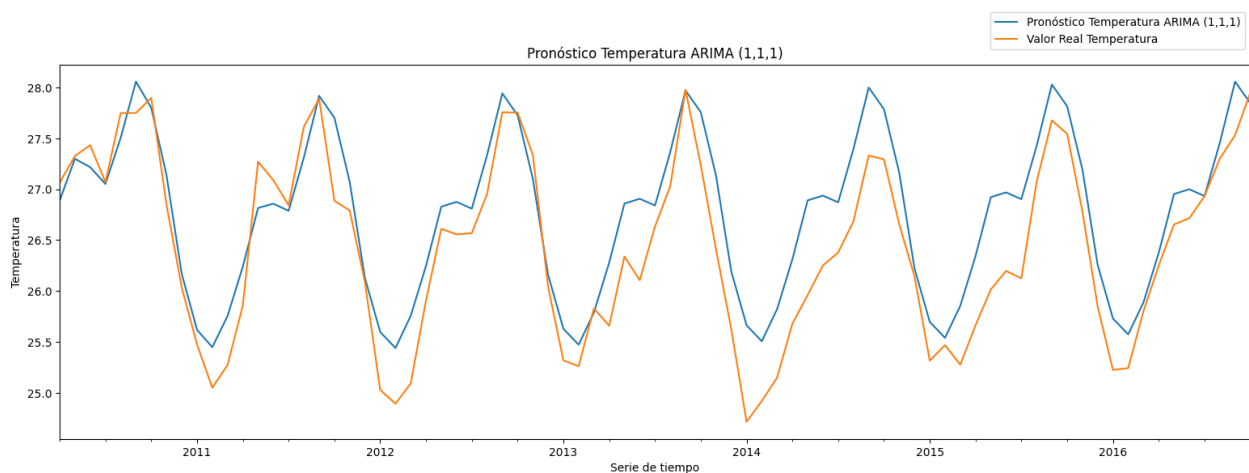
Sin embargo, se presentó un problema en la construcción del primer modelo, relacionado con la cantidad de pasos por temporada que posee, dado que la ejecución de la función *auto\_arima* nunca concluye, debido a la alta cantidad de pasos que posee cada temporada, es decir, 365 días por año.

Según Hyndman, R. (2010), “La función `arima()` permitirá un período estacional de hasta  $m=350$  pero en la práctica normalmente se quedará sin memoria siempre que el período estacional sea superior a aproximadamente 200”, donde  $m$  hace referencia a la cantidad de pasos por temporada. Es por ello que fue necesario cambiar la estructura de los datos de diario a mensual, de manera que cada paso de temporada fuera representado por un mes y así el valor de  $m$  es de 12. Al ejecutar la función `auto_arima`, con los parámetros de estacionalidad, la variable *Temperatura* y los atributos exógenos, se realiza la búsqueda de cuadrícula de los valores  $(p, d, q)$  de la siguiente forma:

```
Performing stepwise search to minimize aic
ARIMA(2,1,2)(1,0,1)[12] intercept : AIC=142.148, Time=3.18 sec
ARIMA(0,1,0)(0,0,0)[12] intercept : AIC=539.851, Time=0.10 sec
ARIMA(1,1,0)(1,0,0)[12] intercept : AIC=255.502, Time=0.51 sec
(...)
ARIMA(2,1,0)(1,0,1)[12] intercept : AIC=146.612, Time=1.05 sec
ARIMA(1,1,1)(1,0,1)[12]           : AIC=inf, Time=0.75 sec
Best model: ARIMA(1,1,1)(1,0,1)[12] intercept
Total fit time: 53.887 seconds
```

La función concluye que el modelo óptimo con los menores valores de AIC es el ARIMA(1,1,1), así como fue estimado mediante las pruebas de ACF, PACF y ADF, previamente. Después de ello, la función construye el modelo en base a los datos de entrenamiento proporcionados. Para desarrollar una predicción, se utiliza el método integrado de *predict*, cuyos parámetros toman la cantidad de períodos que se desea generar, es decir, la cantidad que fue apartada y el modelo jamás ha observado. Es así que se obtiene el primer pronóstico y puede ser contrastado con los datos real del *test\_df*, en la siguiente gráfica:

**Gráfica N°9.** Pronóstico de Temperatura ARIMA (1,1,1).



**Fuente:** Elaboración propia. (2024)



Se puede observar la cercanía con la cual el modelo de ARIMA pronosticó cómo serían los valores de la temperatura a lo largo de los meses en el periodo que jamás le fue proporcionado, representados por la línea azul. En contraste con los valores reales del *test\_df*, evaluando así su desempeño y precisión en la estimación de valores futuros.

Para la determinación de la precisión del modelo, se empleó la función de error medio cuadrático (MSE, por sus siglas en inglés), la cual permite “determinar la diferencia promedio al cuadrado entre el valor previsto y el real”, de acuerdo con Mulani, S. (2023), a partir del cual se calculará la raíz, resultando en el RMSE, siendo este un coeficiente popular para evaluar el desempeño de un modelo predictivo. Este funciona como un estándar para saber si el modelo es lo suficientemente preciso para ser confiable y proporciona una manera fácil de comparar su desempeño con la de otros modelos.

La fórmula del RMSE consiste en la sumatoria de la diferencia entre los valores pronosticados por el modelo y los valores reales para cada punto de la serie temporal, elevados al cuadrado; dividido entre la cantidad de elementos en la serie temporal y dentro de una raíz cuadrática, para así obtener un promedio del error del modelo. El cálculo es representado por la siguiente ecuación:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Pronóstico_i - Actual_i)^2}{N}}$$

Sin embargo, para el desarrollo del cálculo en *python*, se emplearon librerías para la aplicación de tales operaciones a la amplia cantidad de datos de entrenamiento y validación. Utilizando la función *subtract* de la librería *numpy* que toma como argumento los valores del pronóstico y los reales, como series temporales de una sola dimensión, los resta a partir del índice y retorna los resultados. A partir de allí se calcula el cuadrado de todos los valores usando la función *square()* de *numpy* y el promedio de los mismos con la función integrada de *mean()*, sobre ello, se calcula la raíz cuadrada mediante la función de *sqrt()* de la librería *math*. Resultando en:

$$RMSE = 0.44962$$

Es importante tomar en cuenta que el RMSE es una métrica dependiente de la escala a partir de la que se desarrolla, por lo que no existe un estándar o mínimo universal con el cual ser comparado

y a partir de sí solo no es capaz de indicar la precisión del modelo, pero si es útil para comparar el desempeño de dos modelos desarrollados sobre el mismo set de datos.

Es por ello, que el siguiente paso consiste en desarrollar el modelo de LSTM, sus siglas representan *Long Short-Term Memory*, caracterizados por ser “un poderoso enfoque de red neuronal recurrente que se ha utilizado para lograr los resultados más conocidos para muchos problemas con datos secuenciales”, de acuerdo con Kutzkov, K. (2023). Lo cual implica que es una técnica empleada para la solución de otros problemas de pronóstico o predicción y no necesariamente una especializada en la predicción de series temporales como ARMA. De igual manera, es importante recordar que las LSTM son un tipo de red neuronal, por lo tanto, esta no emplea métodos enteramente matemáticos, sino también lógicos mediante el desarrollo de células LSTM, también llamadas funciones, que cumplen con diferentes propósitos de aprendizajes como:

- Una representación compacta de la serie temporal hasta cierto punto.
- Cómo combinar una nueva entrada con la representación de la serie temporal.
- Qué olvidar de la serie.
- Qué retornar como pronóstico en el siguiente paso.

Por tales motivos, es que los modelos LSTM tienen la característica de requerir un alto grado de ajuste o modificación de parámetros para obtener resultados satisfactorios. Para ello, es importante tener en cuenta parámetros como:

- La cantidad de celdas LSTM que puedan representar la secuencia, donde, de acuerdo con Kutzkov, K. (2023), “Es poco probable que unas pocas células LSTM capturen la estructura de la secuencia, mientras que demasiadas células LSTM podrían provocar un sobreajuste.”; por lo que es un proceso delicado de balance y ajuste.
- Es típico que en un principio sea necesario convertir la secuencia de entrada (la serie temporal) en otra secuencia denominada  $h_t$ , la cual representa la serie procesada hasta cierto punto. Es por ello que “diseñar la arquitectura exacta puede requerir un ajuste cuidadoso y muchas pruebas.” (Kutzkov, K., 2023)

Para el desarrollo del modelo LSTM en *python*, fue necesario el preprocesamiento de los datos, de manera que estén ajustados a las necesidades del modelo. Primeramente, se realizó la normalización de los datos de todas las variables meteorológicas a emplear, para lo cual se empleó

la librería *MinMaxScalar* que forma parte del apartado *PreProcessing* en la librería *Sklearn*. En este paso, se incluyeron las variables exógenas desarrolladas previamente, dado que también pueden aportar beneficios en la identificación de patrones dentro del modelo LSTM. El resultado es un arreglo bidimensional donde las columnas están identificadas por un índice numérico y no con un identificador de texto, como en el caso de los sets de datos, debido a que el modelo requiere tal estructura de datos.

Acto seguido, se construyen las secuencias. Estos son conjuntos de un tamaño determinado de valores consecutivos de cada columna de variable, es decir, la primera secuencia de la columna temperatura sería del primer valor al décimo (en este caso), los cuales son utilizados para intentar pronosticar el onceavo valor y el error entre el valor estimado, y el real es tomado en consideración por el modelo a la hora de realizar el mismo proceso con la siguiente secuencia, la cual estaría comprendida desde el segundo valor hasta el onceavo, así sucesivamente a lo largo de todo el modelo.

Seguidamente, los valores normalizados de todas las variables meteorológicas y las secuencias previamente desarrolladas son separados en sets de entrenamiento y prueba, con una distribución de 80/20 respectivamente, la misma proporción usada en los sets de ARIMA. Esto resultó en 4 sets de datos diferentes, donde dos son denominados como *train\_x* y *test\_x*, refiriéndose a los valores de las secuencias previamente desarrolladas, y los otros dos fueron denominados *train\_y* y *test\_y*, los cuales estaban compuestos por los valores normalizados de la variable temperatura, exclusivamente.

En este punto, es donde se crea el modelo LSTM, tomando en cuenta que “es la profundidad de las redes neuronales lo que generalmente se atribuye al éxito del enfoque en una amplia gama de problemas de predicción desafiantes.”, de acuerdo con Brownlee, J. (2019). Por lo tanto, y considerando la complejidad de los datos con los que se trabaja, se diseñó un modelo conformado por múltiples capas internas denominado *deep LSTM*, cuya ventaja es que los valores de entrada alimentados a la red no solo pasan por varias capas de LSTM, sino que también se propagan a través del tiempo dentro de una misma celda de LSTM.

Asimismo, otro parámetro por definir es la cantidad de celdas (o unidades) existentes en cada capa individual, las cuales son las encargadas de recordar los valores de la serie temporal a lo largo

de los periodos establecidos para la construcción del set de datos de secuencia, lo hacen a partir del uso de una puerta de entrada, una de salida y otra de olvido, regulando el flujo de información que atraviesa la celda. En este sentido, la puerta de olvido decide qué información descartar, comparando los valores previos con los actuales en la celda, decidiendo si elimina o mantiene el elemento previo. Por otro lado, la puerta de entrada cumple la misma función, pero antes de la puerta de olvido, por lo que el valor previo de la puerta de olvido es el asignado por la puerta de entrada. Finalmente, la puerta de salida utiliza la misma mecánica para definir qué valor entregar como salida a la siguiente capa del modelo, la cual es proporcionada en forma de secuencia, así como la entrada inicial del modelo.

Sin embargo, realmente no existe una regla universal o técnica general para la definición de la cantidad de unidades o celdas por capas, pero no puede ser menor a la cantidad de variables meteorológicas y cualidades exógenas que determinan la cantidad de celdas que conforman la capa de entrada del modelo. En el caso presente, se tienen 28 cualidades, es a partir de tal número que se experimentará con distintas cantidades comparando su desempeño.

Sin embargo, el método más común está constituido por la experimentación de la anchura vs la profundidad, refiriéndose a la cantidad de celdas por capa como anchura y la cantidad de capas como profundidad, de acuerdo con OverLordGoldDragon (2019). Los beneficios de una mayor *anchura* son la extracción de más cualidades de los datos y la *profundidad* extrae mayor valor de cada una de tales cualidades. En base a lo considerado, se comenzó desarrollando un modelo compuesto por 3 capas internas LSTM intercaladas por capas de *dropout*, las cuales, de acuerdo con Brownlee, J. (2020), consisten en un método de regularización para prevenir que el modelo esté *overfitted*, es decir, que su entrenamiento esté demasiado ajustado a los valores de entrenamiento y que, en vez de pronosticar valores, solo repita los patrones observados.

Para la construcción del modelo, se empleó la librería *tensorflow*, de la cual se importó del apartado *keras*, las funciones *sequential()* para la iniciación del modelo, *LSTM()* para el establecimiento de las capas de memoria, *dropout()* para evitar el *overfitting* y la función *dense()* como salida. Finalmente, se compila el modelo indicando el uso de RMSE como métrica de medición de rendimiento, obteniendo la primera descripción de modelo:

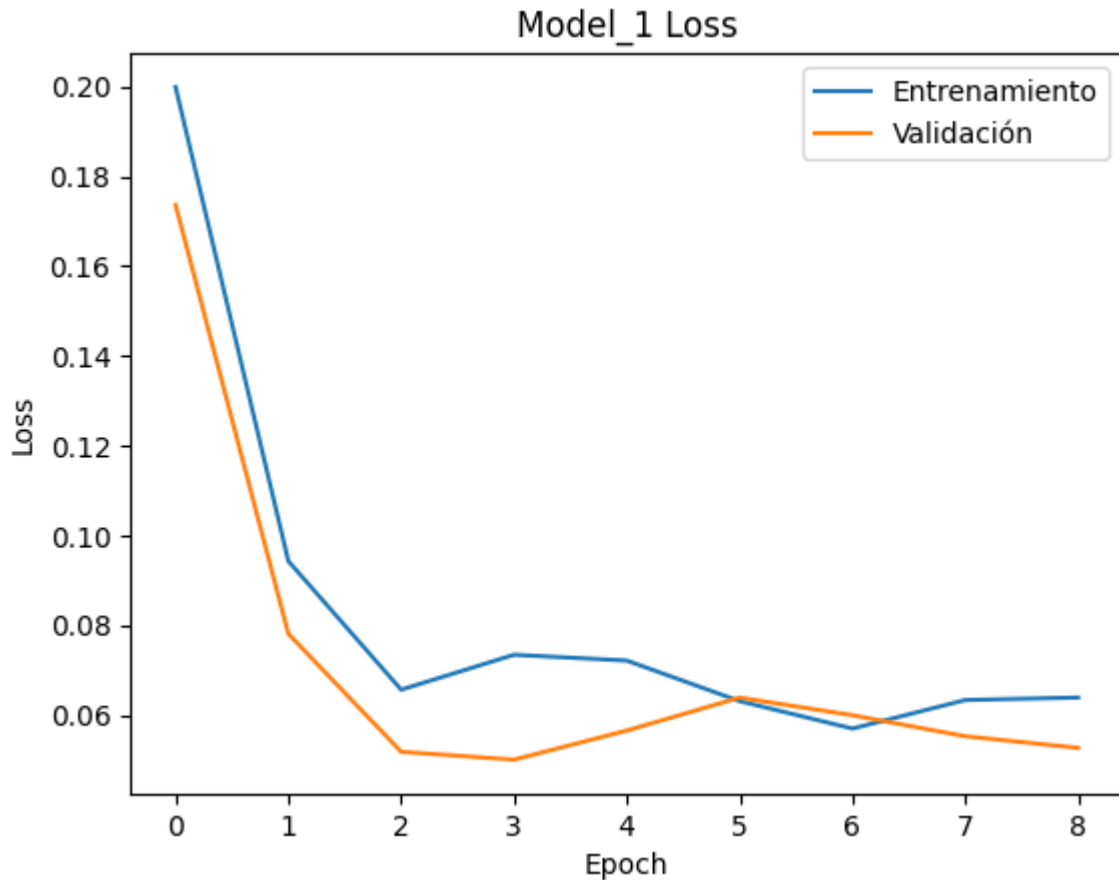
**Tabla N°14.** Descripción Modelo\_1 LSTM.

