UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

PROYECTO DE TESIS 1

Análisis de modelos de Machine Learning aplicados a la predicción de El Niño - Oscilacion del Sur (ENSO) en la Costa Norte del Perú

Autor:

Armando Alberto Lluen Gallardo

Asesor:

Msc. Yuri Nuñez Medrano

1 de julio del 2024

Resumen

El Fenómeno El Niño-Southern Oscillation (ENSO) ejerce un impacto considerable en el clima global y afecta profundamente a la costa Norte del Perú, especialmente por los cambios en las temperaturas superficiales del mar asociados con la corriente de El Niño. Esta investigación evalúa la eficacia de métodos avanzados de Machine Learning, incluyendo Random Forest, redes neuronales convolucionales y recurrentes, para predecir ENSO basándose en datos históricos de Temperatura Superficial del Mar (TSM) a lo largo de la costa peruana. Se emplea además el Índice Costero El Niño (ICEN) para categorizar las fases del fenómeno. Los resultados destacan que los modelos de redes neuronales ofrecen un rendimiento superior en la predicción de ENSO, comparados con otros enfoques evaluados. La discusión se centra en la interpretación de estos resultados y sus implicancias para el pronóstico preciso del Fenómeno El Niño en el contexto peruano, subrayando tanto las ventajas como las limitaciones de las técnicas aplicadas.

PALABRAS CLAVE: El Niño-Southern Oscillation (ENSO), Predicción climática, Temperaturas marinas, Perú, Costa peruana, Inteligencia artificial, Machine Learning, Redes neuronales, Modelos de predicción, Impacto del Niño ENSO, Anomalía térmica, RNN (Redes Neuronales Recurrentes), CNN (Redes Neuronales Convolucionales), Random Forest.

Índice general

Resumen III

1 Introducción 1

1.1 Motivación 1

1.2 Objetivo General 3

1.3 Estructura del Seminario 4

2 Estado del Arte 7

2.1 Marco teórico 7

2.1.1 El Niño - Oscilacion del Sur (ENSO): 7

Componentes 7

Fases del Ciclo ENSO 8

Regiones de Monitoreo ENSO 11

Indices 12

2.1.2 Machine Learning 14

Tipos de Machine Learning 14

Modelos 16

2.2 Revisión de Literatura 31

2.2.1 Métodos de Predicción 32

2.2.2 Modelos con Machine Learning 33

3 Herramientas y Metodología 40

3.1 Área de Estudio 40

3.2 Herramientas 40

3.2.1 Lenguaje de Programación 40

3.2.2 Datos 42

3.2.3 Entorno de Desarrollo 42

3.3 Metodología 43
3.3.1 Preprocesamiento de Datos 43
3.4 Visualización de los Datos 45
3.4.1 Modelamiento 53
División de Datos 53
Modelos utilizados 53
3.4.2 Catalogación del ENSO 56

4 Resultados **58**
4.1 Resultados de los Modelos 58
4.2 Proyecciones 2024 y 2025 64

5 Análisis de Resultados **67**
5.1 Rendimiento y precisión de los Modelos 67
5.2 Pronóstico del ICEN y el ENSO 68
5.3 Proyecciones para el 2024 y 2025 69

6 Conclusiones y Trabajo Futuro **72**
6.1 Conclusiones 72
6.2 Trabajo Futuro 73

Índice de figuras

2.1 Regiones del ENSO. Imagen tomada del Ministerio del Ambiente [19]. 11
2.2 Secuencia del Modelo Boosting [10]. 20
2.3 Secuencia del Modelo XGBoost [69]. 21
2.4 Secuencia del Modelo Random Forest [42]. 23
2.5 Arquitectura de una Neuronna [67]. 25
2.6 Red Neuronal con Capa Densa [25]. 26
2.7 Red Neuronal Convolucional [25]. 27
2.8 Red Neuronal Recurrente Básica [43]. 28
2.9 Neurona artificial LSTM [46] 29
2.10 Neurona artificial GRU [46]. 29
2.11 Funciones de Activación más comunes [46]. 30
2.12 Arquitectura de una Red Neuronal [31]. 31

3.1 Temperatura Superficial del Mar - 2018. 46
3.2 Temperatura Superficial del Mar - 2019. 46
3.3 Temperatura Superficial del Mar - 2020. 47
3.4 Temperatura Superficial del Mar - 2021. 48
3.5 Temperatura Superficial del Mar - 2022. 49
3.6 Temperatura Superficial del Mar - 2023. 50
3.7 TSM y Medias Móviles desde 2017 hasta 2023. 51

4.1 TSM predicha por el modelo Random Forest vs. mediciones reales. 59
4.2 ICEN predicho por el modelo Random Forest. 59
4.3 TSM predicha por el modelo Árbol de Regresión vs. mediciones reales. 60
4.4 ICEN predicho por el modelo Árbol de Regresión. 60
4.5 TSM predicha por el modelo XGBoost vs. mediciones reales. . . . 61

4.6 ICEN predicho por el modelo XGBoost. 61
4.7 TSM predicha por el modelo Red Neuronal Recurrente vs. mediciones reales. 62
4.8 ICEN predicho por el modelo Red Neuronal Recurrente. 62
4.9 TSM predicha por el modelo Red Convolucional vs. mediciones reales. 62
4.10 ICEN predicho por el modelo Red Convolucional. 63
4.11 Proyecciones de la Temperatura Superficial del Mar del 2024 y 2025 con CNN. 64
4.12 Proyecciones de la Temperatura Superficial del Mar del 2024 y 2025 con RNN. 64
4.13 Proyecciones del ICEN del 2024 y 2025con CNN. 65
4.14 Proyecciones del ICEN del 2024 y 2025 con RNN. 65

5.1 ICEN desde el 2000 hasta 2024 según IMARPE [20]. 69

Índice de tablas

2.1 Categorías del Índice Térmico Costero Peruano (ITCP) [20]. 13
2.2 Categorías del Índice Costero El Niño (ICEN). 13
2.3 Categorías del Índice de Oscilación del Sur (ONI). 14

3.1 Temperaturas mensuales (mínima, máxima, promedio) para el año 2017. 45
3.2 Descripción de los datos de temperatura promedio para el año 2017. 46
3.3 Temperaturas mensuales (mínima, máxima, promedio) para el año 2018. 47
3.4 Descripción de los datos de temperatura promedio para el año 2018. 47
3.5 Temperaturas mensuales (mínima, máxima, promedio) para el año 2019. 48
3.6 Descripción de los datos de temperatura promedio para el año 2019. 48
3.7 Temperaturas mensuales (mínima, máxima, promedio) para el año 2020. 49
3.8 Descripción de los datos de temperatura promedio para el año 2020. 49
3.9 Temperaturas mensuales (mínima, máxima, promedio) para el año 2021. 50
3.10 Descripción de los datos de temperatura promedio para el año 2021. 50
3.11 Temperaturas mensuales (mínima, máxima, promedio) para el año 2022. 51

3.12 Descripción de los datos de temperatura promedio para el año 2022. 51
3.13 Temperaturas mensuales (mínima, máxima, promedio) para el año 2023. 52
3.14 Descripción de los datos de temperatura promedio para el año 2023. 52
3.15 Descripción de los datos combinados de temperatura promedio (2017-2023). 52

4.1 Comparación de métricas de modelos. 59

Índice de ecuaciones

2.1 formula de la Impureza de Gini 17
2.2 Fórmula de la Entropía 17
3.1 Rango intercuartílico (IQR) 44
3.2 Media móvil de 7 días 44
3.3 Media móvil de 30 días 44

Índice de Acrónimos

API : Application Programming Interface
AR : Modelos Autoregresivos (Autoregressive Models)
ARIMA : Autoregressive Integrated Moving Average Models
AWS : Amazon Web Services
CART : Classification and Regression Trees
CNN : Redes Neuronales Convolucionales (Convolutional Neural Networks)
CMIP5/6 : Coupled Model Intercomparison Project Phase 5/6
CPU : Central Processing Unit
CSV : Comma-Separated Values
ECMWF : European Centre for Medium-Range Weather Forecasts
EEMD : Ensemble Empirical Mode Decomposition
ENSO : El Niño-Southern Oscillation
ENFEN : Comité Multisectorial Encargado del Estudio Nacional del Fenómeno El Niño
GRU : Gated Recurrent Unit
HTML : HyperText Markup Language
IQR : Interquartile Range
IRI : International Research Institute for Climate and Society
ICEN : Índice Costero El Niño
ID3 : Iterative Dichotomiser 3
IMARPE : Instituto del Mar del Perú
ITCP : Índice Térmico Costero Peruano
JPEG : Joint Photographic Experts Group
JSON : JavaScript Object Notation
LSTM : Long Short-Term Memory
MA : Modelos de Media Móvil (Moving Average Models)
MAE : Mean Absolute Error
MJO : Madden-Julian Oscillation
NCEP : National Centers for Environmental Prediction

Índice de Acrónimos

NOAA	:	National Oceanic and Atmospheric Administration
NP	:	Problema de complejidad no polinomial
OS	:	Oscilación del Sur
PCA	:	Principal Component Analysis
PDF	:	Portable Document Format
RNN	:	Redes Neuronales Recurrentes (Recurrent Neural Networks)
RMSE	:	Root Mean Square Error
SENAMHI	:	Servicio Nacional de Meteorología e Hidrología del Perú
SOI	:	Índice de Oscilación del Sur (Southern Oscillation Index)
SQL	:	Structured Query Language
SST	:	Sea Surface Temperature
SVD	:	Singular Value Decomposition
TSM	:	Temperatura de la Superficie del Mar
TCN	:	Temporal Convolutional Networks
URL	:	Uniform Resource Locator
ZCIT	:	Zona de Convergencia Intertropical

Agradecimientos

Quiero expresar mi sincero agradecimiento a diversas personas e instituciones que han sido fundamentales en la realización de este trabajo de tesis.

A mi familia, especialmente a mi padre Javier Lluen y mi madre Yulitza Gallardo, por su incondicional apoyo económico y emocional a lo largo de esta travesía académica. A mi abuela Juana Gallardo y tía Karen Gallardo por su constante aliento y apoyo desde Piura.

A mis amigos Mariana Jara, Fabiola Lazo, Oscar Escajadillo, Karla Durand y Mary Luz Torres, quienes estuvieron siempre presentes en los momentos de mayor estrés, brindándome su apoyo incondicional y ayudándome a mantener la calma, lo cual fue crucial para continuar y finalizar este proceso académico.

A mi asesor Yuri Nuñez, por su invaluable orientación, paciencia y dedicación. Sus consejos expertos y seguimiento constante fueron esenciales para la elaboración y mejora de este trabajo.

A la Facultad de Ciencias de la Universidad Nacional de Ingeniería, que me ha proporcionado una formación sólida en ciencias de la computación. Los conocimientos adquiridos en esta institución han enriquecido profundamente mi comprensión científica y han fortalecido mi capacidad crítica y analítica, aspectos esenciales para la realización de este trabajo de investigación.

Capítulo 1

Introducción

El Niño - Oscilación del Sur (ENSO) es uno de los eventos climáticos más complejos y recurrentes que afectan al clima global. Sus consecuencias en las costas peruanas son muy notables, especialmente por los cambios en las temperaturas marinas. La predicción precisa del ciclo del ENSO asi como intensidad es vital para enfrentar sus efectos y mitigar los daños que este pueda causar. Aquí es donde la inteligencia artificial emerge como una herramienta prometedora, mejorando la precisión de las predicciones climáticas. Este estudio se enfoca en analizar métodos de Machine Learning para predecir el ENSO basados en datos de la temperatura superficial del mar en las costas del Perú.

1.1. Motivación

A lo largo de la historia hemos vivido con la incertidumbre de la ocurrencia de fenómenos naturales, ante la imprevisibilidad del clima y sus consecuencias materiales y humanas. Pero ahora, debido al impacto del cambio climático, se ha añadido un factor de intensidad al desarrollo industrial, que es el resultado del propio comportamiento contaminante de las personas en todo el mundo. Nuestro país no es ajeno a esto ya que se encuentra ubicado en diferentes regiones geográficas y es propenso a sufrir diversos desastres naturales como terremotos, tsunamis, inundaciones y deslizamientos de tierra.

Por lo tanto, durante la última década, nuestro país ha enfrentado enormes desafíos para gestionar el riesgo y mitigar el impacto de estos eventos. Uno de los mayores desastres naturales que enfrenta nuestro país son las inundaciones, especialmente durante períodos de fuertes lluvias o eventos de El Niño.

Cuando los ríos se desbordan y causan estragos en las comunidades costeras y de las tierras bajas, pueden tener consecuencias catastróficas como la destrucción de viviendas e infraestructura de edificios, pérdida de cosechas, plagas y enfermedades que, junto con la escasez de recursos económicos, causan grandes angustias. . Otro de los peligros naturales más comunes son los deslizamientos de tierra, los cuales son una amenaza constante, especialmente en zonas montañosas y escarpadas, donde la deforestación y las actividades humanas aumentan el riesgo de deslizamientos de tierra. Estos eventos pueden dañar viviendas, bloquear carreteras y aislar comunidades, dificultando los esfuerzos de rescate y recuperación. En este contexto, El Niño se convierte en un gran desafío porque puede causar desastres naturales extremos que afectan a comunidades enteras y economías nacionales.

Cuando vivía en Piura, en la región norte del Perú, vi de primera mano la devastación causada por el clima extremo. Las fuertes lluvias, las inundaciones de los ríos y los deslizamientos de tierra son fenómenos comunes que causan algunas pérdidas económicas y humanas. Vivir la devastación causada por los desastres naturales hizo que me interesara comprender y abordar los riesgos climáticos que enfrentamos como sociedad, y cómo anticipar estos riesgos para evitar pérdidas futuras. Mi motivación para realizar esta investigación surgió de mi preocupación por la falta de preparación y prevención de El Niño en nuestro país. El Niño se percibe como impredecible y repentino, lo que lleva a una peligrosa complacencia entre las autoridades de las zonas más afectadas a la hora de planificar y responder al fenómeno. He visto de primera mano las consecuencias de la falta de preparación en las comunidades locales y los asentamientos humanos, especialmente los más vulnerables.