Model: "sequential_1"		
Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 10, 28)	6384
dropout_7 (Dropout)	(None, 10, 28)	0
lstm_1 (LSTM)	(None, 10, 14)	2408
dropout_7 (Dropout)	(None, 10, 14)	0
lstm_1 (LSTM)	(None, 10, 7)	616
dropout_7 (Dropout)	(None, 10, 7)	0
dense_4 (Dense)	(None, 1)	8

***Fuente:*** Elaboración propia. (2024)

De esta forma, se indica la estructura secuencial del modelo LSTM, cuya primera capa es de tipo LSTM con 28 celdas, como se ve en el último elemento en la columna *output shape*; la cual procesa la cantidad total de parámetros en el set de entrenamiento. Posterior a la construcción del modelo, se procede a su debido entrenamiento, donde es posible indicar la cantidad de *epoch* o iteraciones sobre las cuales se estudian los datos, se evalúa su propio desempeño a partir de métricas de error y vuelve a entrenar. Para cada iteración, la función del entrenamiento indicará un valor denominado *val\_loss*, el cual indica el valor MSE a lo largo de cada interacción y lo comparará con la anterior. En este caso, se implementó la condición de *early\_stopping* o paradas tempranas para evitar el sobre-entrenamiento del modelo, de manera que, en caso de que no vea mejoría durante 5 iteraciones, detendrá el entrenamiento con antelación y retornará el modelo, por lo que la cantidad de epoch no es más que un límite máximo de iteraciones. Los datos de cada iteración son almacenados en un set de datos para ser visualizados en una gráfica, que muestra cómo modelo mejoró en el entrenamiento y la validación a lo largo de todo el proceso, resultando en:

**Gráfica N°10.** Entrenamiento de Modelo\_1 LSTM.

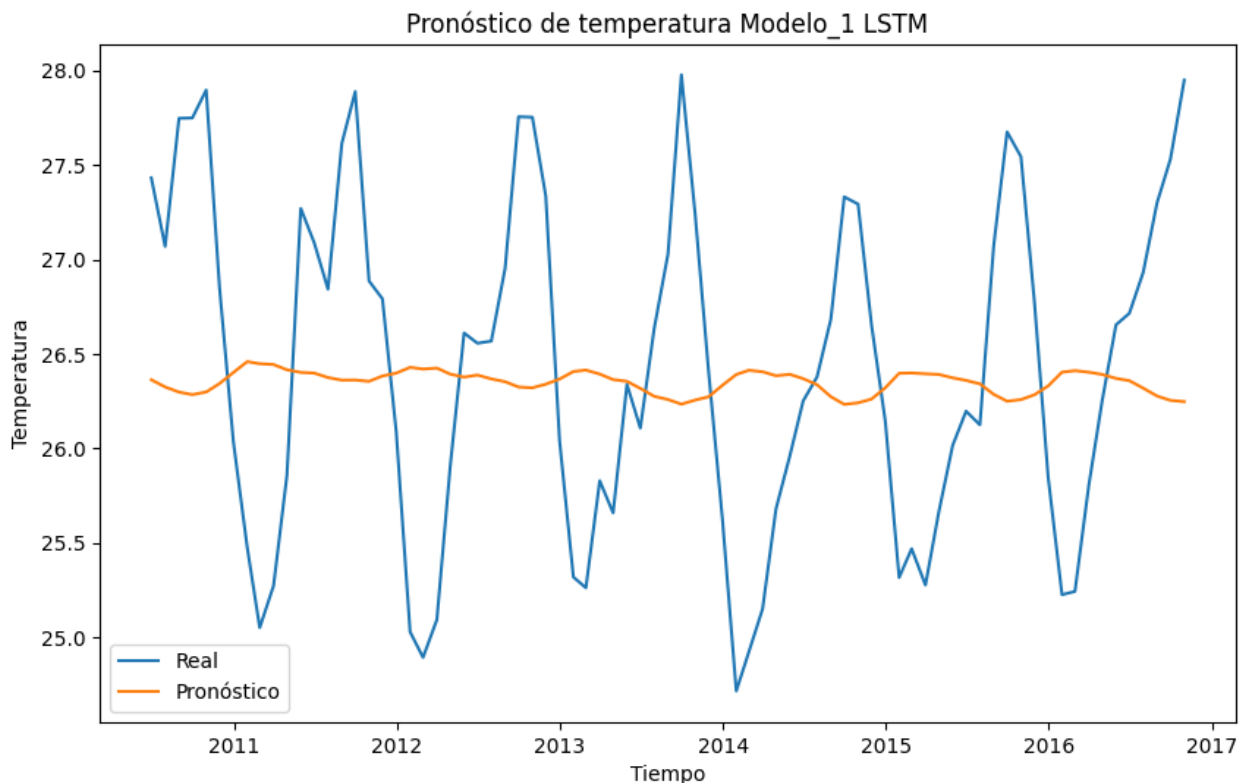


***Fuente:*** Elaboración propia. (2024)

Esta gráfica permite entender el proceso que atraviesa el modelo en el entendimiento de los patrones de los datos meteorológicos, donde la línea azul representa la diferencia entre los valores estimados durante el entrenamiento y los reales, y la naranja la diferencia entre los valores estimados en la validación y los reales. Es importante recordar que los valores de validación solo son usados para corroborar su desempeño en datos que no hayan sido usados para entrenamiento, es por ello que el hecho de que la línea azul está debajo de la naranja la mayoría de la gráfica indica que el modelo está siendo *underfitted*. Asimismo, la existencia de un decrecimiento súbito en ambos valores indica que el modelo no está identificando todos los patrones dentro de los datos y el proceso de aprendizaje no está sucediendo correctamente. Idealmente, se desea observar una curva suave, no una disminución drástica. Por otro lado, no se observó una convergencia entre ambas líneas para el final de la gráfica, ya que eso indicaría que el modelo ya aprendió todo lo que podría haber aprendido.

Por ello, la gráfica previa indica un desempeño mediocre en la primera prueba, de igual manera, se desarrollará un pronóstico y una evaluación de RMSE para confirmar. El pronóstico es ejecutado mediante el método integrado del modelo *predict()* que toma como atributo el set de *test\_x* empleado para la prueba del modelo, resultando en:

**Gráfica N°11.** Pronóstico de temperatura de Modelo\_1 LSTM.



**Fuente:** Elaboración propia. (2024)

En la gráfica previa se contrasta el pronóstico de la temperatura mensual desde el año 2011 hasta el 2017 realizado por el *modelo\_1* de prueba de LSTM, cuyo valor de RMSE fue de 0, , significativamente mayor en comparación al modelo ARIMA. En este caso, la línea azul representa el comportamiento real de la temperatura a lo largo del periodo de prueba, y la línea naranja ilustra el pronóstico realizado. Idealmente, lo que se desea obtener es la mayor similitud posible entre ambas líneas, lo cual evidenciaría un modelo preciso, pero se observa lo opuesto, no hay suficientes similitudes como para considerar al modelo funcional. Es por ello que el *modelo\_1* fracasó en predecir el comportamiento de la variable temperatura.

Sin embargo, es posible inferir, a partir de la gráfica, ciertos aspectos sobre los cuales mejorar el modelo y realizar una segunda prueba, debido a que a pesar de que los puntos altos y los puntos bajos anuales del pronóstico no son cercanos a la realidad, el hecho de que la línea esté ubicada justo en lo que sería el promedio de los valores reales y que las pequeñas ondulaciones que presentan están ubicadas en la misma frecuencia de los valores reales, implica que el modelo es capaz de entender la estacionalidad y los comportamientos relacionados con el mismo, pero no es capaz de recrearlos fielmente. Aunado a lo observado en la gráfica N°11 sobre el entrenamiento del *modelo\_1*, se logra entender que el modelo ha sido *underfitted*. Por tal razón, se pueden tomar dos acciones: aumentar la cantidad de capas o aumentar la cantidad de celdas por capas. Como se partió de una cantidad considerable de 3 capas LSTM, el siguiente cambio fue aumentar la cantidad de celdas por capa a cuatro veces la cantidad previa, tomando en cuenta el significativo error en el pronóstico. Finalmente, se aumentó la paciencia de la condición de *early\_stopping*, para evitar que el modelo detuviera su entrenamiento de manera prematura e intentar conseguir una convergencia entre las líneas de entrenamiento y validación. Al compilar el modelo con tales cambios, se obtiene como resultado:

**Tabla N°15.** Descripción Modelo\_2 LSTM.

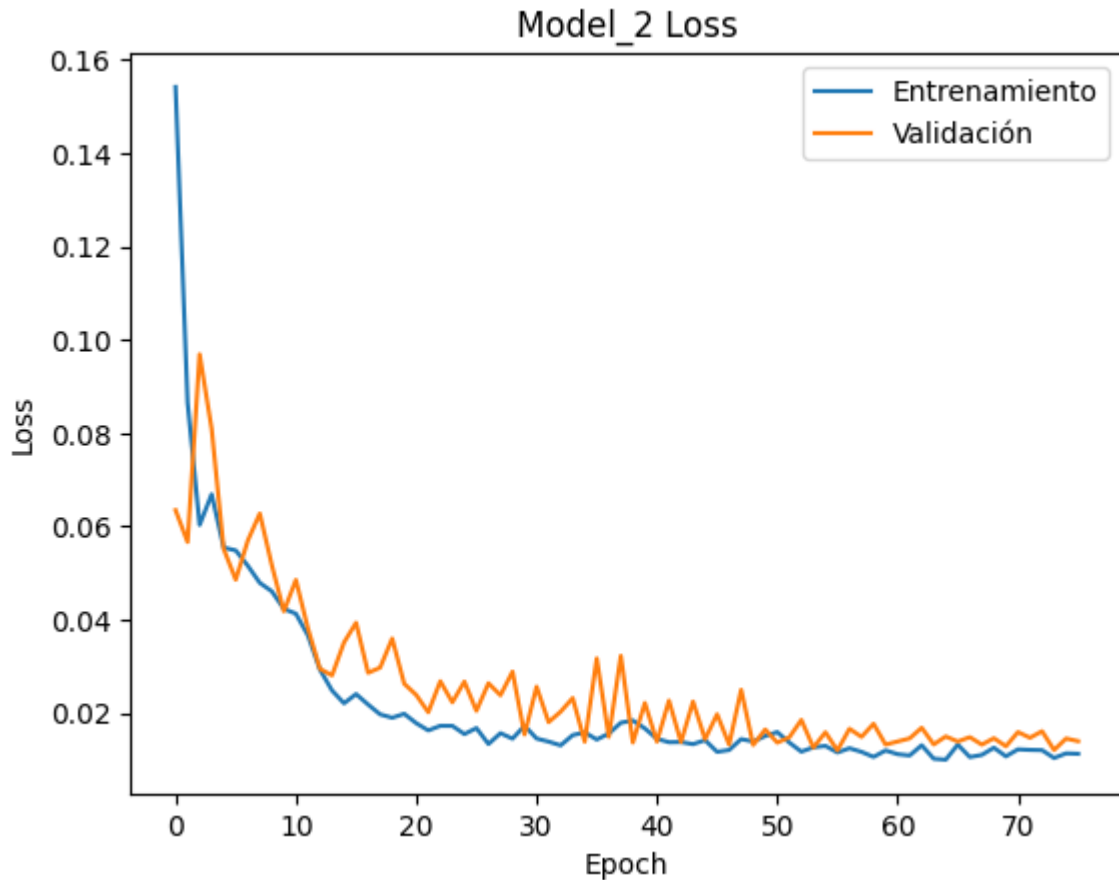
Model: "sequential_2"		
Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 10, 112)	6384
dropout_7 (Dropout)	(None, 10, 112)	0
lstm_1 (LSTM)	(None, 10, 56)	2408
dropout_7 (Dropout)	(None, 10, 56)	0
lstm_1 (LSTM)	(None, 10, 28)	616
dropout_7 (Dropout)	(None, 10, 28)	0
dense_4 (Dense)	(None, 1)	8

***Fuente:*** Elaboración propia. (2024)

Se mantiene la misma estructura de capas, ya que el problema no parece ser su capacidad de identificar los patrones en los datos, sino la de entender la magnitud de los mismos. Igual que antes, se ejecuta el entrenamiento del modelo y se almacenan los valores de desempeño para cada epoch y así observar la curva de aprendizaje del modelo; resultando en la siguiente gráfica:



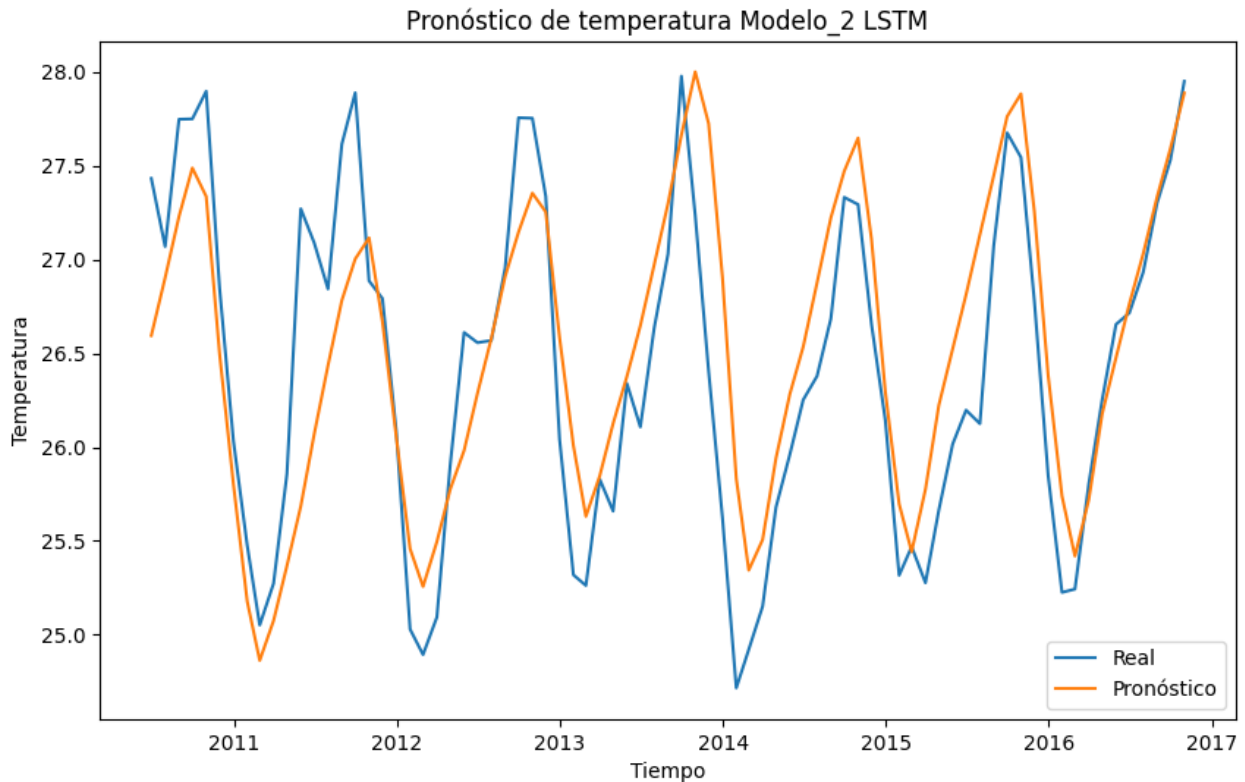
**Gráfica N°12.** Entrenamiento de Modelo\_2 LSTM.