Como residente de una de las zonas más afectadas por este fenómeno, conozco de primera mano el sentimiento de impotencia y vulnerabilidad que crean estos hechos. Me sorprende cómo la gente de estas zonas intenta recuperarse de cada desastre para volver a verse afectada después del siguiente desastre natural. Esta realidad me impulsa a buscar soluciones que puedan ayudar a prevenir o mitigar los efectos de futuros El Niño y otros desastres naturales. Además, la comprensión de que el cambio climático está aumentando la intensidad y frecuencia de estos eventos extremos ha

aumentado la necesidad de encontrar formas efectivas de predecir y responder a los desastres naturales. Teniendo esto en cuenta, creo firmemente que los avances en ciencia y tecnología, especialmente en el campo de la inteligencia artificial, brindan nuevas oportunidades para abordar estas cuestiones de manera más efectiva. Fue la simbiosis entre mi experiencia personal y mi creencia en el poder de la tecnología lo que me motivó a realizar esta investigación. Mi objetivo es hacer todo lo que pueda para crear un futuro más seguro y sostenible para las generaciones futuras, donde los desastres naturales ya no sean una gran amenaza, sino un desafío manejable al que podamos responder con éxito.

1.2. Objetivo General

Investigar y analizar métodos de inteligencia artificial para mejorar la predicción del fenómeno del Niño ENSO basándose en datos de temperaturas marinas en las costas del Perú, con el fin de proporcionar una base sólida para la gestión preventiva y la reducción del impacto de este fenómeno en la región.

Específicamente, los objetivos de este trabajo son:

- Recopilar y analizar datos históricos de temperaturas marinas en las costas del Perú, así como datos asociados con ENSO, para construir un conjunto de datos completo y representativo.
- Implementar y ajustar modelos de machine learning, para predecir las variaciones del fenómeno ENSO en función de las temperaturas marinas de la costa norte del Perú.
- Hacer una proyección sobre el ENSO en la costa norte del Perú para el presente año.
- Proporcionar recomendaciones prácticas y sugerencias para la aplicación futura de métodos de Machine Learning en la predicción y gestión del fenómeno del Niño ENSO en las costas del Perú.

1.3. Estructura del Seminario

- **Introducción:**
  En este capitulo introductorio se expone la motivación para desarrollar el tema así como el objetivo general y específicos y un panorama general del documento.

- **Estado del Arte:**
  La revisión de las tecnicas inteligencia artificial en la predicción del Niño, ademas de las limitaciones y desafíos actuales en este campo. Se identifican lagunas en la investigación existente y se destacan las oportunidades para la mejora de los modelos de predicción.

- **Herramientas y Metodología:**
  Se detallan las herramientas y tecnologías utilizadas en la investigación, incluyendo software de análisis de datos, lenguajes de programación y bibliotecas de inteligencia artificial. Se explican la implementación de modelos y la evaluación de resultados.

  Además, se presenta una descripción de los datos utilizados, incluyendo su origen, frecuencia temporal y características. Se describe el proceso de recopilación, limpieza y preparación de datos, junto con un análisis exploratorio para comprender las tendencias temporales.

  Finalmente, se describe el diseño experimental y la metodología utilizada, que incluye la selección de modelos de Machine Learning, técnicas de evaluación, implementación de modelos ajustados para predecir el Fenómeno El Niño (ENSO), selección de hiperparámetros y métricas para evaluar el rendimiento.

- **Resultados:**
  Se describe los resultados obtenidos de los diferentes modelos aplicados asi como las proyecciones realizadas.

- **Análisis de Resultados:**
  Se presentan y discuten los resultados del Niño ENSO utilizando los diferentes métodos. Se compara el rendimiento de los modelos de

series temporales y redes neuronales, interpretando los resultados y su relevancia para la predicción en la región costera del Perú.

- **Conclusiones y Trabajo Futuro:**
  Se resumen los hallazgos principales y conclusiones de la investigación, reflexionando sobre sus implicaciones prácticas y sugerencias para futuras investigaciones y mejoras metodológicas.

Capítulo 2

Estado del Arte

El ENSO representa el mayor cambio climático en los trópicos que afecta al globo y afecta a los océanos Pacífico e Índico. Estudios recientes muestran que el calentamiento del Océano Pacífico, que comenzó a fines de la década de 1990, ha afectado el clima global, por lo que comprender las interacciones entre estos océanos tropicales es esencial para mejorar las predicciones climáticas actuales y futuras [65,71].

2.1. Marco teórico

2.1.1. El Niño - Oscilacion del Sur (ENSO):

El Niño - Oscilacion del Sur o ENSO por sus siglas en ingles, se refiere a las variaciones en las condiciones atmosféricas y oceánicas, que surgen de las fluctuaciones en la temperatura superficial del mar (TSM) y la presión atmosférica a lo largo del océano Pacífico tropical y que tiene efectos globales. ENSO está compuesto por dos componentes principales que reflejan su naturaleza compleja y acoplada: El Niño (oceánico) y la Oscilación del Sur (atmosférico) [45].

Componentes

1. El Niño (componente oceánica): hace referencia al proceso que ocurre durante la fase cálida de El Niño-Oscilación del Sur (ENSO). Este mecanismo implica el calentamiento anómalo de las aguas oceánicas desde la Línea Internacional de Cambio de Fecha hasta la costa oeste de América del Sur, es decir un calentamiento de alrededor de 3°C

y 5°C en las costas de Colombia, Ecuador y Perú, este calentamiento provoca cambios en la ecología local y regional [11, 66], causando un desplazamiento de las corrientes marinas y cambios en la circulación atmosférica.

2. La Oscilación del Sur (componente atmosférica) : se refiere a las fluctuaciones de la presión atmosférica entre el Pacífico occidental (cercano a Australia e Indonesia) y el Pacífico oriental (cerca de América del Sur). Estas variaciones en la presión atmosférica están estrechamente ligadas al movimiento de los vientos alisios y a la distribución de las temperaturas del mar [11,35].

Fases del Ciclo ENSO

ENSO consta de un ciclo que oscila entre la fase calida (fase El Niño), la fase neutra y la fase fría (La niña) [18]. El Niño y La Niña suelen durar entre 12 y 18 meses y ocurren en períodos de 2 a 7 años, aunque algunos eventos de El Niño y La Niña pueden extenderse más allá de los 24 meses [11].

Las fases de El Niño y La Niña comienzan a manifestarse como anomalías positivas (El Niño) o negativas (La Niña) de la temperatura superficial del mar (TSM) en el Pacífico central y oriental alrededor de julio. Estas anomalías se desarrollan a medida que avanza el ciclo ENSO, alcanzando un pico en el hemisferio norte aproximadamente entre enero y febrero del año siguiente. Posteriormente, las anomalías de la TSM disminuyen durante los meses subsiguientes de marzo a julio/agosto, y el evento de El Niño o La Niña concluye hacia el final del verano [45]. El cambio entre las fases de El Niño y La Niña no es inmediato ni sucesivo. En lugar de eso, estos eventos pueden ser interrumpidos por condiciones "neutralesçuando las TSM en el Pacífico oriental y central están cerca de los valores "normales".

1. **Fase El Niño:** Es la fase cálida de ENSO e históricamente, El Niño se refería a la aparición de aguas inusualmente cálidas frente a la costa de Perú, observadas alrededor de Navidad (de ahí el nombre El Niño, en español, el niño Cristo) [66]. Este fenómeno fue inicialmente identificado

por pescadores peruanos debido al impacto que tenía en la pesca local, lo que localmente le llaman El fenómeno del Niño.

Algunas caracteristicas de esta fase son:

- **Vientos Alisios y Zona de Convergencia Intertropical (ZCIT):** En condiciones normales, los vientos alisios soplan de este (América del Sur) a oeste (Indonesia), convergiendo hacia la ZCIT cerca del ecuador. Durante El Niño, estos vientos se debilitan o invierten su dirección, soplando de oeste a este [44, 66].
- **Anomalías en la Temperatura de la Superficie del Mar (TSM):** El debilitamiento de los vientos alisios reduce la surgencia de aguas frías, facilitando la aparición de anomalías positivas de TSM en el Pacífico oriental. Este calentamiento se manifiesta a través de una diferencia entre el valor observado de la TSM y la media climatológica del lugar [44, 66].
- **Presión Atmosférica y Oscilación del Sur (OS):** El Niño está asociado con una fase negativa de la OS, caracterizada por altas presiones en el oeste del Pacífico y bajas presiones en el este. La alternancia de la presión atmosférica entre el este y el oeste del Pacífico afecta la intensidad de los vientos alisios [44, 66].
- **Impacto en la Precipitación y la Circulación de Walker:** Durante El Niño, la región de la "poza cálida", en el oeste del Pacífico se desplaza hacia el centro y este, llevando la convección atmosférica asociada. Esto altera la circulación de Walker, modificando patrones de precipitación y causando lluvias intensas en el este del Pacífico [44].
- **Termoclina y Ondas Kelvin:** La termoclina, una capa de transición donde la temperatura del agua disminuye abruptamente con la profundidad, se profundiza en el este del Pacífico durante El Niño. Las ondas Kelvin, inducidas por fluctuaciones de los vientos alisios, se desplazan de oeste a este, propagando el calentamiento a lo largo del Pacífico ecuatorial [44, 67].

2. **Fase La Niña:** Es la fase distinguida por un enfriamiento anómalo de las aguas superficiales del océano Pacífico ecuatorial central y oriental [45].

   Algunas caracteristicas de esta fase son:

   - En primer lugar, tenemos la intensificacion del funcionamiento de la celda de Walker, reforzando los vientos alisios, lo cual lleva a la acumulación de aguas cálidas en el oeste del Pacífico y el fortalecimiento de la surgencia frente a las costas de Ecuador, Perú, norte de Chile [44].
   - Los vientos alisios que soplan hacia el oeste a lo largo del Pacífico tropical se intensifican. Esto favore el ascenso de aguas más frías hacia la superficie en el Pacífico ecuatorial. Esto resulta en anomalías negativas de la temperatura superficial del mar (TSM) y el nivel del mar [44,66].
   - La presión atmosférica es más baja de lo normal sobre Indonesia y el norte de Australia, y más alta de lo normal sobre el Pacífico tropical oriental (Índice de Oscilación del Sur positivo) [44,66].
   - El enfriamiento de las aguas en el Pacífico ecuatorial provoca un déficit de precipitaciones en el lado este del Pacífico y condiciones más secas alrededor del trópico y en las latitudes subtropicales de América del Sur, es decir Colombia, Ecuador, Perú y Chile. Al su vez, la lluvia es muy abundante en Indonesia, Malasia y el norte de Australia [11,44].

3. **Fase Neutra:** Segun [34] esta es la etapa en la cual las temperaturas de la superficie del mar en el Pacífico ecuatorial central y oriental están cercanas a sus promedios (+5°C o -5°C), es decir sin anomalías de calentamiento o enfriamiento.

   En especifico podemos decir:

   - La presión superficial en el Pacífico central (Centro y Sur América) es mayor que en el Pacífico ecuatorial occidental (como en Australia). Este gradiente de presión provoca un flujo de aire de oeste a este [34,44].

- La circulación atmosférica de bajo nivel influye directamente en las aguas superficiales. El movimiento del aire de este a oeste a lo largo del Pacífico ecuatorial genera corrientes superficiales en la misma dirección [34, 44].
- Los vientos de nivel bajo llevan las cálidas aguas del Pacífico occidental, lo cual provoca un ambiente muy húmedo, produciendo fuertes lluvias [34, 44].
- Posteriormente, el aire seco termina moviéndose hacia el este sobre el Pacífico oriental y descendiendo sobre las aguas más frías [34, 44].
- La termoclina, que es lavcapa que separa las aguas más cálidas de la superficie de las aguas más frías en las profundidades del Océano Pacífico, se inclina de este a oeste en condiciones neutrales [34, 44].

Regiones de Monitoreo ENSO

Para la identificación de las caracteristicas inusuales, se suele monitorear 4 zonas del oceano pacifico, como se puede observar en la fig. 2.1.

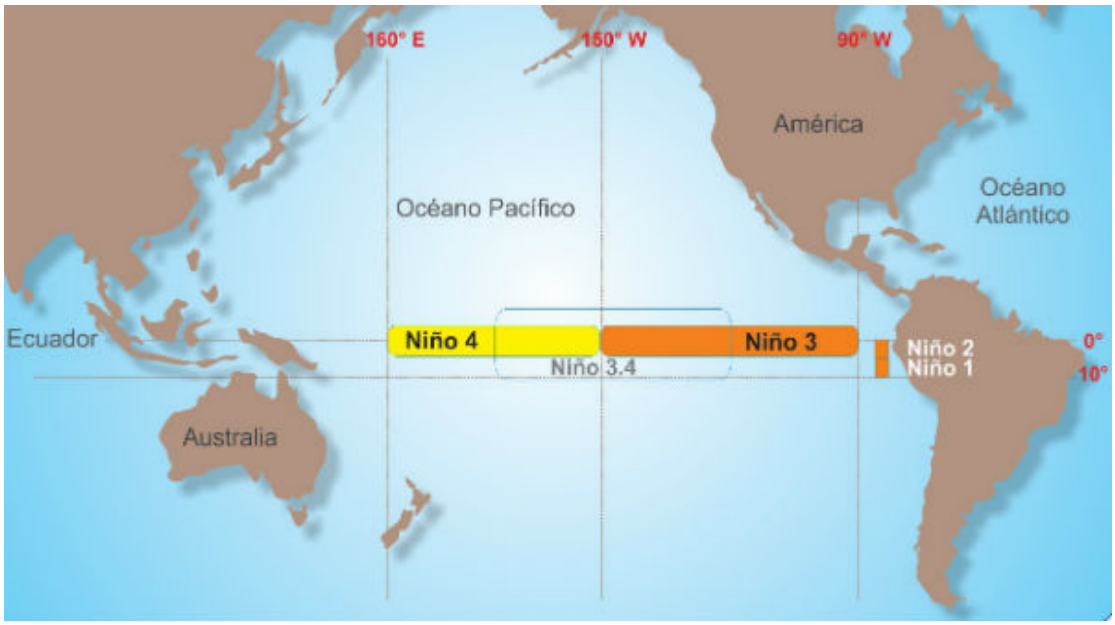

Figura 2.1: Regiones del ENSO. Imagen tomada del Ministerio del Ambiente [19].

Según la NOAA [48] las regiones de monitoreo del ciclo ENSO son:

- **Oceanicas:**

Niño 3 (5N-5S, 150W-90W): Esta región fue en anteriores años la principal área para monitorear ENOS, aunque posteriormente los investigadores se centraron más en otras regiones. Las anomalías de temperatura en esta área afectan el clima global, incluyendo patrones de precipitación y temperaturas en diversas regiones del mundo [67,74].

Niño 4 (5N-5S, 160E-150W): Región más occidental de las regiones Niño del pacifico ecuatorial central. Las anomalías en esta zona influyen en el clima de Asia y Australia [67,74].

Niño 3.4 (5N-5S, 170W-120W): Esta región es considerada el foco principal para monitorear y predecir ENOS, y partir de esta región se calcula varios índices climáticos de ENOS para su monitoreo. Con las anomalías registradas en esta área se calculan varios índices ENSO y las variaciones de temperatura en esta área tienen gran impacto en la climatología global [67,74].

- **Costeras:**
  Niño 1+2 (0-10S, 90W-80W): Esta región es la más oriental del pacifico ecuatorial y la más pequeña, y corresponde a la región costera de América del Sur. Fue en esta región que El Niño fue registrado por primera vez por pobladores locales de Perú. Esta region es importante para detectar el inicio de cualquier fase del ciclo ENSO, debido a que los cambios de temperatura suelen aparecer primero en esta área [67,74].

Indices

En el contexto de poder tener un indicador que pueda medir las anomalias respecto a las temperaturas del mar, diversos organismos internacionales y nacionales han desarrollado diferentes indicadores para aproximar si nos encontramos cerca a una de las fases del ciclo ENSO. Según el Instituto del Mar del Perú - IMARPE [20], se usan los siguientes índices:

1. **Índice Térmico Costero Peruano (ITCP):** El Índice Térmico Costero Peruano (ITCP) es una medida que refleja el impacto de El

Niño-Oscilación del Sur (ENOS) en la variabilidad térmica del océano costero peruano. El ITCP se estimó usando los registros desde 1982 hasta 2014 de la temperatura superficial del mar obtenidos de la NOAA NCDC OISST. Podemos observar como se hace la catalogación en la tabla 2.1.

Condición	**Categoría**	**ΔT máx.**	**ΔT min.**
El Niño	Cálido		>0,4
Neutro		0,4	-0,6
La Niña	Fría	<-0,6	

Tabla 2.1: Categorías del Índice Térmico Costero Peruano (ITCP) [20].

2. **Índice Costero El Niño (ICEN):** El Índice Costero El Niño (ICEN) fue desarrollado por la Comisión Multisectorial encargada del Estudio del Fenómeno El Niño (ENFEN) para evaluar la presencia de El Niño y La Niña en Perú. Este índice se calcula como la media móvil de tres meses de la anomalía de la temperatura superficial del mar en la región Niño 1+2 (90°-80°W, 10°S-0°) basada en los datos NOAA ERSST en el período 1981-2010.Podemos observar como se hace la catalogación en la tabla 2.2.

Condición	**Categoría**	**ΔT máx.**	**ΔT min.**
La Niña	Fuerte		<-1.4
	Moderado	≥-1.4	<-1.2
	Débil	≥-1.2	<-1.0
Neutro		≥-1.0	≤0.4
El Niño	Débil	>0.4	≤1.0
	Moderado	>1.0	≤1.7
	Fuerte	>1.7	≤3.0
	Muy Fuerte	>3.0	

Tabla 2.2: Categorías del Índice Costero El Niño (ICEN).

Otro índice desarrollado por el NOAA es Oceanic Niño Index (ONI), el cual es definido como un mínimo de cinco promedios consecutivos de 3 meses consecutivos de anomalías de la TSM (Según el monitoreo NOAA ERSST.v5) en la región del Niño 3.4 que superan un umbral de +/- 0.5°C.Podemos observar como se hace la catalogación en la tabla 2.3.

Condición	Categoría	ΔT máx.	ΔT min.
El Niño	Fuerte		≥+0.5
	Moderado	≥+0.5	<+1.0
	Débil	≥+1.0	<+1.5
Neutro		≥-0.5	≤+0.5
La Niña	Débil	>-0.5	≤-1.0
	Moderado	>-1.0	≤-1.5
	Fuerte	>-1.5	

Tabla 2.3: Categorías del Índice de Oscilación del Sur (ONI).

2.1.2. Machine Learning

Según IBM [29], el machine learning es una disciplina de la Inteligencia Artificial que se centra en el uso de datos y algoritmos para imitar la forma en la que aprenden los humanos, mejorando gradualmente su precisión. Este campo es altamente interdisciplinario que toma ideas de la optimización, la teoría de la información, la estadística y las matemáticas.

Tipos de Machine Learning

Existen diversos tipos de modelos de aprendizaje automático, clasificados según el grado de intervención humana en los datos sin procesar. Esto puede incluir la provisión de recompensas, la entrega de retroalimentación específica o la utilización de etiquetas para guiar el proceso de aprendizaje [29, 68].

1. **Aprendizaje Supervisado**
   se basa en el uso de conjuntos de datos etiquetados para entrenar algoritmos a fin de clasificar datos o predecir resultados con precisión. Durante este proceso, se proporcionan al modelo tanto las entradas como las salidas deseadas, lo que permite ajustar sus ponderaciones hasta que se han optimizado adecuadamente. Esto se realiza como parte de la validación cruzada para evitar el sobreajuste o el subajuste del modelo. Al entrenarse con datos categorizados, los algoritmos de aprendizaje supervisado pueden generalizar y hacer predicciones precisas sobre datos nuevos y no vistos. Esta técnica es esencial para resolver problemas del mundo real, como clasificar correos no deseados en carpetas separadas de la bandeja de entrada. Entre los modelos utilizados en el aprendizaje

supervisado se encuentran las redes neuronales, la regresión lineal, la regresión logística, los bosques aleatorios y las máquinas de vectores de soporte (SVM) [29,37,68].

2. **Aprendizaje No Supervizado**
   Este tipo de apredizaje no requiere datos de entrenamiento etiquetados y se basa únicamente en las entradas proporcionadas sin objetivos específicos. Utiliza algoritmos para analizar y agrupar conjuntos de datos no etiquetados, descubriendo patrones ocultos o agrupaciones de datos sin intervención humana. Esta capacidad es ideal para el análisis exploratorio de datos, estrategias de venta cruzada, segmentación de clientes, y reconocimiento de imágenes y patrones. Además, el aprendizaje no supervisado puede reducir el número de entidades de un modelo a través de la reducción de la dimensionalidad, utilizando enfoques como el análisis de componentes principales (PCA) y la descomposición de valores singulares (SVD). El principal objetivo es determinar patrones o agrupaciones de datos, organizándolos por similitud. Por ejemplo, la agrupación de imágenes es una técnica de aprendizaje no supervisado que agrupa imágenes de manera que la similitud entre diferentes grupos se minimiza. Los algoritmos como k-Means son frecuentemente utilizados en este contexto [29,37,68].

3. **Aprendizaje por Refuerzo**
   El aprendizaje automático por refuerzo es un modelo de aprendizaje que, a diferencia del aprendizaje supervisado, no se entrena con datos de ejemplo predefinidos,sino que aprende sobre la marcha mediante el método de prueba y error, reforzando secuencias de resultados exitosos para desarrollar la mejor recomendación o política para un problema determinado. Este enfoque permite al algoritmo mapear acciones a situaciones con el objetivo de maximizar la recompensa o la retroalimentación recibida [29,54].

Estos enfoques determinan cómo el modelo interpreta y aprende de los datos, influenciando su capacidad para tomar decisiones precisas y mejorar su desempeño con el tiempo. Además, la elección del tipo de modelo puede

depender del contexto específico y de los objetivos que se deseen alcanzar en el proyecto de análisis de datos.

Modelos

1. **Árbol de Desición** Un árbol de decisión es un modelo de aprendizaje supervisado no paramétrico que se utiliza tanto para tareas de clasificación como de regresión. Este modelo se estructura jerárquicamente en forma de árbol, compuesto por un nodo raíz, ramas, nodos internos y nodos hoja. Cada nodo interno representa una decisión basada en una característica del conjunto de datos, mientras que cada nodo hoja representa el resultado final o la predicción. Los árboles de decisión dividen el espacio de instancias de manera recursiva en subespacios más homogéneos, facilitando la interpretación y visualización del proceso de toma de decisiones [23, 26, 56].

   - **Estructura** Según [23, 57], los Árboles de desición tienen la siguiente estructura:

     a) **Raíz:** El nodo superior del árbol, que representa la característica más importante para la predicción.

     b) **Nodos Internos**: Representan decisiones basadas en características específicas de los datos.

     c) **Hojas:** Los nodos terminales que proporcionan la salida final del árbol, ya sea una clase o un valor continuo

   - **Construcción**

     La construcción de un árbol de decisión se realiza mediante algoritmos llamados ïnductores de árboles de decisión", que generan automáticamente un árbol a partir de un conjunto de datos dado. El objetivo principal es encontrar el árbol de decisión óptimo minimizando el error de generalización, aunque también se pueden definir otras funciones objetivo, como minimizar el número de nodos o la profundidad media del árbol [57].

- **Inducción de un árbol de decisión óptimo**

  Esta es una tarea compleja y se ha demostrado que encontrar un árbol de decisión mínimo consistente con el conjunto de entrenamiento es un problema NP-completo. Debido a esto, los métodos heurísticos son necesarios para resolver el problema de manera eficiente. Estos métodos generalmente se dividen en dos grupos: de arriba hacia abajo (top-down) y de abajo hacia arriba (bottom-up), siendo los primeros los más utilizados [55].Los algoritmos de inducción de árboles de decisión de arriba hacia abajo, como ID3, C4.5 y CART, suelen consistir en dos fases conceptuales: crecimiento y poda. Estos algoritmos son de naturaleza voraz y construyen el árbol de decisión de manera recursiva, dividiendo el conjunto de entrenamiento en cada iteración mediante una función de partición basada en los atributos de entrada [26].

  Algunas metricas que se usa para la division son:
  - Impureza de Gini: Mide la probabilidad de una clasificación incorrecta de una nueva instancia si se clasificó aleatoriamente de acuerdo con la distribución de clases en el conjunto de datos [23, 26, 57], En la ecuación 2.1 podemos apreciar como se expresa matemáticamente.

$$\text{Impureza de Gini}(D) = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.1}$$

  - Entropía: mide la cantidad de incertidumbre o impureza en el conjunto de datos [23, 26, 57], En la ecuación 2.2 podemos apreciar como se expresa matemáticamente.

$$\text{Entropía}(D) = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{2.2}$$

2. **Bagging**

   El método de bootstrap propagation, conocido como bagging y desarrollado por Leo Breiman en 1994, implica dividir el conjunto de entrenamiento mediante muestreo aleatorio para producir subconjuntos distintos. Este enfoque introduce diversidad al crear copias de arranque de los datos de entrada, seleccionando múltiples subconjuntos aleatorios del conjunto de datos original y reemplazándolos. Estos subconjuntos se utilizan para entrenar múltiples modelos base del mismo algoritmo de aprendizaje automático, y luego se combinan mediante votación mayoritaria para obtener un clasificador robusto. Esta técnica es ampliamente utilizada en bosques aleatorios y mejora el rendimiento de los modelos de aprendizaje automático al combinar predicciones de conjuntos de entrenamiento generados aleatoriamente. El principalm argumento de creación de este método es que la perturbación del conjunto de aprendizaje puede llevar a cambios significativos en los predictores, mejorando la precisión del modelo [1,21,30].

   Proceso de funcionamiento:

   De acuerdo con ibm [1], el bagging consta de tres pasos esenciales:

   Primero, mediante el bootstrapping, el algoritmo genera múltiples muestras del conjunto de datos utilizando un método de remuestreo bootstrap que selecciona datos al azar con reemplazo. Esto crea subconjuntos diversos donde ciertas instancias pueden repetirse en diferentes muestras.

   Después, cada una de estas muestras bootstrap se entrena de manera independiente y simultánea utilizando modelos base o débiles (entrenamiento en paralelo).

   Finalmente, las predicciones de estos modelos se combinan mediante votación (ya sea promediando las predicciones en problemas de regresión o seleccionando la clase con mayoría de votos en problemas de clasificación), lo cual mejora la precisión general del modelo al reducir la varianza y mejorar la generalización.

3. **Boosting**

Es un método de aprendizaje en conjunto que combina varios modelos débiles en un modelo robusto para minimizar errores de entrenamiento y mejorar la precisión en predicciones. Este enfoque implica seleccionar una muestra aleatoria de datos, entrenar modelos secuenciales donde cada uno compensa las debilidades del anterior, y combinar las reglas débiles de cada clasificador individual para formar una regla de predicción fuerte [2,3].