***Fuente:*** Elaboración propia. (2024)

La principal diferencia entre el entrenamiento del modelo previo y el actual, es la cantidad de iteraciones que decidió realizar, gracias al cambio del parámetro de paciencia, lo cual logró que el modelo obtuviera una línea de aprendizaje más suave y se acercará más a la convergencia del entrenamiento y la validación, lo cual indicaría que el modelo no tiene más que aprender de los datos. Sin embargo, la línea de validación es inusualmente errática, pero eso puede atribuirse a la cantidad de ruido y valores atípicos encontrados en el estudio de los datos, los cuales presentan casos inesperados de estimación y pronóstico que el modelo falla. De igual manera, lo importante es que la casi convergencia de las líneas indica que el modelo ha superado los imprevistos. Acto seguido, se ejecuta una predicción del modelo para ser comparada con los datos reales de la variable temperatura, resultando en:

**Gráfica N°13.** Pronóstico de temperatura de Modelo\_2 LSTM.



***Fuente:*** Elaboración propia. (2024)

Sin necesidad de indagar mucho en la gráfica, los resultados son considerablemente más prometedores, ya que presentan una mayor similitud con lo que realmente sucedió. Asimismo, la métrica de RMSE apoya las primeras impresiones, indicando un valor de 0.126148, cerca de la mitad de la prueba anterior. De igual manera, en la gráfica se evidencia la considerable mejoría del modelo en seguir los altibajos de la temperatura durante el periodo de prueba, casi alcanzando los picos más bajos y altos. Por tal motivo, tomando el impacto del aumento de las celdas por capa, para intentar aminorar la brecha entre los valores estimados cerca de los límites de la gráfica, se aumentará en menor medida la cantidad de celdas y la paciencia del modelo para entrenar, así como la cantidad máxima de epochs, para lograr conseguir la convergencia entre el entrenamiento y la validación durante la competencia. La construcción del tercer modelo retorna el siguiente resumen de su estructura:

**Tabla N°16.** Descripción Modelo\_3 LSTM.

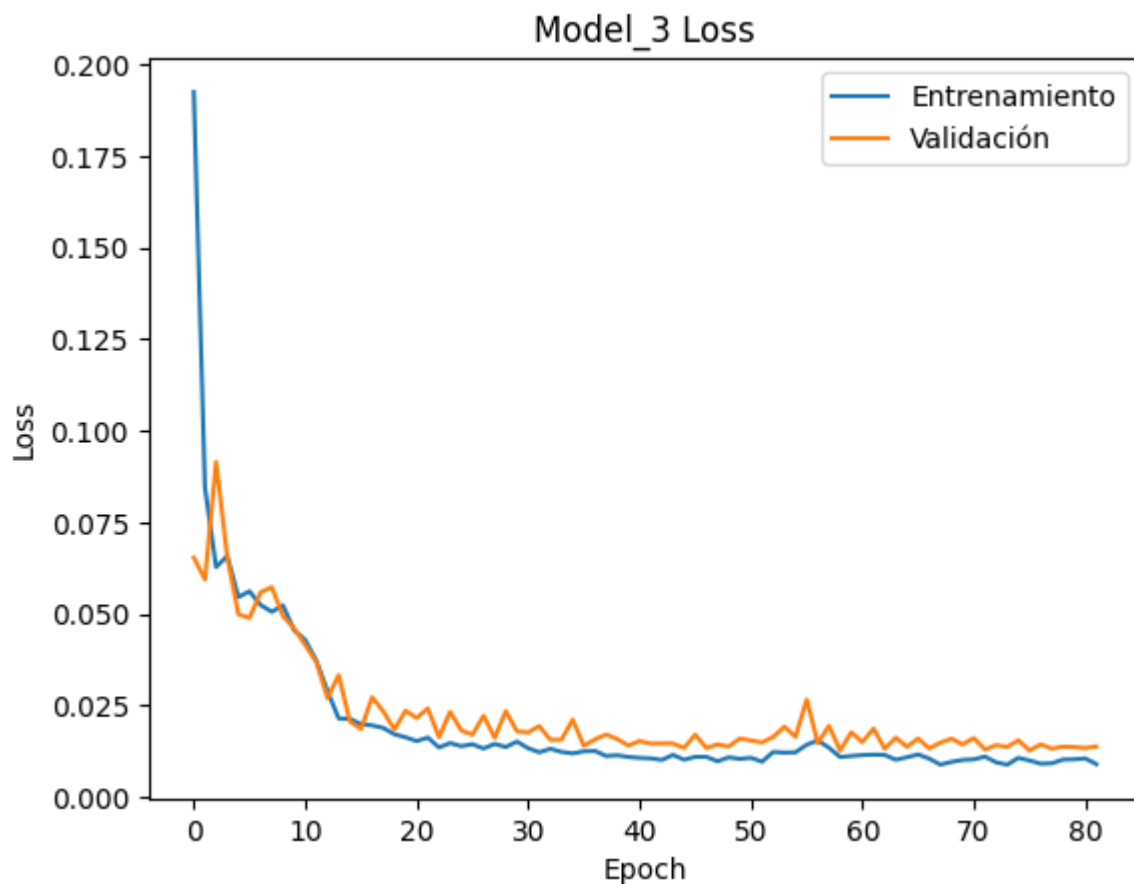
Model: "sequential_3"
-----------------------

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 10, 128)	6384
dropout_7 (Dropout)	(None, 10, 128)	0
lstm_1 (LSTM)	(None, 10, 64)	2408
dropout_7 (Dropout)	(None, 10, 64)	0
lstm_1 (LSTM)	(None, 10, 32)	616
dropout_7 (Dropout)	(None, 10, 32)	0
dense_4 (Dense)	(None, 1)	8

***Fuente:*** Elaboración propia. (2024)

Asimismo, se aumentó la paciencia del modelo a 23 epoch y el límite máximo de los mismos a 130. Posterior a ello, se ejecutó el entrenamiento cuyo desempeño se ve representado en la siguiente gráfica:

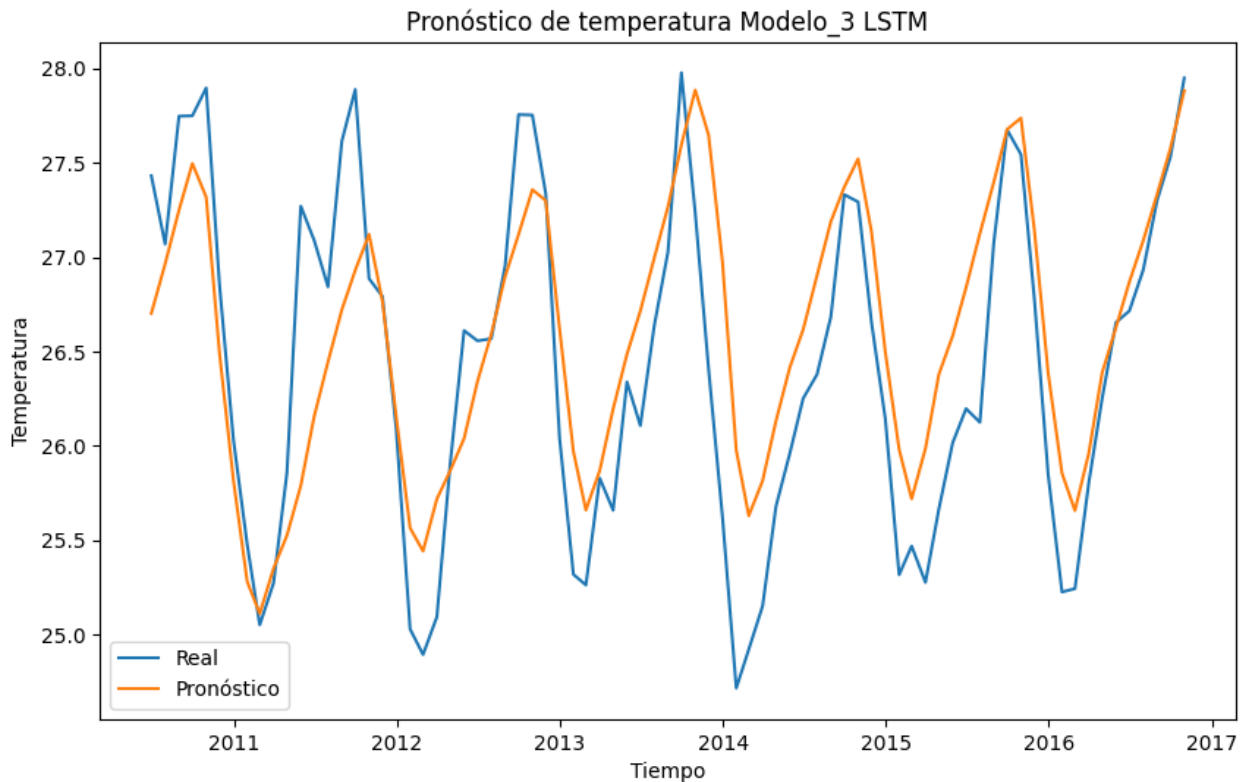
**Gráfica N°14.** Entrenamiento de Modelo\_3 LSTM.



***Fuente: Elaboración propia. (2024)***

En comparación al entrenamiento del modelo previo, la línea de validación evidencia menos fluctuaciones en la prueba de los conocimientos desarrollados por el modelo en su entrenamiento, así como un aumento en la cantidad de epochs que consideró necesario para alcanzar una línea mucho más plana al final de su entrenamiento y validación. No se logró la total convergencia, pero tomando en cuenta la cantidad de valores atípicos, es posible que se deba a la variabilidad entre los sets de distintos tamaños y la diferencia temporal entre ellos. Al ejecutar un pronóstico, resulta la siguiente gráfica:

**Gráfica N°15.** Pronóstico de temperatura de Modelo\_3 LSTM.



***Fuente: Elaboración propia. (2024)***

Primeramente, el valor de RMSE correspondiente al pronóstico ilustrado en la gráfica previa fue de 0.1323, un valor mayor al del *modelo\_2*, indicando una ligera disminución en precisión, la cual puede ser observada en la poca cantidad de detalles que el pronóstico entrega sobre los valores reales de la temperatura. Este comportamiento puede ser atribuido a un posible sobre-entrenamiento del modelo, que disminuye su capacidad para generalizar sobre los valores de la

prueba, empeorando su capacidad de replicar con detalles lo que realmente sucedió. Por ello, se retractarán los cambios al límite de epoch y disminuirá la paciencia del modelo a menos del *modelo\_2*, disminuyendo la posibilidad de incurrir en el mismo error.

De igual forma, dado que la falta de precisión en detalles puede estar atribuida a una sub-óptima observación de las cualidades exógenas de los datos, se añadirá una capa extra de LSTM con la mitad de la cantidad de celdas de la previa, con la intención de mejorar la capacidad de observación y abstracción del modelo, sin incluir la capa de *dropout()* para disminuir factores que puedan afectar el resultado. La implementación de tales cambios resulta en la descripción del siguiente modelo:

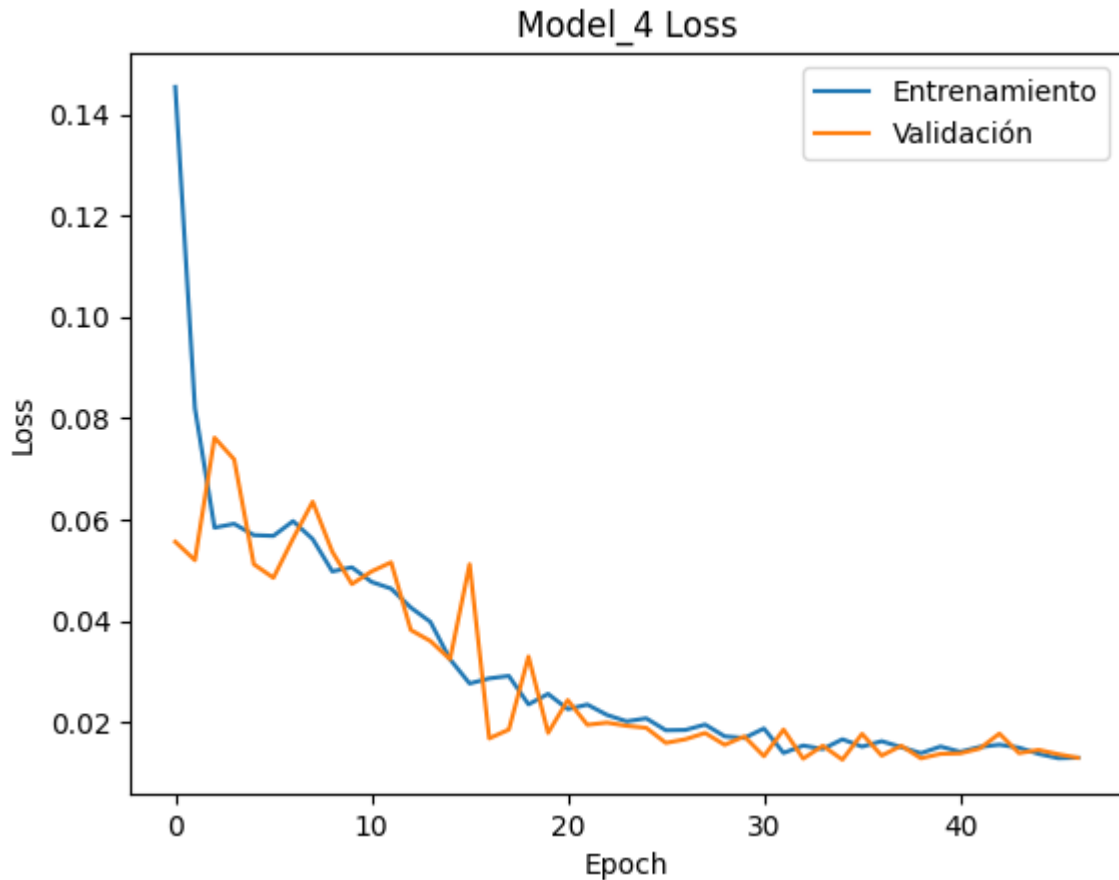
**Tabla N°17.** Descripción Modelo\_4 LSTM.