- **Funcionamiento** Según lo mencionado en [10], el boosting tiene el siguiente funcionamiento:
  - Primero, en la Inicialización, se asigna un peso inicial a cada instancia en el conjunto de entrenamiento, generalmente iguales al inicio para todas las instancias.
  - En el Entrenamiento del primer modelo, se entrena un modelo inicial con los datos de entrenamiento, el cual cometerá algunas predicciones correctas e incorrectas.
  - Luego, se procede al Cálculo de errores, donde se evalúa el rendimiento del modelo inicial en función de los pesos asignados a las instancias. Las instancias mal clasificadas por este modelo recibirán un mayor peso, resaltando su importancia en el siguiente paso.
  - En el Entrenamiento del segundo modelo, se entrena un nuevo modelo que se enfoca más en las instancias con mayores pesos, es decir, aquellas que el modelo anterior clasificó erróneamente. Este proceso se repite en una Iteración definida, donde se ajustan los pesos y se entrenan modelos sucesivos para corregir los errores acumulados por los modelos anteriores.
  - Tras completar las iteraciones, los modelos se combinan mediante una Combinación de los modelos. Esto se logra a través de una suma ponderada de sus predicciones, donde los modelos que han demostrado mejor rendimiento (menos errores en sus predicciones) suelen tener mayor influencia en el modelo final.

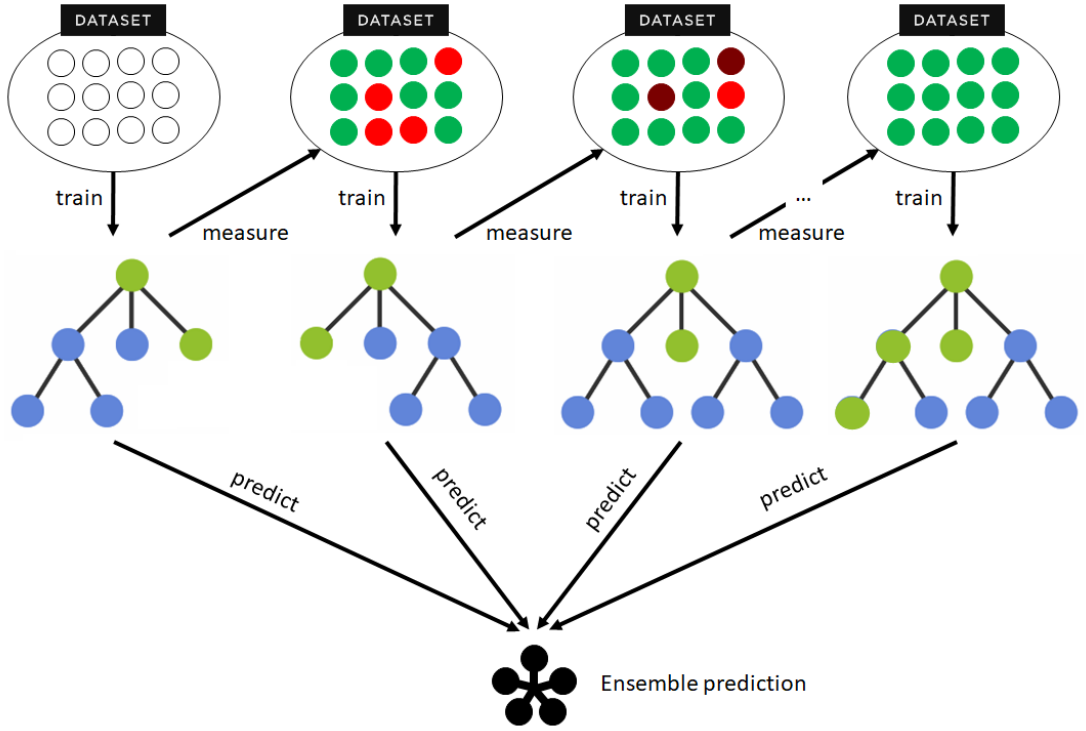

Figura 2.2: Secuencia del Modelo Boosting [10].

- **XGboost** es una versión de gradient boosting optimizada para ser rápida y escalable en términos computacionales. Este algoritmo utiliza eficientemente varios núcleos de la CPU, lo que posibilita el entrenamiento en paralelo durante el proceso de aprendizaje [3]

  - **Funcionamiento**

    Según lo mencionado en [13, 22] y en [15] el funcionamiento de XGBoost se empieza con la creación de un modelo inicial, la cual que realiza predicciones sobre un conjunto de datos. Luego los residuos (diferencias entre las predicciones iniciales y los valores observados) se calculan y se utilizan para entrenar un árbol de decisión. Este árbol se ajusta mediante una función de pérdida específica que minimiza errores y controla la complejidad de los árboles, permitiendo la poda precoz y la regulación de parámetros como la profundidad y la tasa de aprendizaje.

    Durante el proceso de entrenamiento, XGBoost implementa técnicas avanzadas para mejorar la eficiencia y la precisión del modelo. Utiliza métodos de regularización para reducir el

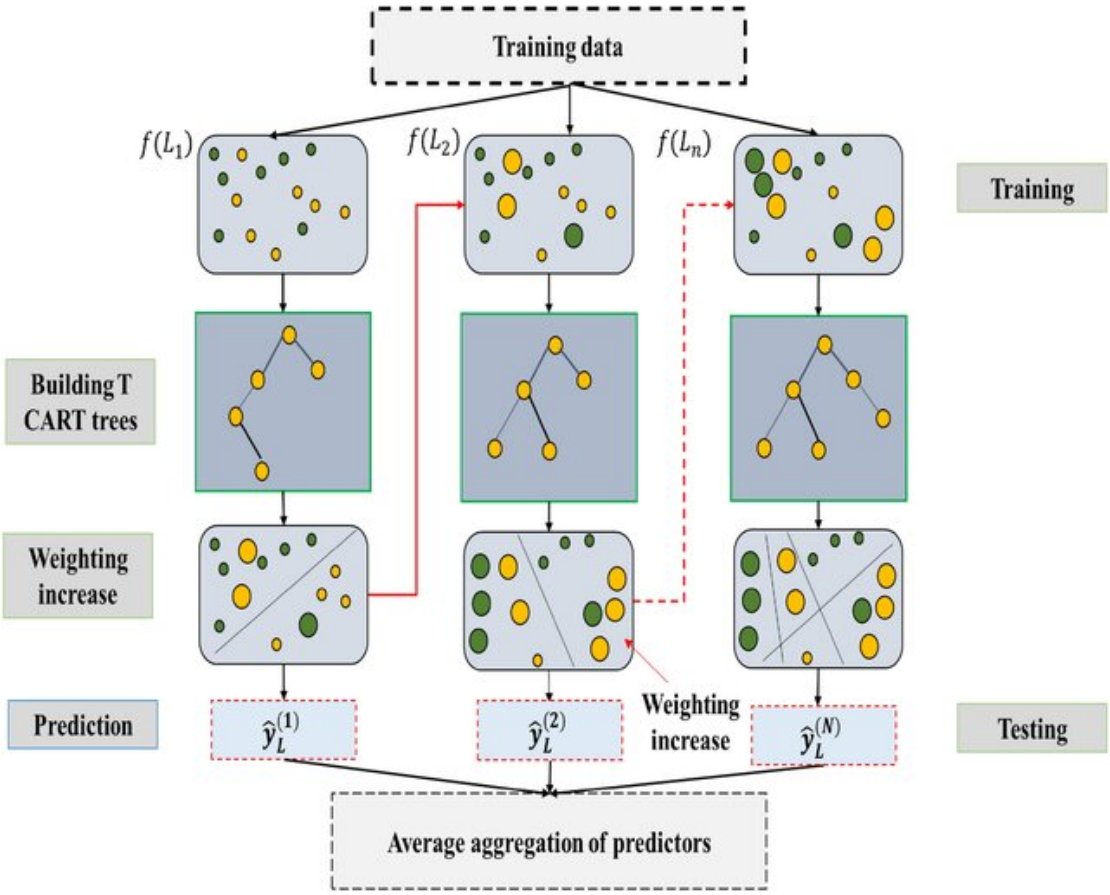

Figura 2.3: Secuencia del Modelo XGBoost [69].

sobreajuste y optimiza la búsqueda de divisiones mediante el uso de estructuras de datos comprimidas que minimizan la necesidad de ordenar datos repetidamente. Además, emplea técnicas de aleatorización como la selección aleatoria de submuestras y la reducción de columnas en niveles de árboles para mejorar la generalización y el rendimiento del entrenamiento.

XGBoost también se beneficia de un manejo eficiente de la escasez de datos, probando múltiples direcciones de división y utilizando algoritmos de boceto para la selección de cuantiles ponderados. Esto optimiza el proceso de aprendizaje paralelo, dividendo los datos en bloques que se pueden procesar simultáneamente, y gestionando el acceso sensible al caché para mejorar la velocidad de cálculo y reducir el tiempo de entrenamiento en conjuntos de datos grandes que pueden no

caber en la memoria principal.

4. **Random Forest**

   Un bosque aleatorio es una técnica de clasificación desarrollada por Breiman como una extensión del bagging. Es una de las técnicas más confiables, utilizando clasificadores base como árboles de decisión [55], siendo el bagging el método más adecuado para crear Random Forest, debido a que los árboles de decisión tienen baja varianza y alta bias. Al clasificar una instancia, las predicciones de los clasificadores base se combinan mediante el voto mayoritario. Los Random Forest tienen varias ventajas sobre otras técnicas de ensamble. No requieren el uso de conjuntos de prueba para la construcción, ya que utilizan instancias fuera de la bolsa (out of bag) para entrenar el clasificador, lo que permite estimar la capacidad de generalización internamente. Además, los Random Forest incluyen métodos para manejar datos faltantes y desequilibrados y son menos sensibles al ruido en los datos en comparación con otros métodos de ensamble [58].

5. **Árbol de Regresión**

   Los árboles de regresión son un tipo de árbol de decisión donde las variables objetivo pueden tomar valores continuos en lugar de etiquetas de clase en las hojas. A diferencia de los árboles de clasificación, los árboles de regresión utilizan criterios de selección de división y criterios de detención modificados para manejar valores continuos [21].

   - Funcionamiento y Estructura: Los árboles de regresión dividen los datos en subconjuntos utilizando nodos, ramas y hojas. Cada nodo interior representa una pregunta sobre un atributo, y las ramas representan las posibles respuestas, conduciendo a nodos hijos o hojas terminales. Estas hojas contienen el valor predicho, que generalmente es el promedio de los valores de la variable objetivo para los datos en esa hoja [28, 63].

     El proceso de particionamiento se conoce como particionamiento recursivo, donde el espacio de datos se subdivide repetidamente

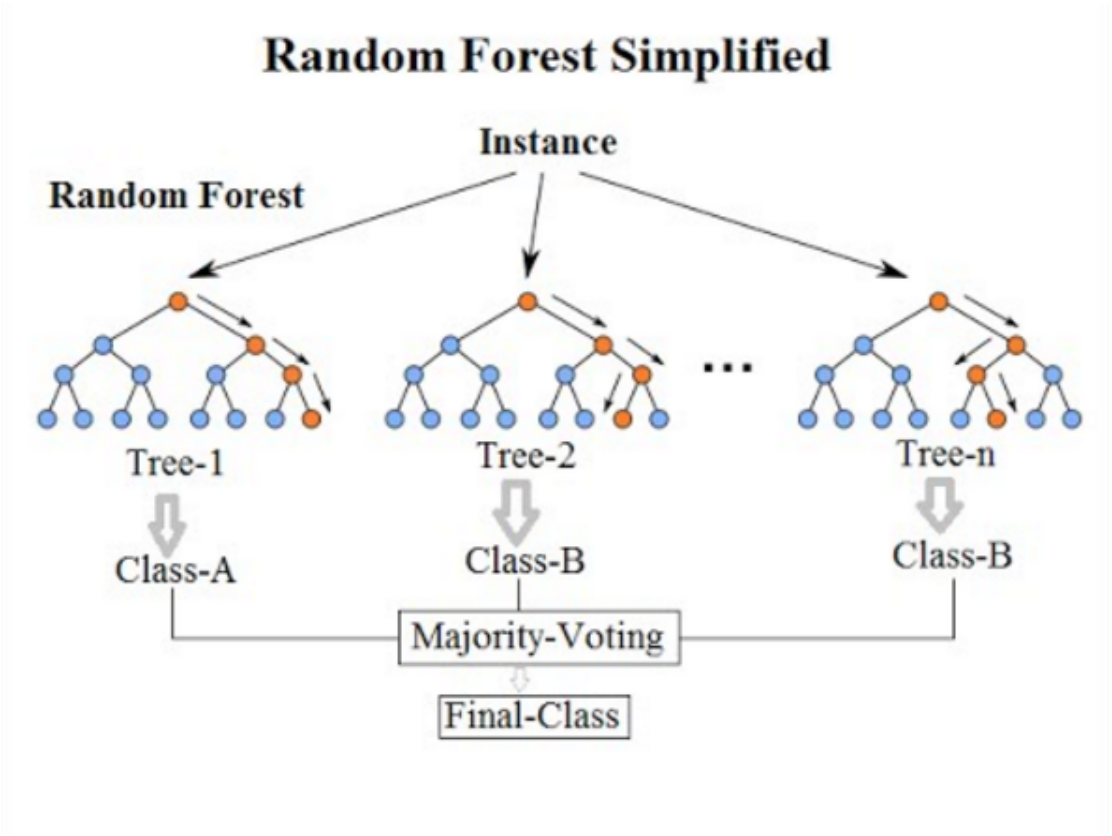

Figura 2.4: Secuencia del Modelo Random Forest [42].

hasta que las subdivisiones son lo suficientemente pequeñas y manejables para ajustarse a modelos simples . En cada paso, se selecciona la división que minimiza la varianza dentro de los nodos hijos, lo que lleva a una mejor predicción de la variable objetivo en cada partición.

- Ventajas y Uso: La principal ventaja de los árboles de regresión es su legibilidad. Los árboles de regresión no solo predicen valores, sino que también explican qué atributos se utilizan y cómo se utilizan para llegar a las predicciones. Esto es especialmente útil para identificar posibles eventos y ver los resultados potenciales de decisiones .