Model: "sequential_4"		
Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 10, 128)	80384
dropout_7 (Dropout)	(None, 10, 128)	0
lstm_1 (LSTM)	(None, 10, 64)	49408
dropout_7 (Dropout)	(None, 10, 64)	0
lstm_1 (LSTM)	(None, 10, 32)	12416
dropout_7 (Dropout)	(None, 10, 32)	0
lstm_1 (LSTM)	(None, 10, 16)	3136
dense_4 (Dense)	(None, 1)	8

***Fuente:*** Elaboración propia. (2024)

La previa tabla evidencia la estructura añadida sobre el *modelo\_3*, con el objetivo de mejorar el nivel de detalle con el que el modelo es capaz de recrear el comportamiento de la variable. Al momento de ejecutar el entrenamiento, se obtiene la siguiente gráfica de desempeño:

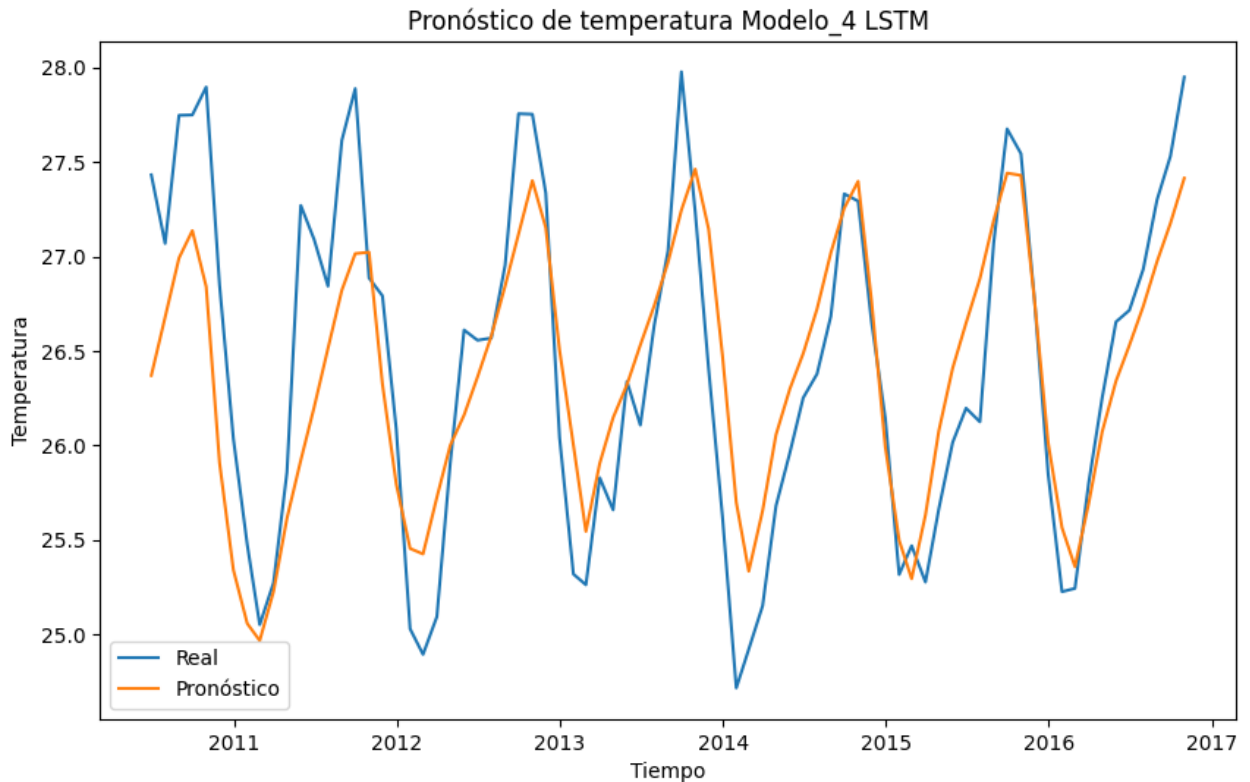
**Gráfica N°16.** Entrenamiento de Modelo\_4 LSTM.



*Fuente: Elaboración propia. (2024)*

El principal aspecto observado en el desempeño del entrenamiento del modelo es la convergencia de las líneas de entrenamiento y validación, indicando que el modelo aprendió lo máximo posible de ese conjunto de datos, lo cual se logró fijando la paciencia del modelo en 12 epochs. Sin embargo, se ejecutaron menos iteraciones de entrenamiento, lo cual podría indicar, nuevamente, underfitting del modelo. Indiferentemente, el correspondiente RMSE es el menor hasta el momento de las pruebas de LSTM, resultando en 0.1191282; por lo que el modelo\_4 indica ser el más preciso de los LSTM. Acto seguido, se observa la estimación del modelo sobre los valores de prueba de temperatura, para observar su exactitud:

**Gráfica N°17.** Pronóstico de temperatura de Modelo\_4 LSTM.



***Fuente:*** Elaboración propia. (2024)

La gráfica previa ilustra que el modelo aún presenta complicaciones con la captación de mayor detalle del comportamiento de la variable a lo largo del periodo de prueba, lo cual, nuevamente, apunta que el modelo sufre de *underfitting*, sin embargo, la gráfica de entrenamiento permite entender que el modelo aprendió todo lo posible de los datos y el modelo LSTM no es capaz de abstraer más información a partir de la estructura mensual de los datos. Por esta razón, se procederá a comparar a los distintos modelos LSTM para seleccionar el de mejor rendimiento:

**Tabla N°18.** Comparación de modelos LSTM.

LSTM	RMSE
Modelo_1	0.22875
Modelo_2	0.12615
Modelo_3	0.13230
Modelo_4	0.11913

***Fuente:*** Elaboración propia. (2024)

Para desarrollar conclusiones a partir de la gráfica previa, es importante recordar que la métrica de rendimiento RMSE consiste en la media de las diferencias cuadráticas entre los valores reales y los valores estimados por los modelos, de manera que un mejor resultado es equivalente a un menor RMSE. Por lo que el modelo LSTM con mejor rendimiento es el último desarrollado, el *modelo\_4*, cuyo rendimiento será comparado con el modelo ARIMA para definir cuál presenta un mayor rendimiento y precisión en cuanto a pronóstico de temperatura en las pruebas realizadas; resultando en:

**Tabla N°19.** Comparación de rendimiento de modelos LSTM y ARIMA.

	LSTM	ARIMA
RMSE	0.11913	0.44963

***Fuente:*** Elaboración propia. (2024)

Nuevamente, el criterio de comparación consiste en que: entre menor sea el valor de RMSE, mayor será la precisión y rendimiento del modelo al ser desarrollado e implementado. Por lo que es fácil concluir, a partir de una comparación directa, que el mejor modelo para aplicar en el contexto de la investigación son las redes neuronales recurrentes de tipo LSTM.

Es a partir de esta conclusión, que se decide desarrollar un modelo de este tipo, empleando las variables meteorológicas seleccionadas dentro del periodo definido previamente. Para ello, se emplea la ventaja principal de los modelos LSTM que, en contraste con ARIMA, no tiene limitaciones en cuanto a cantidad de saltos por frecuencia estacional. De manera que, no se emplearán los datos de dimensionalidad temporal mensual (promedios mensuales), sino los datos estructurados por hora, los cuales poseen una mayor resolución y se presentan como de mayor utilidad al momento de desarrollar el proyecto factible presentado. En efecto, será posible pronosticar los valores meteorológicos por cada hora del día, utilizando la misma estructura de red neuronal recursiva del *modelo\_4* LSTM, con las mismas características exógenas, solo que en formato horario.

#### **4.4 Evaluación del rendimiento del modelo de machine learning en términos de precisión y capacidad de predicción meteorológica en el Estado Nueva Esparta.**

Con el objetivo de identificar una métrica que esté en la capacidad de representar y calificar el desempeño del modelo, se realizó una lista de verificación con la intención de evaluar las fortalezas



y debilidades principales de las técnicas de evaluación de rendimiento de modelos de machine learning más comúnmente utilizados, de manera que, durante su proceso de desarrollo, se pueda cuantificar la capacidad predictiva de las distintas versiones de cada prueba, para una fácil comparación entre los cambios implementados.

Sin embargo, antes de evaluar las cualidades de cada métrica, es crucial definir las necesidades que engloban al modelo que se desea desarrollar, para contrastarlas con las ventajas y desventajas involucradas en el empleo de cualquiera de las posibles métricas y así entender cuál se adapta mejor al contexto. De tal manera, se estima que el modelo presenta las siguientes cualidades:

- Valores atípicos: Debido al comportamiento ocasionalmente errático del tiempo, las variables meteorológicas presentan mediciones fluctuantes que resultan en una gran cantidad de valores atípicos en la serie temporal.
- Escala de medidas: La escala de los valores de variables, como la humedad relativa y la temperatura, presentan una oportunidad para que errores graves desvíen las métricas de rendimiento excesivamente y no sean representativas del verdadero error.
- Estacionalidad: La frecuencia presente en los valores meteorológicos aumenta la posibilidad de la presencia, tanto de sobreestimaciones como de subestimaciones.

A partir de tales cualidades, es posible evaluar cuál métrica se adapta mejor a las necesidades presentadas en el modelo, sus datos y su deseado resultado. En este contexto, se evaluaron las cualidades de 6 distintas métricas, obteniendo los siguientes datos:

**Tabla N°20.** Comparación de métricas de rendimiento.

Métrica de rendimiento	Función	Ventajas	Desventajas
Error absoluto promedio (MAE)	$\frac{\sum_{i=1}^N  Pronóstico_i - Actual_i }{N}$	<ul style="list-style-type: none"> <li>- Fácil de entender.</li> <li>- Mantiene la misma unidad del valor original.</li> <li>- Resistente a valores atípicos por ser absoluto.</li> </ul>	<ul style="list-style-type: none"> <li>- No penaliza errores graves más significativamente.</li> <li>- Sobreestimación y subestimación son tratados igualmente.</li> </ul>

Error absoluto promedio cuadrático (RMSE)	$\sqrt{\frac{\sum_{i=1}^N (\text{Pronóstico}_i - \text{Actual}_i)^2}{N}}$	<ul style="list-style-type: none"> <li>- Penaliza errores graves más significativamente.</li> <li>- Mantiene la misma unidad del valor original.</li> </ul>	- Sensible a valores atípicos por usar potencias.
Error absoluto promedio porcentual (MAPE)	$\frac{\sum_{i=1}^N  \text{Pronóstico}_i - \text{Actual}_i }{\text{Pronóstico}_i}$	<ul style="list-style-type: none"> <li>- Presente la magnitud del error como porcentaje del valor real.</li> </ul>	<ul style="list-style-type: none"> <li>- Problemas con valores cercanos a 0 por división entre cero.</li> <li>- Sobrestimación y subestimación son tratados igualmente.</li> </ul>
Error absoluto promedio porcentual simétrico (SMAPE)	$\frac{\sum_{i=1}^N  \text{Pronóstico}_i - \text{Actual}_i }{\frac{ \text{Pronóstico}_i  +  \text{Actual}_i }{2}}$	<ul style="list-style-type: none"> <li>- Considera la sobrestimación y subestimación individualmente.</li> </ul>	- Menos intuitivo y también tiene inconvenientes con valores cercanos a cero.
Error porcentual absoluto medio (MDAPE)	$\text{media} \left( \frac{ \text{Pronóstico}_i - \text{Actual}_i }{\text{Pronóstico}_i} \right)$	<ul style="list-style-type: none"> <li>- Menos afectado por pocos grandes errores.</li> <li>- La media es una métrica de tendencia central útil para entender el error típico de pronóstico.</li> </ul>	- No provee mucha información sobre la distribución como RMSE
Error absoluto relativo de la media geométrica (GMRAE)	$\left( \prod \frac{ \text{Pronóstico}_i - \text{Actual}_i }{\text{Pronóstico}_i} \right)^{\frac{1}{n}}$	<ul style="list-style-type: none"> <li>- Considera la sobrestimación y subestimación individualmente.</li> <li>- Menos sensible a errores extremos y valores atípicos.</li> <li>- Expresado en porcentaje para fácil</li> </ul>	<ul style="list-style-type: none"> <li>- Puede no revelar cambios sutiles en diferentes desempeños de pronóstico.</li> <li>- Presenta problema con los valores reales incluyen</li> </ul>

		interpretación. - Provee una métrica única y resumida para estimar y comparar rendimiento.	cero. - Menos usada. - Mayor complejidad.
--	--	---	---

***Fuente:*** Elaboración propia. (2024)

Con este cuadro, es posible entender rápidamente los beneficios y contratiempos involucrados en el empleo de las diferentes métricas que pueden ser utilizadas para estimar el desempeño del modelo. Donde, en un principio se describen las cualidades de la técnica MAE, la cual si bien penaliza más gravemente a los errores graves, es altamente sensible a valores atípicos, un factor presente en los datos de entrenamiento pero su importancia se diluye al considerar que todos los modelos pronostican el mismo lapso de tiempo por lo que el error fue constante a lo largo de todas las pruebas.

Por otro lado, la técnica de MAPE, debido a que involucran valores absolutos resolvería el problema presentado por los valores atípicos con el inconveniente de los ceros en los valores reales podría causar errores de división entre cero al momento de estimar la media porcentual de los errores, de la misma manera en la que SMAPE y MDAPE. Por ello, se emplea la métrica MAPE para evaluación de los pronósticos resultantes del modelo desarrollado, que se pueden observar en la siguiente tabla, donde se incluyen dos ejemplos de las múltiples variables que el modelo puede pronosticar:

**Tabla N°21.** Rendimiento de modelo

Variable pronosticada	MAPE	Porcentaje de exactitud
Temperatura	0.03902%	99,9609%
Humedad relativa	0.02680%	99,9732%

***Fuente:*** Elaboración propia. (2024)

Es así, que se puede intuir fácilmente el grado de precisión con la que el modelo es capaz de pronosticar, calculado en relación a los valores reales del mismo lapso de prueba mediante el empleo de la técnica MAPE descrita previamente. Es importante recalcar que no existe un estándar

o un valor mínimo establecido sobre el cual un modelo puede ser designado como preciso o no, sino que depende mucho del grado de precisión que la aplicación específica puede requerir. Sin embargo, entre menor sea la MAPE, que representa el error absoluto porcentual, mejor es el rendimiento del modelo, y tomando en cuenta que es una métrica porcentual, se intuye que su diferencia ( $MAPE - 100\%$ ) representaría la exactitud con la que el modelo estima los valores, representado en la tercera columna de la tabla previa. Por lo tanto, se concluye que el modelo alcanzó la máxima exactitud posible y es viable para una aplicación práctica y operativa en la toma de decisiones de una organización o entidad.

## PARTE V

### CONCLUSIONES Y RECOMENDACIONES

#### 5.1 Conclusiones

A lo largo del desarrollo del proyecto de investigación, se indagó sobre los factores que determinan el comportamiento del tiempo y el clima en el Estado Nueva Esparta, Venezuela, mediante el empleo de técnicas de análisis de grandes cantidades de datos meteorológicos históricos de la región, para la extracción de correlación e interdependencias, patrones recurrentes, tendencias y cualidades de estacionalidad. De esta forma, se logró desarrollar, construir e implementar un modelo adaptado a las cualidades y necesidades de la problemática, capaz de pronosticar eficazmente los valores futuros de las características principales del tiempo en la región, empleando exclusivamente recursos que pueden ser adquiridos y localizados localmente y con el potencial de ser ajustados para pronosticar el clima de regiones específicas con el correcto set de datos históricos. En cuyo contexto, las conclusiones obtenidas de esta investigación ofrecen una perspectiva global en cuanto a la naturaleza del clima y el tiempo, las complicaciones y errores inherentes con su medición, los imprevistos relacionados con el pronóstico de una serie temporal y las técnicas necesarias para alcanzar un modelo de alta exactitud.

Al momento de leer las siguientes conclusiones, será fácil comprender las contribuciones claves de esta investigación en los campos de pronóstico de series temporales con *Machine Learning* y análisis de datos, comenzando con lo relacionado con la selección de variables meteorológicas:

- Relaciones de dependencia: Es crucial definir las cualidades del tiempo que se desea pronosticar con el modelo, de manera que a partir de ellas se desarrolle un criterio de selección, observando sus relaciones con otras variables, con el fin de incluir en el entrenamiento únicamente aquellos valores cruciales para la obtención de un modelo preciso.
- Similitud entre variables: Evitar otorgar demasiada importancia a la correlación entre variables de alta similitud, como “temperatura” y “temperatura aparente”, debido a que su interrelación no es indicativa de un patrón meteorológico sino de la proporción entre una y la otra, en contraste con la relación entre las variables “humedad” y “temperatura” que puede ser indicativa de fenómenos meteorológicos.