  Otra ventaja es que, debido a que las predicciones se basan en promedios simples en las hojas, los árboles de regresión pueden manejar situaciones donde la superficie de regresión verdadera no es suave, permitiendo una respuesta escalonada que puede aproximarse arbitrariamente cerca a la superficie real con suficientes

hojas [28, 63].

- Poda y Validación Cruzada: Un problema común con los árboles de regresión es que pueden crecer demasiado y sobreajustarse a los datos de entrenamiento. Para abordar esto, se utiliza la validación cruzada para podar el árbol. Esto implica dividir los datos en conjuntos de entrenamiento y prueba, construir el árbol completo con el conjunto de entrenamiento y luego podar las hojas que no mejoran la predicción en el conjunto de prueba [28, 63].

6. **Redes Neuronales**

   Según IBM [27], una Red Neuronal Artificial es un modelo de Machine Learning que toma decisiones de manera similar al cerebro humano, mediante el uso de procesos que imitan la forma en que las neuronas biológicas trabajan juntas para identificar fenómenos y llegar a conclusiones.

   Las redes neuronales se basan en datos de entrenamiento para aprender y mejorar su precisión progresivamente. Esta precisión ajustada las convierte en una herramienta potente en informática e inteligencia artificial, permitiendo una clasificación y agrupación de datos rápida. A diferencia del reconocimiento manual realizado por expertos humanos, las tareas de reconocimiento de voz o imágenes pueden realizarse en minutos en lugar de horas.

   a) **Elementos**

   - Neurona o Nodo: Los nodos, también conocido como neurona o unidad, es una unidad básica de procesamiento en una red neuronal [27,75]. Un nodo recibe una o más entradas, las procesa y produce una salida.

     - **Entrada:** Los valores que el nodo recibe de otros nodos o del entorno externo.

- **Salida:** El valor procesado que el nodo envía a otros nodos o a la salida final de la red.

- **Función de Activación:** Una función que transforma la entrada del nodo en su salida.

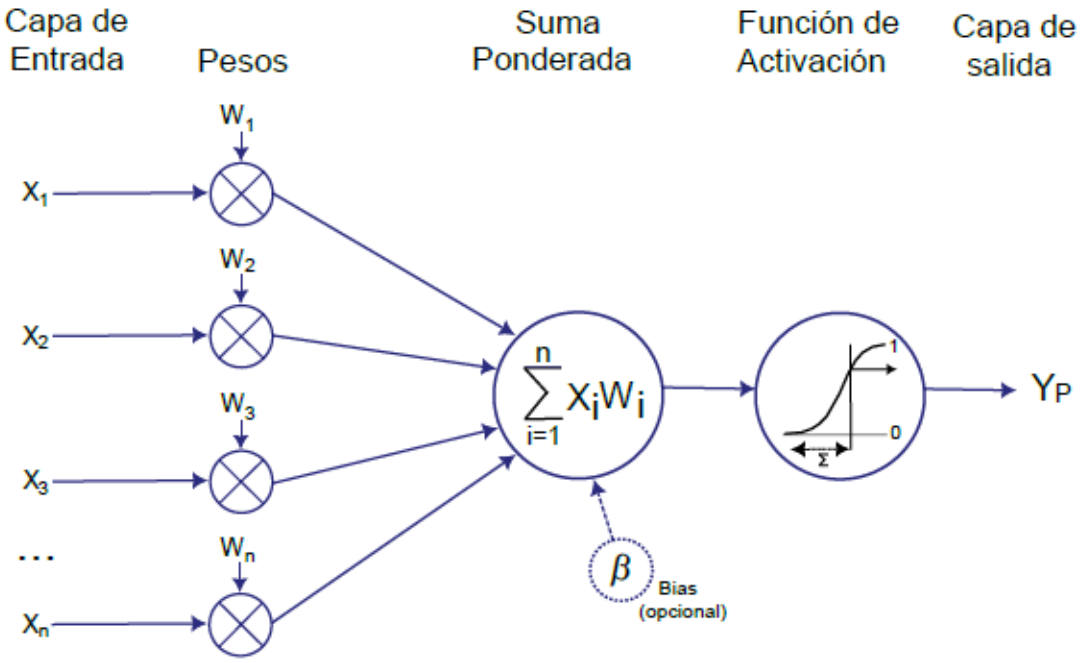

Figura 2.5: Arquitectura de una Neuronna [67].

- Capas: Las Redes Neuronales están organizadas en capas de nodos y según s [12] estas capas puede ser de los siguientes tipo:
  1) Capa de entrada: La información del mundo exterior entra en la red neuronal artificial desde la capa de entrada. Los nodos de entrada procesan los datos, los analizan o los clasifican y los pasan a la siguiente capa [12].
  2) Capa oculta: Las capas ocultas toman su entrada de la capa de entrada o de otras capas ocultas. Las redes neuronales artificiales pueden tener una gran cantidad de capas ocultas. Cada capa oculta analiza la salida de la capa anterior, la procesa aún más y la pasa a la siguiente capa. De acuerdo a lo apreciado en [25] existen diversos tipos de capas ocultas:
     - Capas densas (Redes Neuronales Profundas) Fig. 2.6: también conocidas como capas totalmente conectadas,

han sido un componente fundamental en las redes neuronales desde sus inicios. Una capa densa realiza una transformación lineal de su entrada, seguida de una función de activación [25].

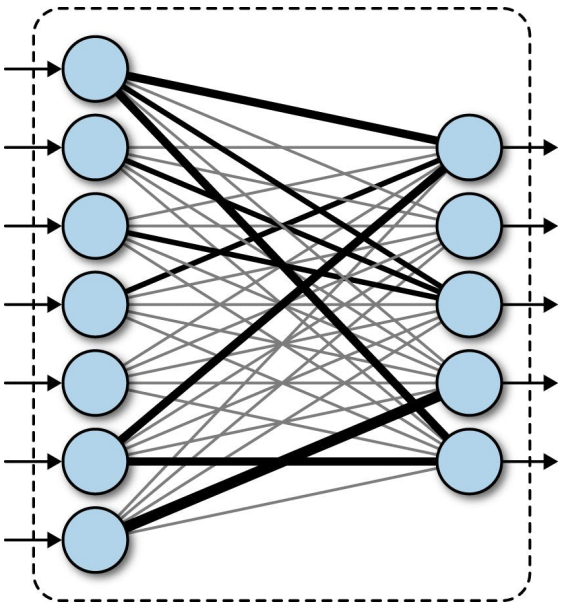

Figura 2.6: Red Neuronal con Capa Densa [25].

- Capas convolucionales (Redes neuronales Convolucionales): introducidas por Yann LeCun en la década de 1980 con la arquitectura LeNet-5, son elementos fundamentales en las redes neuronales convolucionales (CNN) para el reconocimiento de dígitos escritos a mano. Estas capas están diseñadas para aprovechar la estructura espacial de las imágenes, convirtiéndose en el estándar para el procesamiento de imágenes y las tareas de visión por computadora. En una capa convolucional, un filtro o kernel se desliza sobre la imagen de entrada, realizando una operación de convolución que calcula el producto escalar entre el kernel y la región de entrada. Esto produce un mapa de características que representa patrones espaciales en la entrada, En la fig. 2.7 podemos apreciar un poco de su arquitectura.

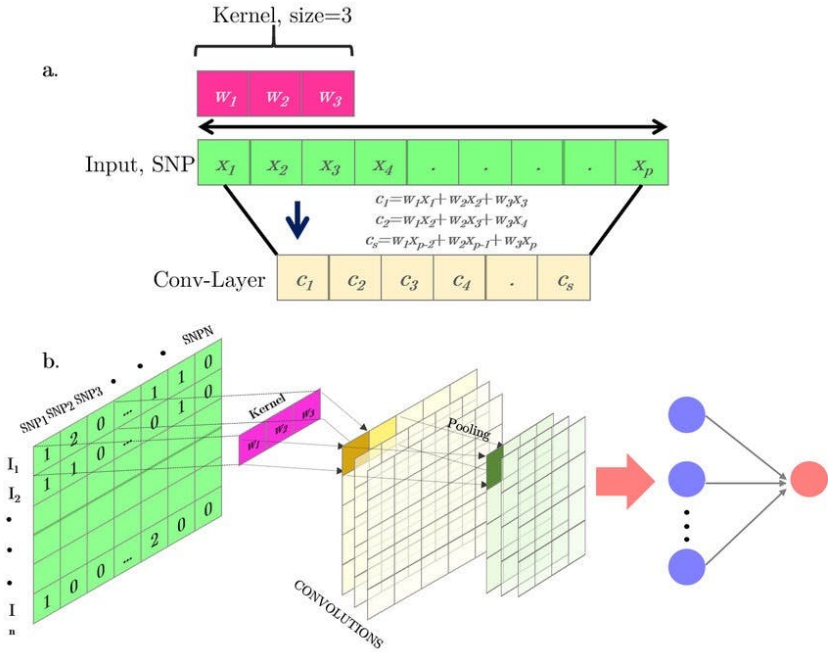

Figura 2.7: Red Neuronal Convolucional [25].

- Capas Recurrentes (Redes Neuronales Recurrentes): surgieron en la década de 1980 y son componentes esenciales en este tipo de redes. Las RNN están diseñadas específicamente para procesar datos secuenciales, siendo particularmente útiles en aplicaciones como el procesamiento del lenguaje natural, la predicción de series temporales y el aprendizaje de secuencias. Un hito notable en este campo es la Red de Elman desarrollada por Jeff Elman. Estas capas recurrentes mantienen un estado oculto interno que se actualiza a lo largo del tiempo, lo que les permite capturar y recordar información de pasos temporales anteriores [25].

  Según [43, 46] existen varios tipos de capas recurrentes, entre las más conocidas tenemos:
  - **Neuronas Recurrentes Básicas** Las neuronas recurrentes básicas son las unidades fundamentales en una RNN estándar. Procesan secuencias de datos manteniendo un estado oculto que se actualiza en cada paso de tiempo.

Estas redes son simples y pueden modelar secuencias de datos, pero tienen dificultades para capturar dependencias a largo plazo debido al problema de desvanecimiento del gradiente.Fig 2.8

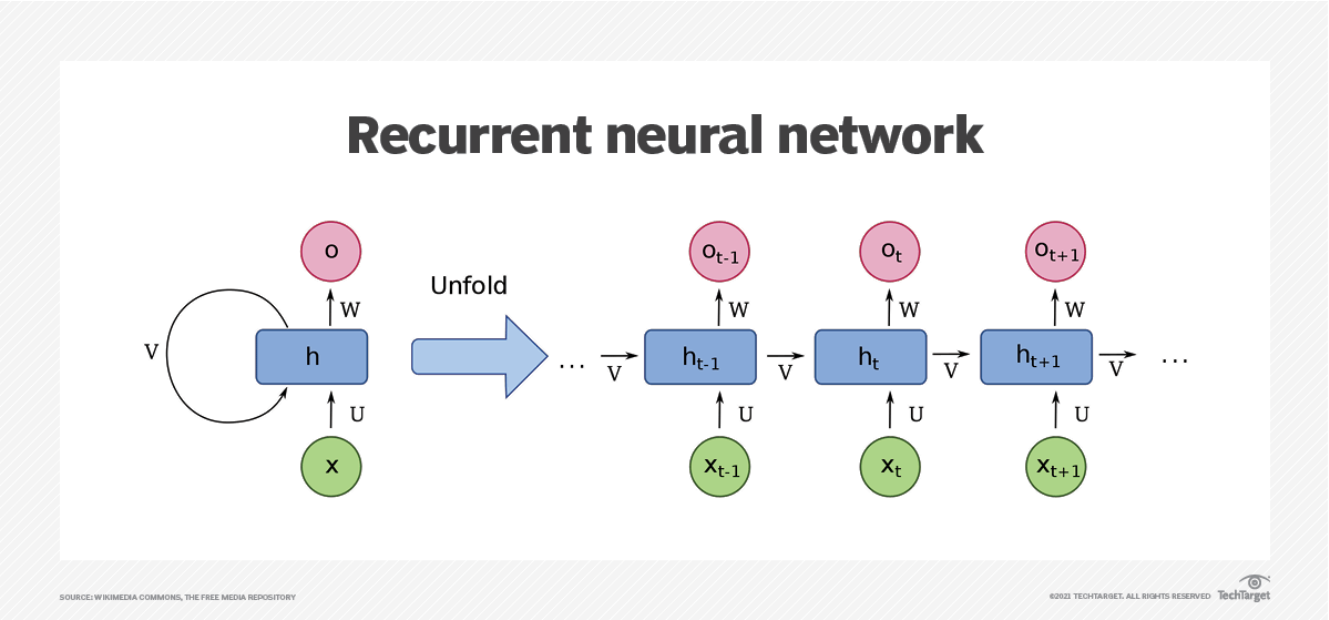

Figura 2.8: Red Neuronal Recurrente Básica [43].

- **Long Short-Term Memory (LSTM)** Introducidas por Hochreiter y Schmidhuber en 1997, las LSTM están diseñadas para superar las limitaciones de las RNN básicas. Utilizan una estructura de tres puertas (puerta de entrada, puerta de olvido y puerta de salida) para controlar el flujo de información, permitiendo mantener y actualizar información durante largos períodos. Esto las hace efectivas para tareas como el procesamiento del lenguaje natural y la predicción de series temporales, aunque su complejidad computacional es mayor [25, 46]. Esto se puede apreciar en la Fig.2.9.

- **Gated Recurrent Unit (GRU)** Las GRU, introducidas por Cho et al. en 2014, son una variante simplificada de las LSTM. Como se puede observar en la fig 2.10 utilizan dos puertas (puerta de actualización y puerta de reinicio) para controlar el flujo de información. Ofrecen un rendimiento

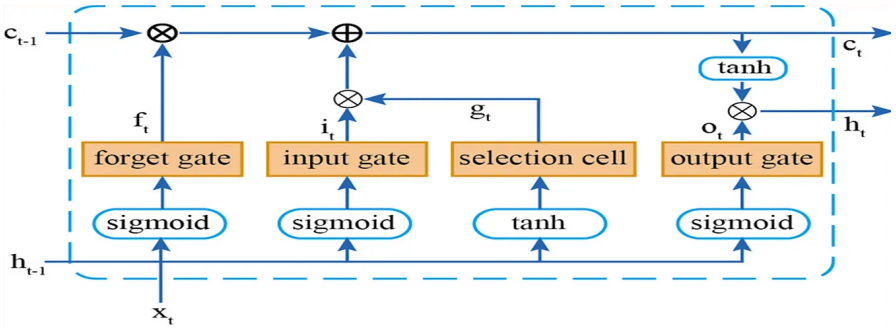

Figura 2.9: Neurona artificial LSTM [46]

similar a las LSTM pero con una estructura más simple y menos parámetros, lo que reduce la complejidad computacional. Son adecuadas para aplicaciones donde se necesita un equilibrio entre rendimiento y eficiencia [25].

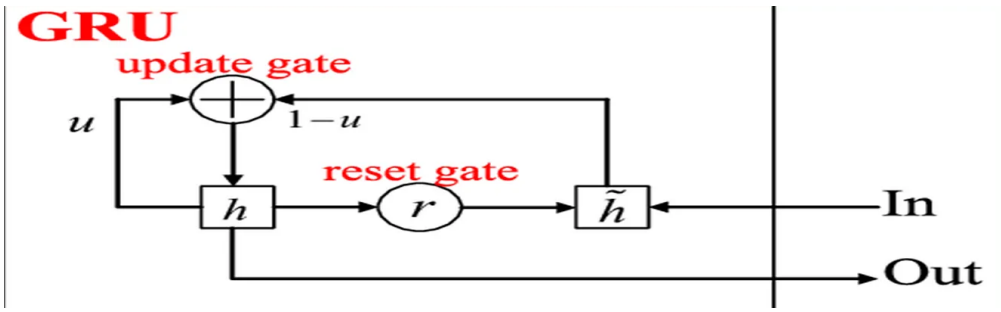

Figura 2.10: Neurona artificial GRU [46].

3) Capa de salida: La capa de salida proporciona el resultado final de todo el procesamiento de datos que realiza la red neuronal artificial. Puede tener uno o varios nodos [12].

- Función de Activación

  La función de activación es la parte más importante de una neurona artificial y determina si la neurona debe activarse. Procesa la salida de la neurona (después de calcular la suma ponderada de las entradas) y convierte esta salida en un valor que puede ser utilizado como entrada para la siguiente capa de la red [46]. En la fig 2.11 podemos apreciar las funciones de

activación más comunes.

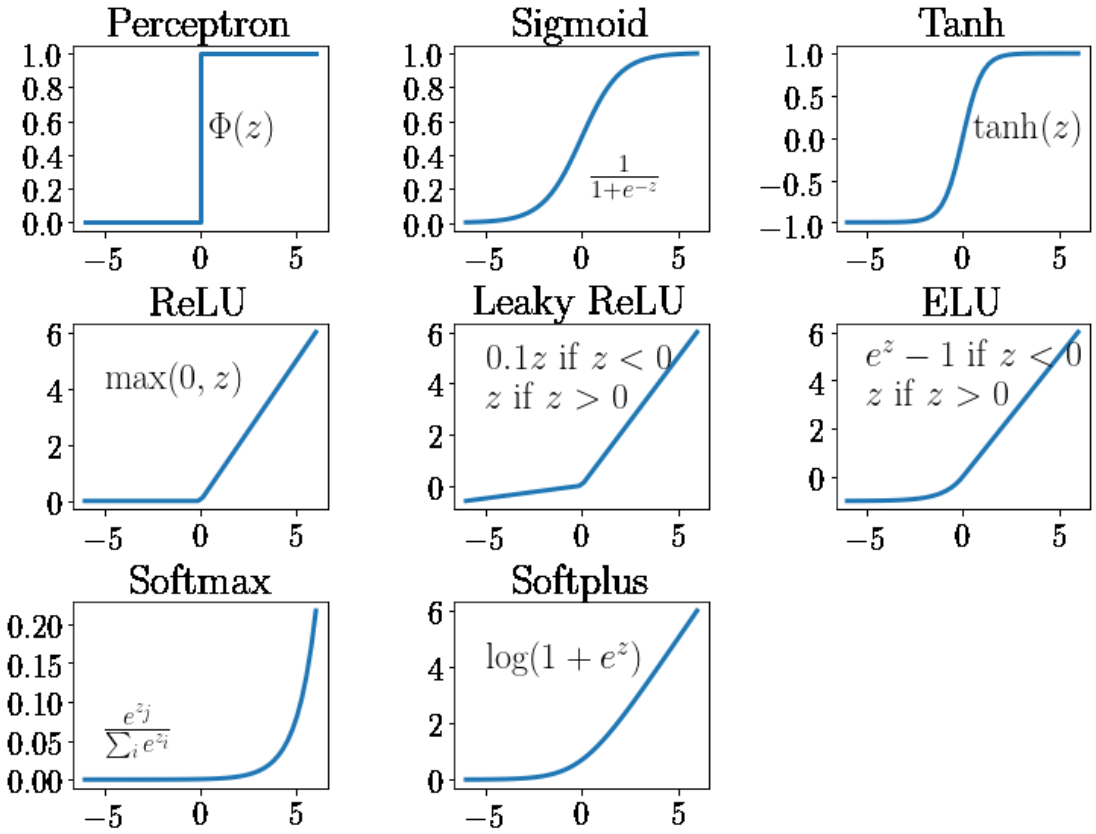

Figura 2.11: Funciones de Activación más comunes [46].

- Weights y Biases

  Los pesos y Biases son parámetros de las redes neuronales que facilitan la identificación de datos en el aprendizaje automático. Los pesos y Biases determinan cómo una red neuronal impulsa el flujo de datos hacia adelante a través de la red, proceso conocido como propagación hacia adelante. Una vez completada esta propagación, la red neuronal ajusta las conexiones utilizando los errores surgidos durante la misma. Luego, el flujo se invierte para recorrer las capas y ajustar los nodos y conexiones necesarios, en un proceso llamado retropropagación [24].

b) **Funcionamiento**

Las redes neuronales artificiales se pueden conceptualizar como gráficos dirigidos ponderados organizados en capas, con nodos

que imitan a las neuronas biológicas y funciones de activación. La primera capa recibe las señales de entrada del entorno, similar a los nervios ópticos en la visión humana. Cada capa posterior recibe la salida de la anterior, y la última capa genera la salida final del sistema. Estas redes son modelos matemáticos que aprenden y han mejorado significativamente las tecnologías de análisis de datos [12,27,67].

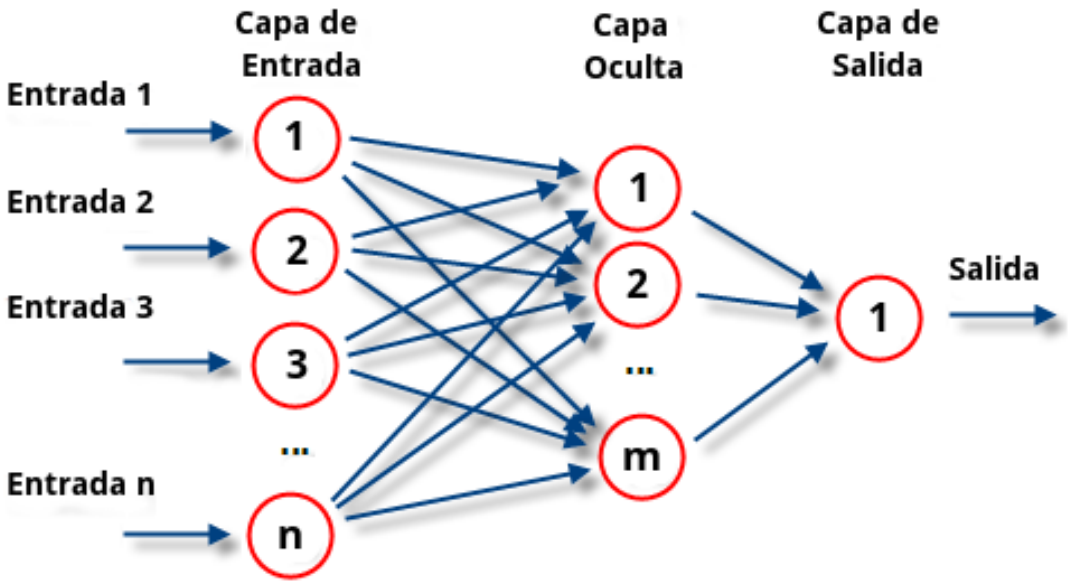

Figura 2.12: Arquitectura de una Red Neuronal [31].

2.2. Revisión de Literatura

Históricamente, las comunidades costeras y los pescadores sudamericanos han observado cambios en las corrientes oceánicas y la temperatura del agua antes de que la ciencia moderna desarrollara herramientas de monitoreo precisas. El conocimiento tradicional se basa en observaciones directas de fenómenos naturales como la temperatura del agua de mar, el comportamiento de las especies marinas y las condiciones climáticas locales [47].

2.2.1. Métodos de Predicción

Los métodos de pronóstico del fenómeno de El Niño y otros eventos climáticos han dependido en gran medida de enfoques estadísticos y dinámicos. A continuación, se describen algunos de los métodos más comunes utilizados:

1. **Metodos Estadísticos** De acuerdo con [17, 72] existen varios métodos o modelos estadísticos que se han utilizando y se siguen utilizando hasta la actualidad, entre los más conocidos tenemos:

   - **Modelos de Series Temporales:** Utilizan datos históricos para identificar patrones y tendencias que puedan predecir eventos futuros.Por ejemplo, Ejemplos: Análisis de regresión, modelos autoregresivos (AR), modelos de media móvil (MA), y combinaciones como los modelos ARIMA (Autoregressive Integrated Moving Average).
   - **Métodos de Análisis de Correlación:** Examina las relaciones estadísticas entre diferentes variables climáticas, como la temperatura de la superficie del mar y la presión atmosférica. Por ejemplo, uso del Índice de Oscilación del Sur (SOI), que mide la diferencia de presión entre Tahití y Darwin (Australia), como indicador de las condiciones de El Niño y La Niña.

2. **Métodos Dinámicos** se centran en capturar la compleja interacción entre el océano y la atmósfera que caracteriza al Fenómeno El Niño-Oscilación del Sur (ENSO). A través de la integración de datos observacionales y la aplicación de principios físicos y matemáticos, estos modelos buscan proyectar cómo evolucionará el ENSO en el futuro, proporcionando así herramientas cruciales para la planificación y la mitigación de riesgos climáticos a escala global [17], entre los más conocidos tenemos:

   - **Modelos de Circulación General (GCMs):** Simulan la atmósfera y los océanos de la Tierra utilizando ecuaciones matemáticas basadas en las leyes de la física. Para ello normalmente dividen la Tierra en

una red tridimensional y calculan las variables climáticas en cada punto de la red.

- **Modelos Acoplados de Océano-Atmósfera:** Integran modelos de circulación atmosférica y modelos de circulación oceánica para simular las interacciones entre el océano y la atmósfera. Por ejemplo, modelos como el NCEP (National Centers for Environmental Prediction) y el modelo de acoplamiento del Centro Europeo de Predicción a Medio Plazo (ECMWF).

3. **Métodos Empíricos** constituyen un enfoque crucial en la climatología moderna. Ese basan principalmente en observaciones y patrones encontrados en los datos históricos sin necesariamente tener un modelo explícito del proceso físico subyacente para predecir la evolución de ENSO. Estos datos pueden ser la temperatura del mar, presión atmosférica y otros indicadores clave, estos modelos empíricos pueden ofrecer proyecciones útiles sobre la probabilidad y la intensidad de futuros eventos de El Niño y La Niña. [65]El más utilizado es el Índice de Oscilación del Sur (SOI) y el Índice de El Niño (ONI) para realizar predicciones basadas en observaciones históricas. Los índices se calculan a partir de medidas como la presión atmosférica (en el caso del SOI) y la temperatura del océano (en el caso del ONI), proporcionando indicaciones sobre la probabilidad de eventos [48, 49].

2.2.2. Modelos con Machine Learning

Estos modelos se basan en patrones observados en datos históricos (como la relación entre variables climáticas y la ocurrencia del ENSO). Algunos modelos utilizados y ejemplos de ello son:

1. **Modelo Basados en Árboles de Regresión:**

   En un estudio publicado en MENDEL — Soft Computing Journal [70], se utilizó la técnica de aprendizaje automático supervisado mediante árboles de regresión para desarrollar un modelo predictivo de la fase del fenómeno El Niño-Oscilación del Sur (ENSO) utilizando datos del

periodo 1950-2022. El rendimiento del modelo, validado con las métricas MAE, ME y RMSE, mostró una mejora continua al aumentar la cantidad de datos de entrenamiento, reduciendo los errores significativamente desde 1953 hasta 2022. Se observó que la mayoría de los valores del ENSO durante el período de estudio eran neutros, y las características más relevantes para la predicción fueron los valores del ENSO del mes inmediatamente anterior. El modelo demostró ser consistente y confiable para predicciones a 12 meses, con un bajo costo computacional, aunque su precisión disminuye para pronósticos más largos.