Asimismo, fue posible desarrollar conclusiones la determinación del periodo de los datos:

- Desviaciones artificiales: Considerando que el error es un factor inherente en la medición, es crucial verificar que este permanece constante a lo largo de toda la serie temporal en todas las variables empleadas, para evitar incluir tendencias de origen artificial.

Seguidamente, en el preprocesamiento y la determinación de la técnica de machine learning, se pueden recalcar los siguientes puntos:

- Valores atípicos: La naturaleza compleja y, parcialmente impredecible del tiempo, es propensa a producir valores atípicos con frecuencia. Por lo que es necesario saber identificar cuáles son de origen natural y cuáles son artificiales, minimizando el efecto que el analista pueda tener en la integridad y fidelidad de los valores meteorológicos.
- Continuidad de serie temporal: El impacto de la interrupción de una serie temporal sobre valores meteorológicos es mayor que la estimación a partir de valores adyacentes al segmento vacío, debido a la importancia que la frecuencia estacional tiene en la predicción de valores meteorológicos, ya que el clima está estrechamente relacionada con la temporada del año.
- Preparativos de ARIMA: La implementación de un modelo ARIMA para el pronóstico de una serie de tiempo es relativamente simple, en contraste con las pruebas matemáticas que lo anteceden y la interpretación de sus resultados; por lo que el apropiado y completo estudio de los datos y sus cualidades es crucial para la correcta implementación de modelos ARIMA.
- Cualidades de soporte: La construcción de variables exógenas a partir del método de retraso y la descomposición de estructuras complejas como la fecha, son un paso importante para la explotación de todo patrón posible dentro de los datos. Su determinación depende de la naturaleza del fenómeno a pronosticar, como el tiempo depende del momento y la estación, las variables estaban relacionadas a lo mismo.
- Naturaleza de modelos LSTM: La implementación de un modelo de red neuronal recursiva de tipo LSTM conlleva preprocesamiento de datos más complejo, puesto que el desarrollo de secuencias, transformación de dimensiones y normalización de datos constituyen puntos donde se pueden cometer errores complicados de diagnosticar e identificar. Por esta razón, LSTM es un método más frágil que requiere una atención más minuciosa.

- Abstracción de datos: La capacidad de comprender las gráficas de entrenamiento para los modelos LSTM es de vital importancia en el diagnóstico temprano de problemas, la identificación de zonas de mejora y la visualización del impacto de campos implementados, presentándose como una de las principales herramientas en el ajuste de modelos LSTM.

Finalmente, se desarrolló la siguiente conclusión al momento de evaluar el rendimiento del modelo;

- Medición de error: Las métricas para la medición del error entre el pronóstico y los valores reales de la variable no son universalmente comparables entre sí ni entre modelos, ya que dependen de la escala en la que la métrica existe y la cantidad de valores sobre la cual se estima. En efecto, modelos de distinto tamaño o pronóstico no son fácilmente comparables entre sí.

Es importante recalcar que el logro de los objetivos específicos a partir de los resultados obtenidos en la investigación sirvió como un portal entre el análisis teórico y la implementación práctica. En tal sentido, las previas conclusiones fundamentan la investigación y son ejemplo de los logros alcanzados en el estudio, resaltando:

- Pronóstico del tiempo: Se logró desarrollar un modelo de *Machine Learning* con la capacidad de pronosticar los valores de temperatura de las condiciones del tiempo en el futuro próximo del Estado Nueva Esparta, empleando únicamente datos históricos meteorológicos, los cuales pueden ser recolectados y almacenados localmente. De este modo, se proporciona un antecedente para el desarrollo de herramientas inteligentes para el pronóstico de factores nacionales, únicamente empleando recursos de la región.
- Comparación de modelos: El estudio, más allá de anteceder a la construcción del modelo, proporciona un valor añadido mediante la comparación de los distintos modelos LSTM y ARIMA, contrastando sus limitaciones, ventajas y desventajas, proporcionando fundamentos para la selección de un modelo en base a las cualidades de los datos referentes al problema por resolver y los pasos requeridos para la definición de parámetros necesarios en su correspondiente construcción.

Primordialmente, la intención alrededor de estas conclusiones es invitar al lector a considerar las implicaciones de embarcarse en el desarrollo de un proyecto con similares requerimientos,

instigando e incitando a la solución de problemas complejos y sin aparente solución, así como el mejoramiento de sistemas preexistentes mediante la implementación de *Machine Learning*.

## 5.2 Recomendaciones

En este apartado, se especifican las medidas por implementar originadas directamente de los descubrimientos producto del proceso investigativo, cumpliendo la función de una guía metodológica y estratégica destinadas a orientar a nuevos desarrolladores de modelos de *Machine Learning* en el proceso de construcción de soluciones adaptadas a las necesidades de las problemáticas en cuestión.

El uso de estas directrices darían origen a un proceso de construcción de modelo más eficaz, directo y simple, evitando el trabajo redundante o innecesario causado por inconvenientes inesperados. Estas medidas están destinadas a ser aplicadas, comprensibles, indicativas y eficaces.

- Evaluar la disponibilidad de datos meteorológicos: Ya sean datos históricos o en tiempo real, es necesario saber qué datos son accesibles, debido a que es a partir de tal oferta que se debe desarrollar un modelo capaz de explotar las relaciones entre las variables con mayor interdependencia. Para evitar construir un modelo que no se ajuste a la información disponible en cuanto a la región de aplicación.
- Tomar en consideración el origen de los datos históricos: Para así saber si no es automatizado o de integridad certificada, es necesario desarrollar pruebas que corroboren su completitud, mediante la verificación de valores vacíos y saltos en la serie de tiempo.
- Asegurarse de determinar apropiadamente la frecuencia estacional: Esta es necesaria para la obtención de pronósticos correctos con un modelo ARIMA a partir de series temporales, de la misma manera en la que es una limitación física. Por lo que es crucial identificar la frecuencia de pasos sobre la cual los datos se repiten a sí mismos.
- Considerar que no existe una metodología universal: Sino que, para el desarrollo de modelos LSTM, se utilizan una serie de reglas empíricas para el punto de partida y es la prueba y el error, fundamentados en la teoría de cómo funciona el modelo, que suele proporcionar resultados satisfactorios. Es por ello que el registro de cambios y su impacto es importante para la mejora como desarrollador.



Al implementar estas recomendaciones, las partes involucradas en el desarrollo de modelos de ML, no solo estarán en la capacidad de cumplir con los objetivos planteados, sino que también podrán alcanzarlos de manera eficiente, evaluando los requerimientos y limitaciones con antelación para tomar acciones decisivas y fundamentadas en el proceso innovativo y creativo que conlleva la tarea embarcada.

## **PARTE VI**

### **PROPUESTA**

Esta sección presenta la parte principal del estudio, identificada como la propuesta, la cual está estructurada para justificar su importancia y delimitar su potencial impacto en su área de implementación. Describiendo detenidamente la correspondiente factibilidad técnica, operativa y económica de la misma iniciativa, englobada en objetivos generales y específicos que especifican su correcta implementación y uso. Finalmente, se presenta la estructura del código desarrollado para construir el modelo LSTM.

#### **6.1 Importancia de la propuesta**

El desarrollo de un modelo de machine learning para el pronóstico del clima local es de vital importancia para el mantenimiento de las prevenciones necesarias para la ejecución de actividades dependientes o vinculadas al estado del tiempo. Por lo que la implementación de esta herramienta permitiría el aprovechamiento de datos abstraídos localmente, que proporcionan resultados de alta precisión para la región donde pronósticos exactos son necesarios, presentándose como una solución interna integral para organizaciones cuyas operaciones se vean en la necesidad de planificar alrededor de las condiciones del tiempo.

Asimismo, el conocimiento sobre los futuros fenómenos meteorológicos potenciales en el futuro, podrían jugar un papel importante en la toma de decisiones informadas dentro del sector público, principalmente para la administración de recursos de la nación, y en la planificación de mantenimiento de la infraestructura del país, eventos públicos, operativos de seguridad y prevención contra catástrofes meteorológicas.

Igualmente, proporciona una herramienta de bajos requisitos, armada a partir de recursos de uso gratuito en hardware común, para entidades científicas o educativas, por lo que se busca producir nuevos conocimientos en estudios dependientes o derivados de valores meteorológicos específicos y altamente localizados en el futuro, así como también en estudios sobre las implicaciones meteorológicas de cambios culturales, sociales y económicos.

#### **6.2 Viabilidad de la propuesta**

Seguidamente, se presenta un análisis evaluando todos aquellos aspectos cruciales de la propuesta relacionados con su grado de factibilidad que son de suma importancia para determinar si es aplicable y con qué eficacia dentro del entorno real, considerando los aspectos técnicos, operativos y económicos que son claves para respaldar la viabilidad de la propuesta.

### 6.2.1 Factibilidad técnica

A partir de todos aquellos elementos que fueron necesarios para el diseño e implementación del modelo LSTM, se ha definido la factibilidad técnica, abordando las consideraciones necesarias para que se pueda implementar de manera sólida y efectiva el proyecto en cuestión. En otras palabras, se han determinado los recursos tecnológicos necesarios para llevar a cabo el proyecto y si estos cumplen adecuadamente.

**Tabla N°22.** Requerimientos técnicos.

Requerimiento	Descripción
Equipo de desarrollo	<p>El computador cuenta con las siguientes características:</p> <ul style="list-style-type: none"> <li>- Windows 10 Pro</li> <li>- Intel Core i5-4570, 4 núcleos, 4 hilos, 3.2Ghz.</li> <li>- 16 Gb RAM DDR3</li> <li>- 526 Gb SSD</li> <li>- Radeon R7 350x, 4Gb 128bit GDDR5, 1000 MHz</li> </ul>
Servidor local de desarrollo y entrenamiento	<p>Es servidor cuenta con las siguientes características:</p> <ul style="list-style-type: none"> <li>- Ubuntu Linux 22.04</li> <li>- Intel Core i7-14700K, 20 núcleos, 28 hilos, 3.4Ghz.</li> <li>- 64GB de RAM DDR5</li> <li>- 1TB SSD M.2</li> <li>- NVIDIA GeForce RTX 2080 Ti, 11Gb 352bit GDDR6.1, 1665 MHz.</li> </ul> <p>Considerados como los requisitos mínimos para el entrenamiento del modelo.</p>
Conexión a internet	De fibra óptica con una velocidad mínima de 10mb y una ideal de 50mb.

Lenguaje de programación	Python 3.11.8
IDE	Visual Studio Code 1.78.2
Repositorio	GitHub
API Climatológica	Open-Meteo Historical Weather API

***Fuente:*** Elaboración propia. (2024)

Asimismo, es importante recalcar que la API empleada posee una licencia de GNU Affero General Public License v3.0, la cual es “Es una licencia copyleft gratuita para software y otros tipos de trabajos, diseñados específicamente para garantizar cooperación con la comunidad en el caso del software de servidor de red” (Zippenfenig, P. 2022) Por lo tanto, es posible utilizar Open-Meteo con propósitos comerciales o académicos, así como en el proyecto de investigación presente. Seguidamente, se describen las dependencias necesarias en el desarrollo y funcionamiento de la propuesta:

**Tabla N°23.** Dependencias de la propuesta y versiones.

Nombre	Versión
openmeteo_requests	2.31.0
request_cache	1.2.0
retry_request	2.0.0
pandas	2.2.1
numpy	1.26.0
datetime	5.5
sklearn.metrics	1.4.1.post1
sklearn.preprocessing	1.4.1.post1
matplotlib.pyplot	3.8.3
tensorflow.keras.models	2.15.2
tensorflow.keras.layers	2.15.2
tensorflow.keras.callbacks	2.15.2

***Fuente:*** Elaboración propia. (2024)

### 6.2.2 Factibilidad operativa

Esta parte hace referencia al personal involucrado en el desarrollo, implementación y mantenimiento del modelo LSTM a lo largo del tiempo, siendo parte fundamental para la realización de la investigación y el cumplimiento de sus respectivos objetivos. Para este proyecto, se consideró necesario utilizar personal para su correspondiente mantenimiento, ajuste y actualización de datos históricos y para monitoreo de precisión y rendimiento.

**Tabla N°24.** Requerimientos operativos.

Personal	Descripción
Personal de mantenimiento, ajuste y monitoreo. (Científico de datos Junior)	Personal encargado de la revisión periódica de la exactitud del modelo para corroborar la integridad de sus resultados, así como del preprocesamiento e integración de nuevos datos históricos meteorológicos y ejecución del entrenamiento para su actualización. Así como del monitoreo periódico de la precisión de las estimaciones del modelo con respecto a los valores reales y su correspondiente rendimiento de ejecución de pronósticos.
Desarrollador de modelo ML. (Científico de datos Senior)	Personal encargado de la construcción del modelo predictivo de ML, encargado del preprocesamiento y ajuste de datos, definición de la estructura del modelo, ejecución de entrenamiento inicial y ajuste de pesos. Evaluación de desempeño e implementación en sistemas.

***Fuente:*** Elaboración propia. (2024)

El personal de monitoreo estaría encargado de desarrollar evaluaciones semanales sobre la precisión con la que el modelo fue capaz de pronosticar los valores de la semana entre el último monitoreo y el actual, de manera que se compare el crecimiento del error a lo largo del tiempo y se evalúe la necesidad de realizar mantenimiento y ajustes sobre el modelo, incorporando nuevos datos históricos del tiempo ya transcurrido para actualizar la información sobre la cual el modelo desarrolla estimaciones, ejecutando su entrenamiento y pruebas de precisión y rendimiento. Mantenimiento que sería realizado mensualmente o cuando el error del modelo supere el mínimo establecido.

### 6.2.3 Factibilidad económica

En este apartado se van a detallar todos los aspectos económicos considerados para la implementación del modelo LSTM, de esta manera se puede determinar si el proyecto es financieramente viable.

#### 6.2.3.1 Costo de inicio del proyecto.

**Tabla N°25.** Inversión inicial del proyecto.

Descripción	Cantidad	Inversión (US\$)
Servidor local de entrenamiento	1	1120
Desarrollador e implementación de modelo ML. Científico de Datos Sr. (Por hora)	48	960
Equipo de desarrollo	1	350
Instalación de conexión de fibra óptica de 100mb.	1	100
Total		2530

*Fuente: Elaboración propia. (2024)*

Este apartado abarca los costos involucrados en el inicio del proyecto, incluyendo la inversión relacionada con el hardware necesario y el desarrollo del modelo LSTM.

#### 6.2.3.1 Costo mensual

**Tabla N°26.** Costos periódicos del proyecto.

Descripción	Inversión mensual (US\$)	Inversión anual (US\$)
Personal de mantenimiento, ajuste, actualización y monitoreo. Científico de datos Jr.	150	1800
Servicio de internet por fibra óptica de 100mb	35	420
Costo de licencia de software (Código Abierto)	0	0

Total	185	2220
-------	-----	------

***Fuente: Elaboración propia. (2024)***

En la tabla previa, se proporcionan valores estimados en cuanto al costo incurrido mensualmente al mantenimiento del modelo después de la inversión inicial necesaria para el comienzo del proyecto. Es importante tener en cuenta que los montos son valores aproximados y pueden variar dependiendo de la disponibilidad y las necesidades de procesamiento que la cantidad de valores por estimar presenten.

### **6.3 Metodología**

Para el desarrollo del modelo predictivo LSTM se empleó la metodología de Kanban, la cual es definida por Rehkopf, D. (2024) como “una herramienta ágil de gestión de proyectos diseñada para ayudar a visualizar el trabajo, limitar el trabajo en curso y maximizar la eficiencia” por lo que es la herramienta ideal para la descomposición de un proyecto altamente complejo en sus tareas primarias y así organizar su ejecución por orden de prioridad y jerarquía.

En el presente proyecto, para la utilización de la metodología Kanban, se decidió utilizar la plataforma gratuita de Trello, la cual permite desarrollar un tablero donde se representa visualmente cada actividad por realizar y su progreso a lo largo del proceso de desarrollo. Así como se observa en el Anexo N°1.

### **6.4 Objetivos de propuesta**

#### **6.4.1 Objetivo general**

Desarrollar un modelo de *machine learning* de tipo red neuronal recursiva *Long-short term memory* adaptado al pronóstico de las condiciones meteorológicas en base a los datos históricos meteorológicos del Estado Nueva Esparta, Venezuela.

#### **6.4.2 Objetivos específicos**

- Definir las dependencias, llamada a API y estructuración de datos.
- Desarrollar las cualidades exógenas de soporte a las variables meteorológicas.
- Preprocesar los datos meteorológicos históricos para su empleo como datos de entrenamiento y evaluación de desempeño de un modelo LSTM.

- Construir el modelo LSTM mediante definición de sus capas, dimensiones de entrada y salida y celdas por capa.
- Entrenar los modelos para cada pronóstico en base a las variables meteorológicas, las cualidades exógenas, los parámetros de paciencia e iteraciones.
- Evaluar el rendimiento y exactitud de los valores pronosticados y los valores reales.
- Construir el repositorio público de Github para el almacenamiento del proyecto.

## 6.5 Estructura de programación de la propuesta

### 6.5.1 Definición de dependencias, llamado a API y estructuración de datos.

Para la ejecución de la API y llevar a cabo ciertas manipulaciones sobre los datos ahora y más adelante, es necesario asegurarse de que ciertas dependencias estén instaladas en el entorno de desarrollo de Python, las cuales permitirán la ejecución de la petición y el manejo de caché, así como de reintentos, respectivamente, siendo estas: *openmeteo\_requests*, *requests\_cache* y *retry\_requests*. Es así, que se procede con la importación de las dependencias necesarias para la construcción del modelo, las cuales están descritas en la siguiente tabla:

**Tabla N°27.** Dependencias de la propuesta.

Nombre	Descripción
<i>openmeteo_requests</i>	Utilizada para la estructuración de la request en la API, con las variables deseadas, coordenadas y lapso temporal como parámetros
<i>request_cache</i>	Permite el empleo de caché para el almacenamiento temporal de los datos de la sesión en caso de ser necesarios
<i>retry_request</i>	Proporciona métodos para el reintento periódico de peticiones a la API, en caso de fracasar.
Pandas	Crucial para la estructuración de los datos históricos en dataframes para su rápida retribución y manipulación a través de métodos integrados y optimizados



Numpy	Permite la utilización de arreglos lineales numpy para la introducción de datos escalados al modelo LSTM.
Datetime	Dependencia que incluye el formato y los métodos necesarios para la manipulación de datos de tipo fecha, para la descomposición de las variables exógenas.
sklearn.metrics	Apartado de librería Sklearn con métodos integrados para el cálculo de métricas de desempeño del modelo y su entrenamiento.
sklearn.preprocessing	Otro apartado de la librería Sklearn que ofrece métodos integrados para el preprocesamiento de datos, más específicamente, su escalado relativo al mínimo y máximo.
matplotlib.pyplot	Librería que permite el desarrollo de gráficas de rendimiento y comparativas de datos durante el desarrollo del modelo.
tensorflow.keras.models	Apartado de la librería tensorflow que incluye las clases y los métodos necesarios para la construcción del modelo.
tensorflow.keras.layers	Incluye las clases de las capas que conforman al modelo de pronóstico LSTM.
tensorflow.keras.callbacks	Proporciona la capacidad de implementar los parámetros de paciencia en el entrenamiento del modelo para evitar sobre-entrenamiento

***Fuente:*** Elaboración propia. (2024)

La API a utilizar, de acuerdo con lo estipulado con anterioridad, será Open-Meteo, debido a su amplia oferta en cuanto a diversas variables meteorológicas, tomando en cuenta las variables a utilizar en el modelo. De manera que el request de la API requirió los parámetros de latitud y longitud del Estado Nueva Esparta, la fecha de inicio y fin de los datos meteorológicos, las variables meteorológicas por obtener y la zona temporal. Es así que se puede ejecutar la petición a la API.

A partir de ello, se retribuyen los datos de la respuesta y se convierten individualmente en arreglos unidimensionales usando el formato de *numpy*, para después ser incorporados a un

*dataframe* (df) de *pandas* que incluya todas las variables adjuntadas a un índice de tipo *datetime*, que incluye la hora, día, mes y año, donde cada columna representa una variable meteorológica en el instante de tiempo que indica el índice. De esta forma, se obtuvo un df con un total de 561048 filas y 13 columnas.

### 6.5.2 Desarrollo de las cualidades exógenas de soporte a las variables meteorológicas.

Las variables exógenas corresponden a otras variables que puedan ser obtenidas a partir de las ya existentes que facilitan al modelo la identificación de nuevos patrones, proceso denominado como *ingeniería de características*. Primeramente, se debe descomponer cualquier tipo de variable compleja o estructurada en varias variables distintas unas de otras, debido a que el modelo no podrá diferenciar valores complejos como lo podrían ser fechas enteras. Por lo que el índice de tipo *datetime* se convirtió en varias columnas que descomponen sus componentes, resultando una columna para el día de la semana, el día del mes, el mes del año, el año y la hora del día, facilitando la información necesaria al modelo para identificar patrones inherentes al momento en el día, mes y año.

De la misma manera, se desarrollaron cualidades exógenas de tipo retardadas. Las cuales consisten en retardar a lo largo de determinados saltos de tiempo (o filas) cada cualidad meteorológica y asignarla a una columna distinta, con el propósito de que para cada fila existan columnas que indican el valor de cada una de las variables meteorológicas un día, semana, mes y año en el pasado. Proporcionando así múltiples referencias extras para que el modelo logre pronosticar más precisa y rápidamente los valores del futuro. Este proceso resulta en un aumento significativo en la cantidad de columnas del *dataframe* para el entrenamiento, siendo las siguientes:

**Tabla N°28.** Cualidades exógenas.

Variable exógena	Descripción
Year	Contiene el año de cada medición individual
Month	Contiene el mes de cada medición individual
Day	Contiene el día del mes de cada medición individual
day_of_week	Contiene el día de la semana, donde el 1 es lunes y domingo es 7

retardado_7_Temperatura	Contiene el valor de la temperatura de una semana antes en el tiempo
retardado_7_Humedad_rel	Contiene el valor de la humedad de una semana antes en el tiempo
retardado_7_Punto_Dew	Contiene el valor del punto dew de una semana antes en el tiempo
retardado_7_et0_Evot	Contiene el valor de la evapotranspiración de una semana antes en el tiempo
retardado_7_Deficit_VP	Contiene el valor del déficit de presión de vapor de una semana antes en el tiempo
retardado_30_Temperatura	Contiene el valor de la temperatura de un mes antes en el tiempo
retardado_30_Humedad_rel	Contiene el valor de la humedad relativa de un mes antes en el tiempo
retardado_30_Punto_Dew	Contiene el valor del punto dew de un mes antes en el tiempo
retardado_30_et0_Evot	Contiene el valor de la evapotranspiración de un mes antes en el tiempo
retardado_30_Deficit_VP	Contiene el valor del déficit de presión de vapor de un mes antes en el tiempo
retardado_365_Temperatura	Contiene el valor de la temperatura de un año antes en el tiempo
retardado_365_Humedad_rel	Contiene el valor de la humedad relativa de un año antes en el tiempo
retardado_365_Humedad_rel	Contiene el valor del punto dew de un año antes en el tiempo
retardado_365_Punto_Dew	Contiene el valor del punto dew de un año antes en el tiempo
retardado_365_et0_Evot	Contiene el valor del déficit de presión de vapor de un año antes en el tiempo

***Fuente:*** Elaboración propia. (2024)

Estas variables de soporte permitirán al modelo identificar más patrones ocultos en los datos, con mayor impacto en este caso en particular debido a la inherente dependencia de las condiciones del clima con el momento en el año y el día. Por lo que las variables con un retardo anual servirían de referencia para pronosticar más precisamente las condiciones meteorológicas.

### **6.5.3 Preprocesamiento de los datos meteorológicos históricos para su empleo como datos de entrenamiento y evaluación de desempeño de un modelo LSTM.**

El primer paso correspondiente al preprocesamiento de datos es el escalamiento de los mismos a partir de los valores máximos y mínimos de cada variable, donde la media es representada por el valor 0 y la diferencia entre cada valor y la media es representada proporcionalmente entre 0, 1 y -

1, dependiendo de si es positiva o negativa. Debido a que el modelo tiene inconvenientes identificando patrones y relaciones entre variables en diferentes escalas, por lo que el proceso de escalarlos permite unificar la escala entre todos los factores involucrados en el pronóstico. En este caso, se utilizó la función *MinMaxScaler()* en el apartado de *sklearn.preprocessing*. A partir de lo cual se crea un objeto *scaler* que incluye el método de *fit\_transform*, que toma como parámetros el *dataframe* con los datos para entrenamiento, retornando un *dataframe* de forma equivalente con los datos ya escalados relativos a su media. Asimismo, ajusta el objeto *scaler* para convertir las estimaciones del modelo de vuelta a la escala original para ser observadas.

A partir de este punto, es posible desarrollar las secuencias y sus etiquetas. Las cuales son empleadas para que el modelo, durante su entrenamiento, desarrolle estimaciones sobre una sola variable tomando en cuenta una secuencia de todas las variables a la vez, en este caso las secuencias tienen un largo de 24 valores por columna y el modelo debe estimar el valor número 25 de la variable por estimar para compararlo con el verdadero valor, es decir, la primera etiqueta y modificar sus pesos internos en base al error observado, así sucesivamente. El resultado del proceso son dos *dataframes*, uno para las secuencias y otro para las etiquetas, los cuales son convertidos en arreglos bidimensionales del tipo *numpy*, necesarios debido a las limitaciones de compatibilidad en el entrenamiento del modelo. Es a partir de tales arreglos que se construyen los conjuntos de entrenamiento y prueba, para las secuencias y etiquetas por separados, resultando en 4 arreglos bidimensionales denominados:

**Tabla N°29.** División *train* y *test*.

Conjunto de datos	Descripción	Forma
train_x	Contiene el 80% de todos los valores en forma de secuencia sin la variable por pronosticar (etiqueta), destinados al entrenamiento.	(385295, 10, 29)
train_y	Contiene el 80% de las etiquetas (la variable por pronosticar) destinados al entrenamiento del modelo.	(385295,)
test_x	Contiene el restante 20% de las secuencias de todas las	(96324, 10, 29)

	variables (sin la etiqueta), para probar el aprendizaje del modelo entre cada epoch.	
test_y	Contiene el 20% de las etiquetas destinadas a probar el aprendizaje del modelo.	(96324,)

**Fuente:** Elaboración propia. (2024)

En la previa tabla se observa la división de los datos ya escalados y organizados en secuencias para la ejecución del entrenamiento y su puesta a prueba, donde los conjuntos de datos *train* serán empleados en el entrenamiento del modelo para la identificación de patrones y la construcción de las capacidades predictivas del modelo, y los conjuntos de datos *test* son utilizados en la puesta a prueba de los conocimientos del modelo para el ajuste de sus pesos internos durante el entrenamiento. Tomando en cuenta que los conjuntos con el sufijo *x*, en el apartado de forma, poseen tres valores; la cantidad total de filas (dependiente del porcentaje), el largo de las secuencias, la cantidad de columnas (variables meteorológicas y exógenas), por lo que poseen la mayor cantidad de datos para el entrenamiento y en base a la cual el modelo desarrolla estimación del futuro, cuando los conjuntos con sufijo *y* solo poseen un valor, la cantidad de filas, debido a que este son los valores reales de la variable por estimar para su comparación.

#### 6.5.4 Construcción del modelo LSTM mediante definición de sus capas, dimensiones de entrada y salida y celdas por capa.

El modelo se desarrolló empleando las librerías de *tensorflow.keras*, dentro de los apartados *models* y *layers*. Comenzando con la capa de inicialización *sequential* que define la estructura secuencia lineal del modelo desarrollado, continuando con las capas de *LSTM*, seguidas por *Dropout*, finalizando con la capa de *Dense*, que representa la capa de salida del modelo. Su estructura se visualiza en la siguiente tabla:

**Tabla N°30.** Descripción modelo LSTM.

Model: "model_LSTM"		
Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 10, 128)	6384
dropout_7 (Dropout)	(None, 10, 128)	0

lstm_1 (LSTM)	(None, 10, 64)	2408
dropout_7 (Dropout)	(None, 10, 64)	0
lstm_1 (LSTM)	(None, 10, 32)	616
dropout_7 (Dropout)	(None, 10, 32)	0
dense_4 (Dense)	(None, 1)	8

***Fuente: Elaboración propia. (2024)***

Para el pronóstico de cada modelo, se construyó un objeto del tipo modelo con la misma estructura, cuya diferencia se centra en la información que le es proporcionada al momento de entrenarlo, que define cuál variable meteorológica pronostica. La estructura del modelo expuesto en la previa tabla es la misma que el modelo de la tabla N°10, del modelo 3, que se determinó ser el mejor para esta determinada aplicación.

#### **6.5.5 Entrenamiento de los modelos para cada pronóstico en base a las variables meteorológicas, las cualidades exógenas, los parámetros de paciencia e iteraciones.**

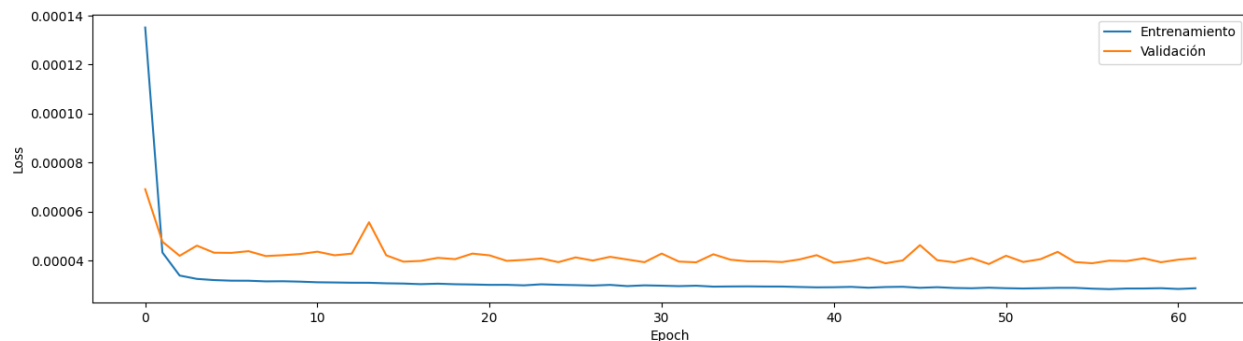
Para desarrollar pronósticos, los modelos deben ser entrenados individualmente para los valores de cada variable específica. Por ello, para la demostración de su funcionamiento, se entrenaron únicamente dos, en este caso: la variable temperatura y humedad relativa. Sin embargo, de acuerdo con su diseño y los estudios realizados, el modelo es capaz de pronosticar todas las variables incluidas en él, previamente identificadas.

De tal manera, el entrenamiento de los modelos consiste en un proceso iterativo, donde las iteraciones son denominadas *epochs*, mediante los cuales el modelo evaluará las secuencias en el conjunto de *train\_x* para luego estimar las etiquetas en el conjunto *test\_x*, y modificar sus pesos y estructura interna en base a la diferencia con los valores reales, obteniendo el error promedio de sus estimaciones. Luego, dentro del mismo epoch, se estiman los valores del conjunto de etiquetas *test\_y*, en base al conjunto de secuencias *train\_y*, de manera que se obtiene el error promedio del modelo durante su validación, tomando en cuenta que el modelo jamás interactúa con los valores reales del conjunto de validación, por lo que su estimación es imparcial y es así que se evalúa su desempeño en un contexto real. En tal sentido, durante cada iteración, el entrenamiento retornará un valor de error promedio (también llamado pérdida), para el entrenamiento y la validación, de

manera que se puede observar la curva de aprendizaje del modelo, lo cual permite diagnosticar problemas con el modelo con anticipación.