2. **Modelo Basados en Regresión Lineal:**

   Según lo presentado por Sébastien Pascal, el uso de un modelo de regresión lineal con un predictor basado en la temperatura superficial del mar (SST) del Océano Índico meridional mejora la predicción del ENSO. Este estudio compara tres precursores del ENSO que pueden predecir a través de la barrera de persistencia de primavera: el contenido de calor anómalo del océano superior ecuatorial del Pacífico, la anomalía del esfuerzo del viento ecuatorial zonal en el extremo oeste del Pacífico y las anomalías de la SST en el sudeste del Océano Índico (SEIO) durante el final del invierno boreal. Un análisis de correlación muestra que los inicios de El Niño (La Niña) son precedidos por anomalías significativas de SST frías (cálidas) en el SEIO después del cambio de régimen climático de 1976-77. Los modelos de regresión lineal basados en la relación retrasada entre Niño3.4 SST y los predictores sugieren que las anomalías de SST en el SEIO son el predictor más robusto del ENSO. Los modelos predictivos para el periodo 1977-2001 demuestran que el uso de SEIO SST mejora significativamente la habilidad predictiva. El modelo UWPAC-SEIO indica que aproximadamente la mitad de la variabilidad interanual de Niño3.4 SST puede predecirse con una antelación de diez meses. Aunque el rendimiento predictivo podría ser inferior para años posteriores, este estudio apoya la inclusión de las SST del SEIO como un parámetro crucial en la evolución del ENSO, sugiriendo que podría mejorar la habilidad predictiva de los modelos actuales [64].

3. **Modelo basado en Random Forest:**

   Según lo presentado en el estudio sobre predicción climática subestacional en la cuenca del río Upper Colorado, se empleó un clasificador de bosque aleatorio para mejorar las predicciones climáticas en esta región vulnerable. El modelo se entrenó y probó con datos de 1982-2010 y se verificó con datos de 2017-2019, clasificando las condiciones en secas, húmedas y normales. Los resultados indican que el modelo tiene mejor habilidad que las predicciones del Centro de Predicción Climática (CPC) y sugiere una relación entre la fase de MJO y la habilidad de la predicción, aunque con un ligero sesgo hacia condiciones húmedas, posiblemente debido al cambio climático. Futuras investigaciones incluirán el uso de probabilidades del bosque aleatorio para comparaciones más directas con las predicciones del CPC [9].

4. **Modelos basados en redes neuronales LSTM (Long Short-Term Memory)**

   - El estudio liderado por Dong-Hoon Kim, Il-Ju Moon, Chaewook Lim y Seung-Buhm Woo sostiene que mediante el uso de redes neuronales recurrentes LSTM y una innovadora función de pérdida ponderada, es posible mejorar significativamente la predicción del fenómeno El Niño-Oscilación del Sur (ENSO). Esta metodología ajusta la función de pérdida original para reducir el peso relativo de eventos normales de alta frecuencia, lo cual ha demostrado una notable mejora en la precisión de las predicciones de ENSO, especialmente en años anormales como El Niño fuerte o La Niña fuerte, que tienen impactos significativos pero son menos frecuentes. Los resultados muestran un incremento considerable en la exactitud de las predicciones para horizontes temporales de 1 a 12 meses, destacando una mejora progresiva a medida que se extiende el período de predicción. Este enfoque no solo optimiza la capacidad predictiva de ENSO, sino que también sugiere aplicaciones potenciales en la predicción de otros eventos climáticos extremos de baja frecuencia pero alto impacto, ampliando las

posibilidades para la gestión de riesgos climáticos y la planificación estratégica a largo plazo en diversas regiones del mundo [41].

- El estudio realizado por Cao Xiaoqun, Guo Yanan, Liu Bainian, Peng Kecheng, Wang Guangjie, y Gao Mei explora el uso de redes neuronales recurrentes con LSTM para la predicción del ENSO. Los resultados mostraron que el modelo basado en LSTM ofrecía predicciones altamente correlacionadas con los datos observacionales, subrayando la capacidad de las técnicas de aprendizaje profundo para mejorar la precisión en la predicción de eventos ENSO. El análisis comparativo con métodos de regresión estadística tradicionales muestra que las LSTM tienen un rendimiento superior, particularmente en la predicción de la región Niño3.4 utilizando datos del Centro Nacional de Predicción Ambiental (NCEP). Los resultados sugieren que las técnicas de aprendizaje profundo tienen un gran potencial en la predicción de eventos climáticos extremos, destacando la necesidad de conjuntos de datos de alta calidad para construir modelos predictivos fiables. En futuras investigaciones, se propone la introducción de métodos de regularización para mejorar aún más la precisión de los modelos [14].
- En el artículo "Forecasting ENSO using convolutional LSTM network with improved attention mechanism and models recombined by genetic algorithm in CMIP5/6"publicado en Information Sciences por Y. Wang et al., se presenta un enfoque avanzado para la predicción del Fenómeno El Niño-Oscilación del Sur (ENSO). El estudio utiliza una combinación de redes LSTM convolucionales y un mecanismo de atención mejorado para mejorar la extracción de características espaciotemporales tanto locales como globales. Para mejorar la calidad de los datos de entrenamiento, se aplicó un algoritmo genético para recombinar y filtrar modelos del CMIP5/6, resultando en un conjunto de datos de entrenamiento más robusto y preciso. Comparando con otros modelos de aprendizaje profundo como CNN, CNN + LSTM, ResNet + LSTM, Transformer

y Swin Transformer, el modelo propuesto demostró un rendimiento superior con el menor error cuadrático medio (MSE) y el coeficiente de correlación (CC) más alto. Este enfoque no solo destaca por su capacidad para mejorar las predicciones a largo plazo de ENSO, sino que también subraya la importancia de personalizar cada módulo de aprendizaje profundo según las características específicas de los datos oceánicos y climáticos. Además, el estudio sugiere futuras investigaciones que podrían integrar variables adicionales y técnicas como redes neuronales gráficas para seguir avanzando en la precisión y comprensión de fenómenos climáticos complejos como ENSO [73].

5. **Método basado en Red Convolucional Temporal EEMD**

   En el artículo "Temporal Convolutional Networks for the Advance Prediction of ENSO", se presenta un enfoque innovador para mejorar la precisión en la predicción de El Niño-Oscilación del Sur (ENSO). El estudio propone utilizar una combinación de Redes Convolucionales Temporales (TCN) y la técnica de descomposición empírica en modos ensembles (EEMD) para abordar la complejidad de los índices Niño 3.4 y de Oscilación del Sur (SOI), conocidos por su alta variabilidad y mezcla de componentes de baja y alta frecuencia. El método EEMD-TCN primero descompone estos índices en subcomponentes más manejables y luego aplica TCN para predecir cada subcomponente con anticipación. Los resultados muestran que esta estrategia mejora significativamente la precisión de las predicciones tanto a corto como a largo plazo, superando a métodos convencionales como LSTM y EEMD-LSTM. Específicamente, se observó que las predicciones a un mes mostraron la mayor precisión, con correlaciones significativas (PCC) y errores cuadráticos medios (RMSE) favorables para el índice Niño 3.4 y el SOI.A pesar de estos avances, el estudio identifica desafíos durante períodos de fuertes El Niño o La Niña, donde los errores de predicción aún pueden ser significativos, especialmente en predicciones a largo plazo. Esto subraya la necesidad de seguir desarrollando modelos que puedan manejar mejor estas condiciones extremas. Además, el artículo destaca la eficiencia

operativa y la estructura más simple de TCN en comparación con LSTM, lo cual es crucial para la escalabilidad y aplicación práctica en predicciones climáticas a gran escala [76].

6. **Modelo basado en redes neuronales y series temporales**

   La tesis "Pronóstico ENOS en las regiones Niño 3.4 y Niño 1+2, utilizando redes neuronales profundas con secuencias espacio temporales"presentada por Kennedy Richard Gomez Tunque, aborda el uso de redes neuronales profundas para la predicción del fenómeno El Niño-Oscilación del Sur (ENOS) en las regiones Niño 3.4 y Niño 1+2. La investigación se centra en la aplicación de modelos de redes neuronales que analizan datos espacio-temporales, buscando mejorar la precisión de los pronósticos de anomalías de temperatura superficial del mar (TSM) en estas áreas críticas. El estudio compara el desempeño de las redes neuronales profundas con métodos tradicionales de predicción, demostrando la capacidad de las redes para capturar patrones complejos y mejorar las predicciones a largo plazo. Los resultados sugieren que las redes neuronales profundas pueden ser una herramienta efectiva para el pronóstico de ENOS, contribuyendo a una mejor preparación y respuesta ante eventos climáticos extremos [67].

Capítulo 3

Herramientas y Metodología

En este capítulo se describe las herramientas usadas en esta investigación y la dividiremos en dos secciones, la primera describirá herramientas para la predicción del Fenómeno del Niño (ENSO) en la costa norte del Perú y la segunda parte expone la metodología utilizada para este objetivo

3.1. Área de Estudio

La costa norte del Perú, específicamente la región conocida como Región del Niño 1+2, es el área focal de este estudio. Esta región es crítica para la monitorización del Fenómeno del Niño debido a su influencia directa en los patrones climáticos locales y regionales. Los límites geográficos de la costa norte del Perú están delimitados por las siguientes coordenadas:

1. **Latitud:** Desde -6.0 hasta -3.5 grados
2. **Longitud:** Desde aproximadamente -81.0 hasta -79.0 grados

Estas coordenadas comprenden una porción significativa del mar peruano del norte, el cual es monitoreado para estudiar los efectos del ENSO.

3.2. Herramientas

3.2.1. Lenguaje de Programación

El lenguaje de programación utilizado es Python debido a su amplia gama de bibliotecas científicas y de machine learning como:

- Pandas: Desempeñó un papel importante en la manipulación y análisis de datos utilizando el marco de datos flexible para facilitar la limpieza, transformación y exploración inicial de datos.
- NumPy: se utiliza para realizar operaciones matemáticas y gestionar matrices de manera eficiente utilizando una amplia gama de funciones matemáticas y funciones de álgebra lineal esenciales para procesar grandes conjuntos de datos numéricos.
- Scikit-learn: para implementar modelos de aprendizaje automático como algoritmos de clasificación, regresión, agrupamiento y reducción de dimensionalidad. Además, Scikit-learn facilita la creación y evaluación de modelos predictivos al proporcionar herramientas para validar modelos y evaluar su desempeño.
- XGBoost: Se utiliza para implementar modelos de gradiente potenciados y es reconocido por su eficiencia y alto rendimiento en competencias de aprendizaje automático. XGBoost proporciona una implementación optimizada del algoritmo de impulso, mejorando la precisión y la velocidad del entrenamiento del modelo. TensorFlow y Keras: se utilizan para implementar redes neuronales profundas, incluidas redes neuronales recurrentes (RNN) y redes neuronales convolucionales (CNN).

- TensorFlow proporciona una plataforma potente y flexible para crear y entrenar modelos de aprendizaje profundo, mientras que Keras simplifica la creación y experimentación de modelos de redes neuronales con su API de alto nivel.
- Matplotlib: se utiliza para visualizar datos y resultados de forma eficaz proporcionando una variedad de gráficos estáticos para facilitar la interpretación y presentación de los resultados.

Estas bibliotecas facilitan el análisis y la manipulación de datos, la creación y evaluación de modelos de aprendizaje automático, y la implementación de redes neuronales y otros algoritmos avanzados. Su sintaxis sencilla y clara

también permite desarrollar y mantener el código de manera eficiente, haciendo de Python una opción ideal para proyectos de investigación y desarrollo en el ámbito del análisis climático y la predicción del fenómeno de El Niño (ENSO).

3.2.2. Datos

La base de datos de temperaturas superficiales del mar (TSM) de los años 2017 a 2023 fue obtenida del Instituto Humboldt de Investigación Marina y Acuícola (IHMA) [32].Incluye las siguientes columnas:

- **Latitud:** En números reales reflejando coordenadas geográficas
- **Longitud:** En números reales reflejando coordenadas geográficas
- **Fechas** cada día del año es una columna y en esta se encuentra la temperatura medida ese dia en ese punto geográfico específico

3.2.3. Entorno de Desarrollo

Se escogio trabajar con Pycharm en su version 2024.1 con una licencia educativa [36]. Este entorno desarrollado por JetBrains se utiliza habitualmente para el desarrollo de aplicaciones en Python, principalmente por grandes empresas como Twitter, Facebook, Amazon y Pinterest [4].

PyCharm ofrece numerosas ventajas para los desarrolladores. Su editor de código inteligente facilita la escritura de código de alta calidad, destacando palabras clave, clases y funciones con diferentes colores que mejoran la legibilidad y facilitan la detección de errores. Además, cuenta con funciones de autocompletar que agilizan el desarrollo y es compatible con bibliotecas científicas como Matplotlib, NumPy y Anaconda, siendo una herramienta valiosa para proyectos de Data Science y Machine Learning [4].

3.3. Metodología

3.3.1. Preprocesamiento de Datos

La presentación de las bases de datos proporcionados por El Instituto Humboldt de Investigación Marina y Acuícola (IHMA) no es la más adecuada para el uso de este estudio, por lo que en primera instancia empezaremos con una transformacion de esta base de datos a otra nueva que sirva para el analisis que vamos a hacer.A continuación se describe paso a paso el método empleado:

1. **Recolección de Datos:** Se obtuvo las bases de datos del IHMA sobre la temperatura superficial del mar (TSM) por año, por ubicacion geográfica y por fecha a lo largo de todo el mar peruano [33].

2. **Eliminacion de Datos:** Primero se procedio hacer filtrar cualquier valor nulo que pudiera existir en las columnas de temperatura, asegurando así la integridad de los datos.

   Posteriormente, se delimitaron los datos correspondientes únicamente a la región del mar del norte del Perú, utilizando las coordenadas de latitud (-6.0 a -3.5 grados) y descartando datos fuera de este rango geográfico.

3. **Generación de Nuevas Bases de Datos:** Para cada año, se calculó el promedio diario de temperaturas. Este proceso consistió en agrupar todas las observaciones por fecha y calcular la temperatura promedio para todas las latitudes dentro de la región especificada.

   A partir de los pasos anteriores, se obtuvieron 7 nuevas bases de datos, cada una conteniendo dos columnas:

   - **Fecha:** Representa el día específico para el cual se calculó el promedio.
   - **Temperatura Promedio**: Es el valor numérico que representa la temperatura promedio diaria en la costa norte del Perú para esa fecha.