Antes del entrenamiento, es necesario definir el objeto *early\_stopping*, que indica la paciencia del entrenamiento de los modelos. Esto quiere decir que el modelo dejará de entrenar cuando deje de observar mejora durante una cierta cantidad de epoch, para evitar el sobre-entrenamiento, lo cual provocaría que el modelo esté mejor desarrollado para estimar los valores de entrenamiento y no los valores en su aplicación real. Finalmente, es posible entrenar los modelos y obtener la gráfica de entrenamiento:

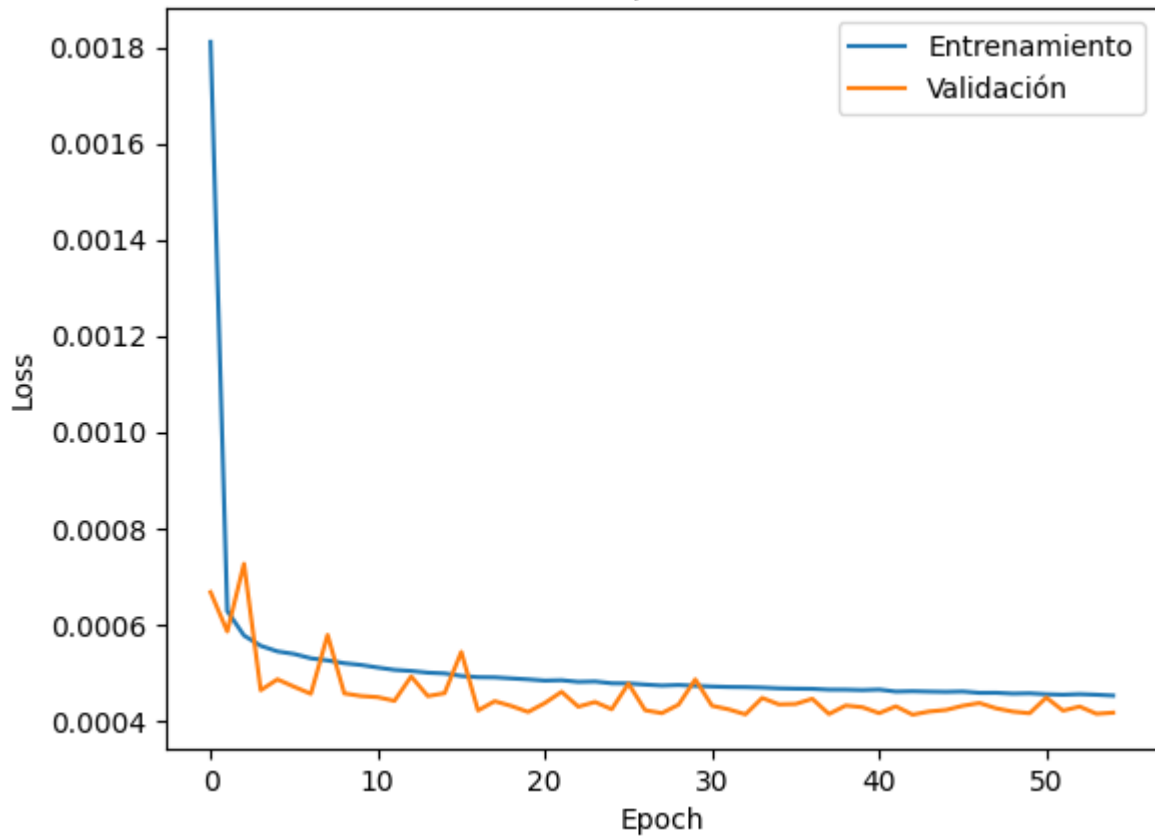
**Gráfica N°18.** Entrenamiento del modelo de humedad relativa LSTM.



**Fuente:** Elaboración propia. (2024)

En la gráfica previa, se observa el desempeño del modelo de humedad durante su entrenamiento, donde el eje x se refiere a la cantidad de iteraciones o epochs que el modelo necesitó para alcanzar su mejor rendimiento, cuyo error promedio era poco mayor a  $0.00004$  MSE. Asimismo, la gráfica refleja una curva de aprendizaje suave con un desempeño de entrenamiento y validación uniforme, con una muy baja diferencia entre ambos valores, indicando que el modelo logró abstraer toda la información posible de los datos de entrenamiento. Seguidamente, se observa la gráfica del entrenamiento del modelo de temperatura:

**Gráfica N°19.** Entrenamiento del modelo de temperatura LSTM.



***Fuente:*** Elaboración propia. (2024)

Similar al entrenamiento del modelo de humedad relativa, el entrenamiento del modelo de temperatura presenta una curva suave de aprendizaje durante su entrenamiento y un mejoramiento progresivo en su validación, casi finalizando en una convergencia entre ambas gráficas, lo cual representaría el resultado ideal. El modelo concluyó su entrenamiento a un punto similar al anterior, alrededor de los 50 epochs, indicando una extensión razonable de tiempo para haber aprendido todo lo posible de los datos proporcionados.

#### **6.5.6 Evaluación del rendimiento y exactitud de los valores pronosticados y los valores reales.**

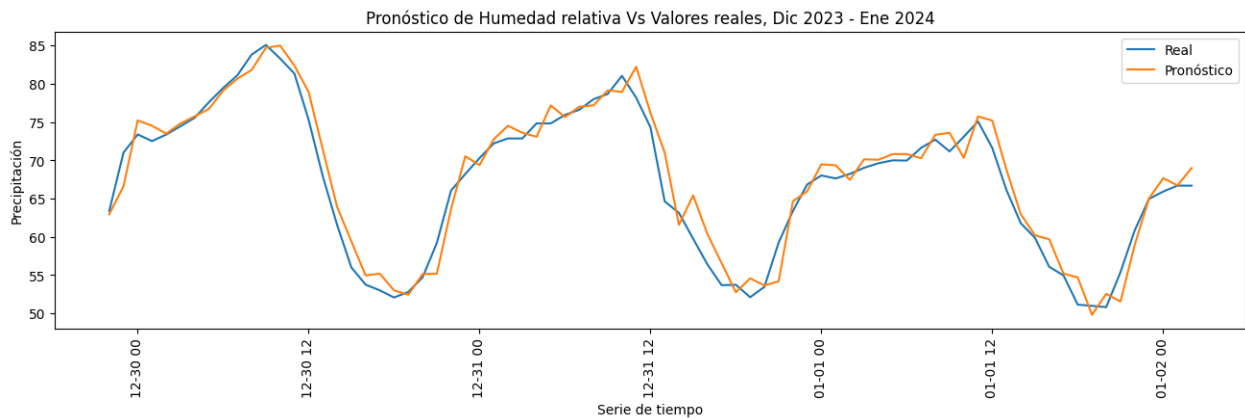
El resultado de ambos entrenamientos es la capacidad de desarrollar predicciones sobre los datos de validación para cada modelo (*test\_x*), se realiza mediante la ejecución del método integrado *model.predict()*, utilizando como parámetros los datos de validación. Por lo que su estimación se



extiende hasta el punto final de tales datos y así compararlos con los valores reales mediante una gráfica de ambos valores, para ambos modelos.

Comenzando con el modelo de humedad relativa, se ejecuta su pronóstico. Sin embargo, de la misma manera en la que el modelo requiere los datos escalados para su entrenamiento, sus resultados también estarían escalados con una media relativa a 0, lo cual no sería fácil interpretar directamente, por lo que se utiliza el objeto *scaler* ajustado a los datos escalados, para retornar los datos obtenidos del modelo a su escala original para ser interpretados fácilmente y comparados con los valores reales; resultando en la siguiente gráfica:

**Gráfica N°20.** Pronóstico de humedad relativa del modelo LSTM.



**Fuente:** Elaboración propia. (2024)

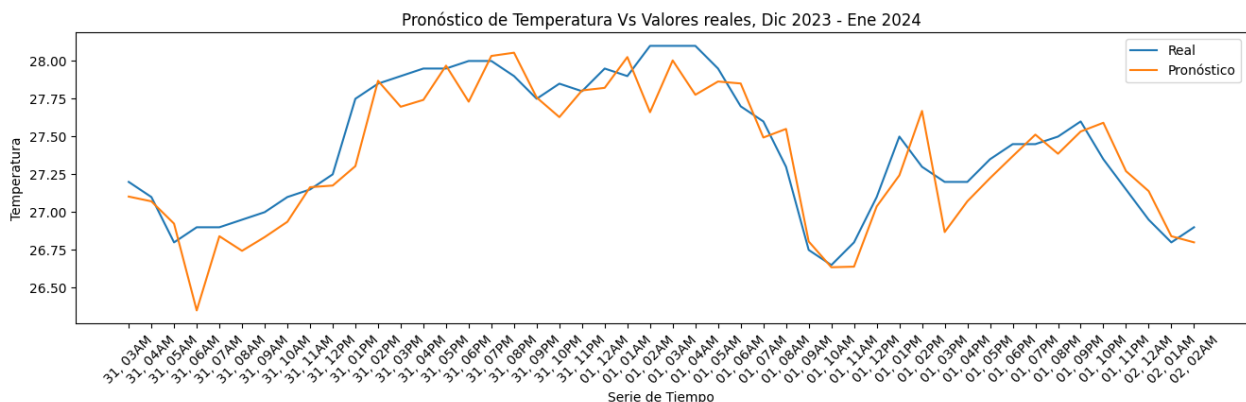
En la gráfica previa, se observa el contraste entre las estimaciones del modelo y los valores reales, de los últimos cuatro días en el *dataframe* original, correspondientes a los días 30 de diciembre de 2023 hasta el 2 de enero del 2024 en lapsos de 6 horas. Donde la línea naranja representa la variación de la humedad relativa durante ese periodo y la línea azul indica las estimaciones del modelo sobre esas horas, observándose que existe una considera sobreestimación por parte del modelo.

Seguidamente, se evalúa el desempeño del modelo de humedad calculando la métrica error absoluto promedio porcentual (MAPE). Para lo cual se emplea la librería métricas de sklearn, que incluye la librería *mean\_absolute\_percentage\_error*, resultando en 0.02680%. Este valor, como representa el error porcentual promedio, puede ser utilizado para obtener la exactitud de las estimaciones del modelo, resultando en: 99,9732% de exactitud. Corroborando la previa

conclusión del desempeño, que el modelo pronostica con alta precisión la frecuencia de la humedad relativa.

Seguidamente, se evalúa el desempeño del modelo de temperatura, el cual es evaluado de la misma manera y sobre el mismo lapso de tiempo. En este caso, sí es crucial que el modelo logre estimar la variación de la temperatura con una alta precisión tanto en valor como en frecuencia, sin embargo, la evidente dependencia de la temperatura con el instante temporal y la gran cantidad de cualidades exógenas proporcionan una grán cantidad de información para entrenar al modelo en el pronóstico de la temperatura. De la misma manera, se ejecuta la función *predict()* y se resaltan los valores resultantes para poder ser fácilmente entendidos, resultando en los siguientes valores:

**Gráfica N°21.** Pronóstico de temperatura del modelo LSTM.



**Fuente:** Elaboración propia. (2024)

Es así, que en la gráfica previa se observa la cercanía con la que el modelo es capaz de pronosticar la temperatura por hora de los siguientes 3 días, donde la línea azul representa los valores reales y la línea naranja representa el pronóstico que el modelo desarrolló sobre ese mismo periodo. Presentando una gran precisión en cuanto a la frecuencia y la escala con la que pronostica los valores, por lo que, considerando limitaciones y requerimientos, el desempeño del modelo de temperatura se encuentra dentro de los requerimientos. De la misma manera, se realiza el cálculo del error absoluto promedio porcentual (MAPE), que indicaría el porcentaje de error promedio que tiene las estimaciones del modelo con respecto a los valores reales, resultando en: 0.03902%, proporcionado un valor de exactitud de 99,9609%, representando un gran nivel de precisión en las estimaciones del modelo sobre los valores reales.

### **6.5.7 Construir el repositorio público de Github para el almacenamiento del proyecto.**

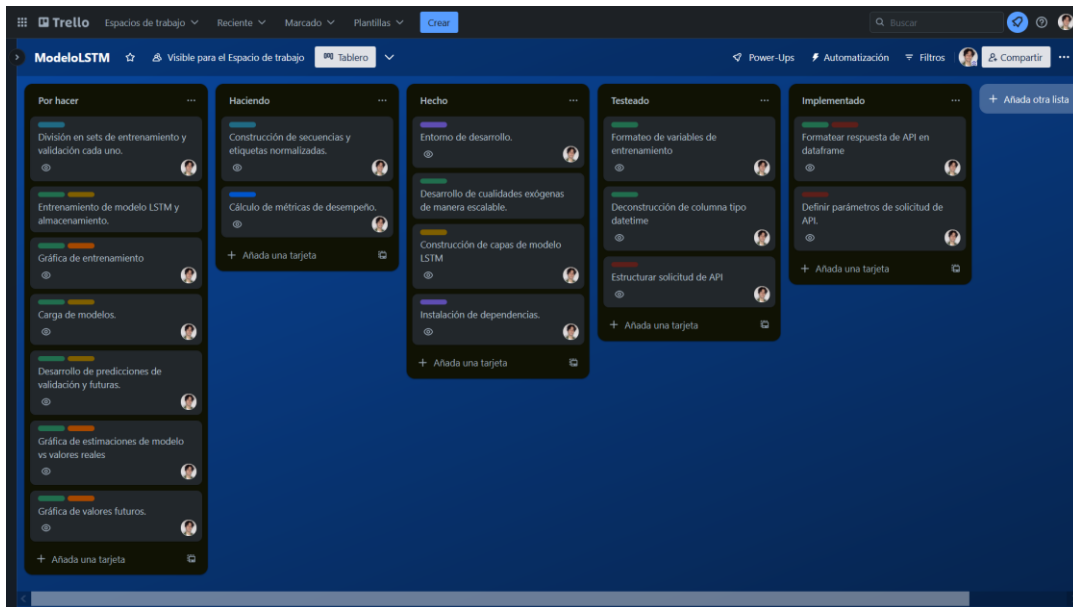
Finalmente, para la publicación del proyecto final, se desarrolló un repositorio público de Github que pueda almacenar y unificar el progreso y desarrollo que el sistema pueda experimentar durante el tiempo, por parte de la comunidad, primeramente. Por ello, la licencia de uso que acompaña el proyecto es MIT, debido a que “ofrece a los desarrolladores y organizaciones una opción de licencia permisiva y no restrictiva que fomenta el intercambio abierto de código” (AppMaster, 2023). La cual permite a usuarios de la comunidad utilizar, modificar y distribuir libremente el software, sin tomar en consideración regalías o restricciones legales.

Para lograr tal objetivo, se desarrollo el archivo *readme.md* el cual busca describir a detalle el funcionamiento del sistema. Primeramente, se proporciona una descripción general del proyecto y todo lo que lo compone, que pueda hacer el rol de una pequeña introducción para cualquier usuario de *github* interesado en el proyecto. Seguidamente, se explica el funcionamiento del sistema a detalle, abarcado todos los pasos involucrados en la solicitud de la API de datos meteorológicos históricos Open-Meteo, en el preprocesamiento de datos y el desarrollo de variables exógenas, la construcción y entrenamiento del modelo y, finalmente, el desarrollo de predicciones. Así como puede ser observado en el Anexo N.2. Donde se observa que el repositorio es identificado por el nombre *PronosticoTiempoLSTM*, el cual engloba las principales cualidades del proyecto.

Asimismo, el repositorio contiene los archivos en formato *Jupyter Notebook* (.ipynb), del entrenamiento de múltiples modelos desarrollados para pronosticar una variable en específico, así como el código necesario para guardar y cargar tales modelos ya entrenados y así desarrollar pronósticos futuros de las variables individualmente.

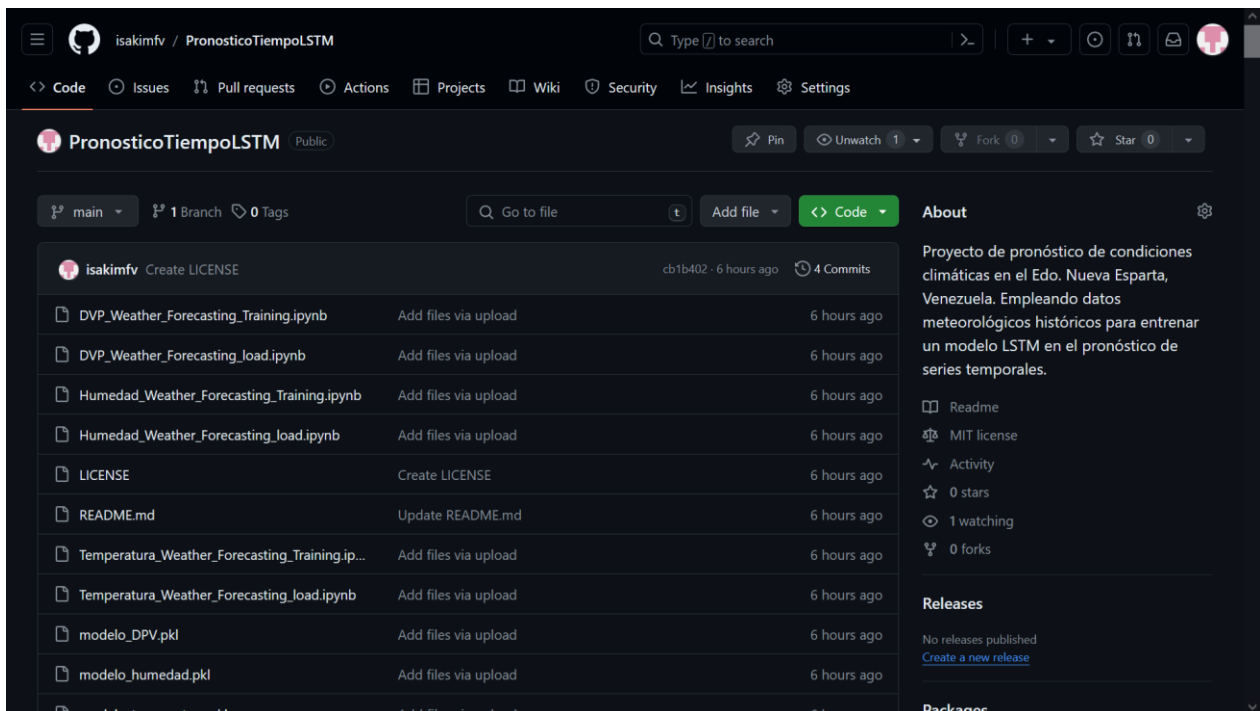
## ANEXOS

### Anexo N °1. Tablero Kanban de Trello durante el desarrollo de la propuesta.



*Fuente: Elaboración propia. (2024)*

### Anexo N °2. Archivos del repositorio PronosticoTiempoLSTM.



*Fuente: Elaboración propia. (2024)*

## REFERENCIAS

- Aditya, Y. (2023, 8 diciembre). *How Do You Validate A Time Series Model?*.  
<https://www.linkedin.com/advice/1/how-do-you-validate-time-series-model-printhow-skills-statistics?lang=es&originalSubdomain=es>
- Administration, National Oceanic and Atmospheric [NOAA]. (2019). *GFS*. Obtenido de EMC Home: [https://www.emc.ncep.noaa.gov/emc/pages/numerical\\_forecast\\_systems/gfs.php](https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php)
- Ahmed, I. (2023, 31 mayo). What are ACF and PACF Plots in Time Series Analysis? *Medium*.  
<https://ilyasbinsalih.medium.com/what-are-acf-and-pacf-plots-in-time-series-analysis-cb586b119c5d>
- AppMaster (2023). Licencias de código abierto. Recuperado el 09 de noviembre de 2023, de <https://appmaster.io/es/blog/licencias-de-codigo-abierto>
- Amat, J y Escobar, J. (2021). *Exogenous variables - Skforecast Docs*.  
[https://joaquinamatrodrigo.github.io/skforecast/0.10.1/user\\_guides/exogenous-variables.html#:~:text=Exogenous%20variables%20\(features\),enhance%20the%20accuracy%20of%20forecasts.](https://joaquinamatrodrigo.github.io/skforecast/0.10.1/user_guides/exogenous-variables.html#:~:text=Exogenous%20variables%20(features),enhance%20the%20accuracy%20of%20forecasts.)
- Arias, F. G. (2012). *El proyecto de Investigación*. Caracas: Episteme. Arthur, S. (1959). Algunos estudios sobre aprendizaje automático utilizando el juego de damas. *Revista IBM de investigación y desarrollo*, págs. 206-226.
- AWS. (s. f.). *¿Qué es una API?*. Amazon Web Services, Inc. <https://aws.amazon.com/es/what-is/api/>
- Blanco, J. I. (2023, 28 abril). “¿Por qué la normalización es clave e importante en Machine Learning y Ciencia de Datos? *Medium*. <https://jorgeiblanco.medium.com/por-qu%C3%A9->

[la-normalizaci%C3%B3n-es-clave-e-importante-en-machine-learning-y-ciencia-de-datos-4595f15d5be0](#)

Britannica, T. (5 de Abril de 2023). *Vilhelm Bjerknes*. Obtenido de Encyclopedia Britannica:

<https://www.britannica.com/biography/Vilhelm-Bjerknes>

Brownlee, J. (2019, 14 agosto). *Stacked Long Short-Term memory Networks*.

MachineLearningMastery.com. <https://machinelearningmastery.com/stacked-long-short-term-memory-networks/>

Brownlee, J. (2020, August 14). *How to Check if Time Series Data is Stationary with Python*.

MachineLearningMastery.com. <https://machinelearningmastery.com/time-series-data-stationary-python/>

Caraballo, C., Iglesias, L., García, F. (2010, 16 diciembre). *Reflexiones acerca del objeto de investigación y el campo de acción en una investigación*.

<https://mendive.upr.edu.cu/index.php/MendiveUPR/article/view/345>

Cázarez, L., Christen, M., Jaramillo E., Villaseñor L. y Zamudio, L (1999). *Técnicas actuales de investigación documental*. Editorial Trillas. Quinta Edición.

Chairani, S. (2022). The Correlation between Rainfall, Temperature, Relative Humidity, and Rice

Field Productivity in Aceh Besar. *IOP Conference Series: Earth and Environmental Science*, 1071(1), 012030. <https://doi.org/10.1088/1755-1315/1071/1/012030>

Chaudhuri, S. (2023, October 30). Assessment of accuracy metrics for time series forecasting.

*Medium*. <https://medium.com/analytics-vidhya/assessment-of-accuracy-metrics-for-time-series-forecasting-bc115b655705>

Danielkievych, A. (2023, 19 de julio). Top 3 Machine Learning Time Series Techniques: pros and

cons. *Forbytes*. <https://forbytes.com/blog/machine-learning-time-series-techniques/>

- Delua, J. (2021, 12 de marzo). Aprendizaje supervisado versus no supervisado: ¿cuál es la diferencia? *IBM*. <https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>
- Dhingra, D. (Julio 20 de 2023). *All you need to know about your first Machine Learning model – Linear Regression*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/>
- Fadul, A. O. (2004). *Diseño estructurado de algoritmos*. Alexander Oviedo Fadul.
- Faggella, D. (26 de Febrero de 2020). *Emerj*. Obtenido de What is machine learning?: <https://emerj.com/ai-glossary-terms/what-is-machine-learning/>
- Garcia, R., & Gross, R. (1994). *Diccionario práctico de español moderno*.
- Gnoza, N. Barberena, M. (2018). *Estudio de factibilidad del uso de Machine Learning con múltiples fuentes de datos en el pronóstico del tiempo*. [Tesis de grado inédita]. Universidad ORT Uruguay.
- Gordon, J. (2023, 5 diciembre). *Practical Guide for Feature Engineering of Time Series Data*. dotData. <https://dotdata.com/blog/practical-guide-for-feature-engineering-of-time-series-data/>
- Jaramillo, I., & Ramírez, R. (2006). *Método y conocimiento: metodología de la investigación : investigación cualitativa/investigación cuantitativa*. Universidad Eafit.
- Korstanje, J. (2023, 31 julio). *How to select a model for your Time Series Prediction Task [Guide]*. neptune.ai. <https://neptune.ai/blog/select-model-for-time-series-prediction-task>
- Krishnan, S. (2022, 6 agosto). How to determine the number of layers and neurons in the hidden layer?. *Medium*. <https://medium.com/geekculture/introduction-to-neural-network-2f8b8221fbd3#:~:text=Number%20of%20Neurons%20In%20Input,as%20a%20regressor%20or%20classifier.>

- Kutskov, K. (2023, 31 julio). *ARIMA vs Prophet vs LSTM for Time Series Prediction*. neptune.ai. <https://neptune.ai/blog/arima-vs-prophet-vs-lstm>
- Latam, A. (2023, March 24). *Mejora del análisis con Boxplot*. Alura. <https://www.aluracursos.com/blog/mejora-del-analisis-con-boxplot>
- Ley de Aeronáutica Civil (2001, 8 de septiembre) Normas Legales, N° 37293, Gaceta Oficial de la República Bolivariana de Venezuela. Ley de Meteorología e Hidrología Nacional (2006, 22 de diciembre) Normas Legales, N° 5833, Gaceta Oficial de la República Bolivariana de Venezuela.
- Luca (2020) ¿Qué es Python?. *Luca-d3.com*. <https://web.archive.org/web/20200224120525/https://luca-d3.com/es/data-speaks/diccionario-tecnologico/python-lenguaje>
- Luna, J. P., & Gallo, G. I. (2018). *Incremental: Una Visita Guiada Al Mundo Emprendedor*.
- Malkari, N. (2023, April 17). Seasonal Decomposition - Nikhil Malkari - medium. *Medium*. <https://medium.com/@nikhilmalkari18/seasonal-decomposition-425a2d7490e8>
- Martins, J. (2022, 6 diciembre). Cómo entender los procesos iterativos (con ejemplos) [2022] • Asana. *Asana*. <https://asana.com/es/resources/iterative-process>
- Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y aplicaciones*. Santa Fe.
- Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Universidad Tecnológica Nacional – Facultad Regional Rosario. [https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5\\_anio/orientadora1/monograias/matich-redesneuronales.pdf](https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf)
- Medina, J. Dr. Durante, C. (2007). *Cálculo del parámetro de radioatenuación troposférica en banda de 0.4 a 60 ghz en el estado Nueva Esparta* [Tesis de magíster inédita]. Universidad Dr. Rafael Belloso Chacín.



Merri-Webster. (s.f). Aprendizaje. *Definición & Significado*. Recuperado el 15 de noviembre de 23, de <https://www.merriam-webster.com/dictionary/learning>

Mulani, S. (2023, 16 febrero). RMSE - Root Mean Square Error in Python - AskPython. *AskPython*.  
<https://www.askpython.com/python/examples/rmse-root-mean-square-error>

Nilsson, N. (1998). Artificial intelligence: a new synthesis.

Norving, S. R. (2008). *Inteligencia Artificial: Un enfoque moderno*. Madrid: PEARSON EDUCACIÓN, S.A.

Nvidia. (s.f.). *Machine learning - What it is and why does it matter*. Obtenido de Nvidia:  
<https://www.nvidia.com/en-us/glossary/data-science/machine-learning/>

Onretrieval. (2023, 29 enero). *¿Qué es Siri y para qué sirve?* OnRetrieval.  
<https://onretrieval.com/que-es-siri-y-para-que-sirve/>

Orellana, J. (2019). *Ucuenca*. Obtenido de ¿Qué es el machine learning y por qué es popular?:  
<https://www2.ucuenca.edu.ec/component/content/article/233-espanol/investigacion/blog-de-ciencia/1222-machine-learning?Itemid=437>

Organización Meteorológica Mundial. (s.f.). *Clima*. Obtenido de World Meteorological Organization: <https://public.wmo.int/en/our-mandate/weather>

OverLordGoldDragon (2019) What is the rule to know how many LSTM cells and how many units in each LSTM cell do you need in Keras?. Stack Overflow.  
<https://stackoverflow.com/questions/59072728/what-is-the-rule-to-know-how-many-lstm-cells-and-how-many-units-in-each-lstm-cel>

*Oxford Languages and Google - Spanish | Oxford Languages*. (2022, 15 febrero).  
<https://languages.oup.com/google-dictionary-es/>

Pozo, T. G. (2008). *Física y Química, 3 ESO: ciencias de la naturaleza*.

- Real Academia Española: *Diccionario de la lengua española*, 23.<sup>a</sup> ed., [versión 23.6 en línea].  
<https://dle.rae.es> [20 de noviembre de 2023]
- Rob J Hyndman (2010, 29 de septiembre). *Forecasting with long seasonal periods*. .  
<https://robjhyndman.com/hyndsight/longseasonality/>
- Rodríguez, R., Benito, A., y Portela, A. (2004). *Meteorología y Climatología: semana de la Ciencia y la Tecnología 2004*.
- Sampieri, R., Collado, F., Lucio, B. (2003). *Metodología de la investigación*.
- Setchell, H. (s. f.). *About*. ECMWF. <https://www.ecmwf.int/en/about>
- Shah, D (2021). "Pronóstico de temperatura a corto plazo usando LSTMS, y CNN". [Tesis de grado]. Instituto de tecnología de New Jersey. <https://digitalcommons.njit.edu/theses/1840>
- Software DELSOL. (2021, 5 enero). *Retroalimentación ¿Qué es?, ¿Qué tipos existen?* Software del Sol. <https://www.sdelisol.com/glosario/retroalimentacion/>
- Solé, R. V., & Manrubia, S. C. (2009). *Orden y caos en sistemas complejos. aplicaciones*. Univ. Politèc. de Catalunya.
- Solis, D. C. (2023, 24 octubre). Datasets: qué son y cómo acceder a ellos. *OpenWebinars.net*.  
<https://openwebinars.net/blog/datasets-que-son-y-como-acceder-a-ellos/>
- Tamayo, M., Tamayo, Y. (2001). *El proceso de la investigación científica*. Editorial Limusa.
- Tendencias de la Tecnología - Inteligencia Artificial*. (s.f.). Obtenido de OMPI:  
[https://www.wipo.int/tech\\_trends/es/artificial\\_intelligence/](https://www.wipo.int/tech_trends/es/artificial_intelligence/)
- Teseo. Marín, D., Pineda, I. (2019). *Modelo predictivo Machine Learning aplicado a análisis de datos Hidrometeorológicos para un SAT en Represas*. [Tesis de grado]. Universidad Tecnológica del Perú.
- Zippenfenig, P. Open-Meteo.com Weather API [Computer software].  
<https://doi.org/10.5281/zenodo.7970649>

Zippenfenig, P. (2022). *open-meteo/LICENSE at main · open-meteo/open-meteo*. GitHub.

<https://github.com/open-meteo/open-meteo/blob/main/LICENSE>