4. **Detección de Outliers:** Mediante el rango intercuartílico (IQR) se procedio a calcular los outliers en las temperaturas promedio calculadas. Los pasos fueron los siguientes:

- Se calculó el IQR para la distribución de temperaturas.El IQR se calcula como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de la distribución de datos:

$$\text{IQR} = Q3 - Q1 \tag{3.1}$$

  Los outliers se identifican como aquellos valores que están por debajo de $Q1 - 1.5 \times$ IQR o por encima de $Q3 + 1.5 \times$ IQR.

- Si se detectaron outliers,se corrigen reemplazando los valores atípicos con el promedio de las temperaturas de los días anterior y posterior al outlier,

5. **Generación de una nueva Base de Datos** Una vez los outliers han sido corregidos, las 7 bases de datos resultantes, correspondientes a los años 2017 al 2023, fueron unidas en un solo DataFrame y exportadas a un solo archivo para el posterior analisis de los modelos.

   Finalmente, se realizó una visualización integral de los datos procesados. Se generó un gráfico que muestra la evolución de las temperaturas promedio de 2017 a 2023. Además, se incluyeron características adicionales como las medias móviles (método estadístico utilizado para suavizar las fluctuaciones en los datos y destacar tendencias a largo plazo) de 7 días(eq 3.2) y 30 días(3.3).

$$\text{MA}_7(t) = \frac{1}{7} \sum_{i=t-6}^{t} x_i \tag{3.2}$$

$$\text{MA}_{30}(t) = \frac{1}{30} \sum_{i=t-29}^{t} x_i \tag{3.3}$$

Este proceso de preprocesamiento asegura la calidad y la coherencia de los datos utilizados en el análisis y predicción de ENSO en el mar de la

costa norte del Perú, proporcionando una base sólida para los modelos de machine learning aplicados posteriormente.

3.4. Visualización de los Datos

Se muestran en las figuras 3.1, 3.2, 3.3, 3.4, 3.5 y 3.6 los valores TSM graficados en su respectivo y en las tablas 3.1, 3.3, 3.5, 3.7, 3.9, 3.11 y 3.14 un análisis descriptivo de los valores por cada mes tales como las temperaturas máximas y mínimas y la temperatura promedio, asi como la descripción anual.

En las tablas 3.2, 3.4, 3.6, 3.8, 3.10, 3.12, 3.14 se muestra el resumen descriptivo de todos los datos en sus respectivos años, se muestran metricas como la mediana, media y desviación estandar.

Mes	Temperatura Mínima	Temperatura Máxima	Temperatura Promedio
Enero	23.54	27.75	25.87
Febrero	27.62	28.78	28.24
Marzo	28.54	29.47	29.01
Abril	25.96	28.79	27.20
Mayo	23.55	34.74	26.90
Junio	22.04	23.43	22.63
Julio	20.93	22.21	21.60
Agosto	19.67	21.13	20.33
Septiembre	18.66	19.85	19.29
Octubre	18.87	19.67	19.23
Noviembre	19.44	20.61	20.10
Diciembre	20.79	22.65	21.46

Tabla 3.1: Temperaturas mensuales (mínima, máxima, promedio) para el año 2017.

Descripción	Valor
Count	365.000000
Mean	23.462352
Std Dev	3.780280
Min	18.655079
25 %	20.186352
50 % (Mediana)	22.140140
75 %	26.938290
Max	34.735517

Tabla 3.2: Descripción de los datos de temperatura promedio para el año 2017.

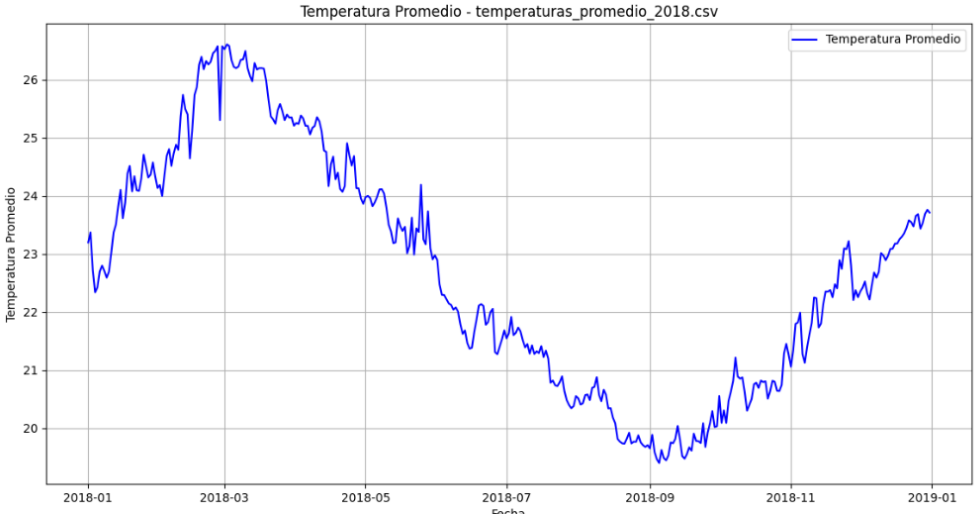

Figura 3.1: Temperatura Superficial del Mar - 2018.

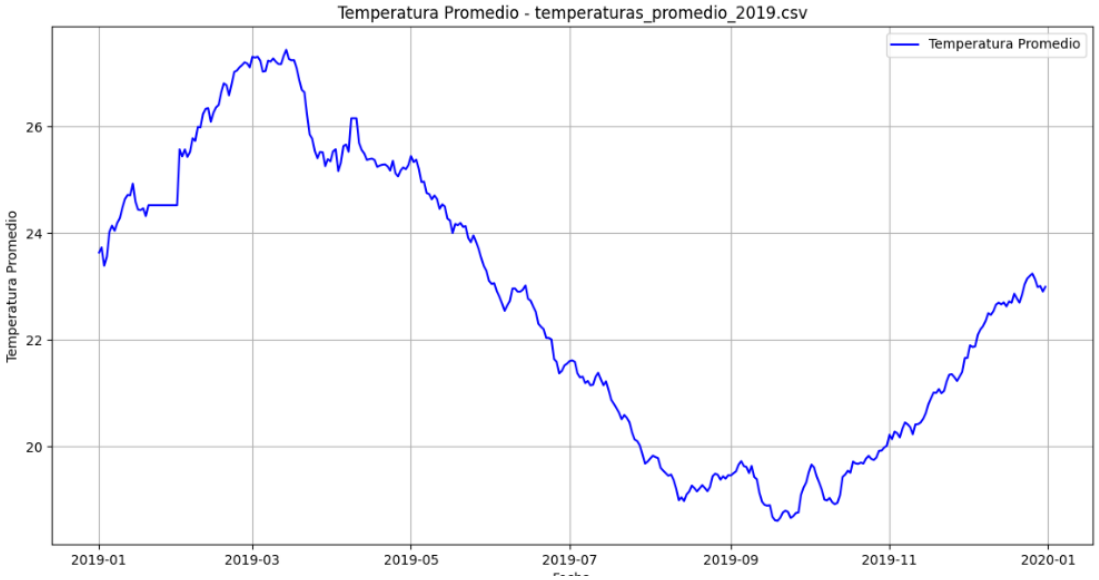

Figura 3.2: Temperatura Superficial del Mar - 2019.

Mes	Temperatura Mínima	Temperatura Máxima	Temperatura Promedio
Enero	22.34	24.71	23.66
Febrero	24.00	26.57	25.48
Marzo	25.21	26.61	25.93
Abril	23.86	25.38	24.72
Mayo	22.91	24.19	23.54
Junio	21.28	22.90	21.89
Julio	20.34	21.92	21.14
Agosto	19.68	20.88	20.17
Septiembre	19.40	20.29	19.75
Octubre	20.09	21.45	20.71
Noviembre	21.06	23.22	22.14
Diciembre	22.21	23.76	23.10

Tabla 3.3: Temperaturas mensuales (mínima, máxima, promedio) para el año 2018.

Descripción	Valor
Count	365.000000
Mean	22.669951
Std Dev	2.025804
Min	19.402040
25 %	20.812368
50 % (Mediana)	22.479368
75 %	24.167971
Max	26.606617

Tabla 3.4: Descripción de los datos de temperatura promedio para el año 2018.

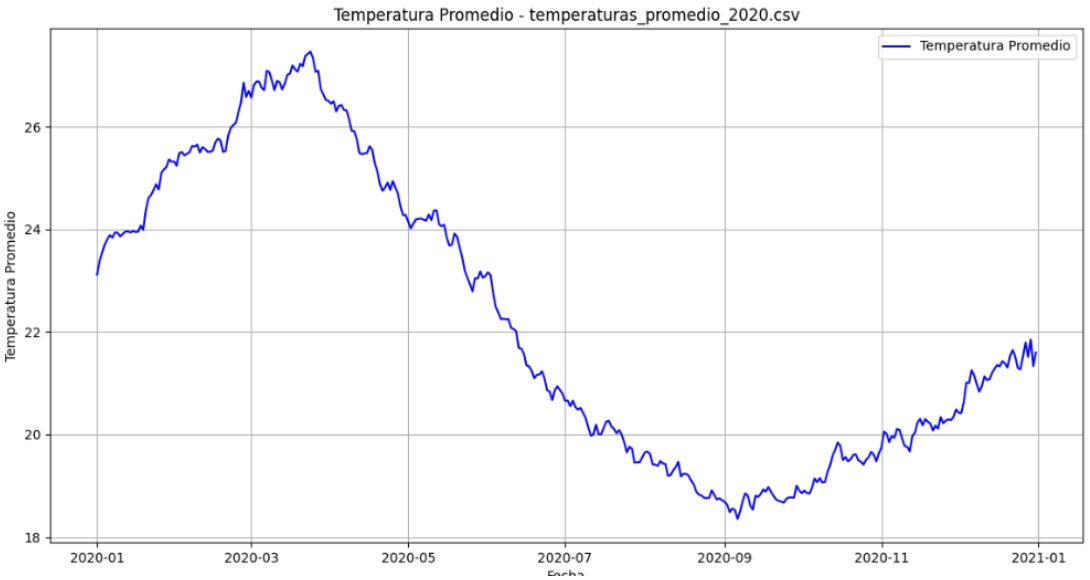

Figura 3.3: Temperatura Superficial del Mar - 2020.

Mes	Temperatura Mínima	Temperatura Máxima	Temperatura Promedio
Enero	23.39	24.93	24.36
Febrero	25.43	27.21	26.38
Marzo	25.26	27.44	26.65
Abril	25.07	26.16	25.44
Mayo	23.11	25.45	24.33
Junio	21.37	23.06	22.42
Julio	19.68	21.62	20.83
Agosto	18.98	19.83	19.37
Septiembre	18.61	19.72	19.11
Octubre	18.92	20.02	19.51
Noviembre	20.14	21.66	20.77
Diciembre	21.66	23.24	22.63

Tabla 3.5: Temperaturas mensuales (mínima, máxima, promedio) para el año 2019.

Descripción	Valor
Count	365.000000
Mean	22.626575
Std Dev	2.677148
Min	18.605579
25 %	19.917709
50 % (Mediana)	22.667133
75 %	25.128951
Max	27.442703

Tabla 3.6: Descripción de los datos de temperatura promedio para el año 2019.

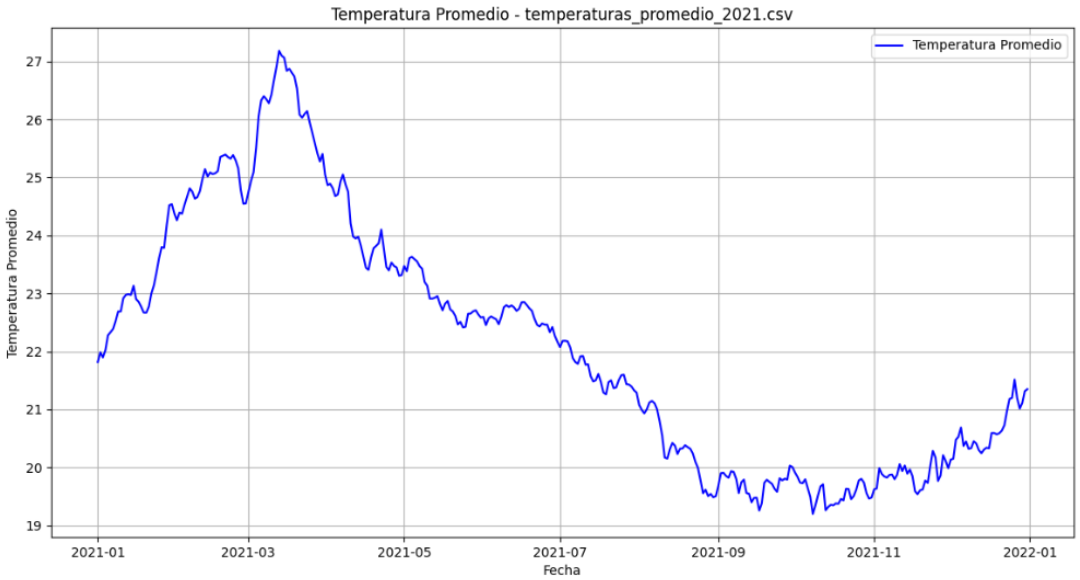

Figura 3.4: Temperatura Superficial del Mar - 2021.

Mes	Temperatura Mínima	Temperatura Máxima	Temperatura Promedio
Enero	23.12	25.36	24.27
Febrero	25.24	26.86	25.79
Marzo	26.50	27.46	26.96
Abril	24.28	26.49	25.45
Mayo	-85.66	24.37	20.20
Junio	20.67	23.16	21.65
Julio	19.45	20.66	20.10
Agosto	18.72	19.67	19.15
Septiembre	18.35	19.00	18.73
Octubre	18.85	19.84	19.38
Noviembre	19.67	20.49	20.10
Diciembre	20.42	21.85	21.23

Tabla 3.7: Temperaturas mensuales (mínima, máxima, promedio) para el año 2020.

Descripción	Valor
Count	366.000000
Mean	21.900805
Std Dev	6.296284
Min	-85.663126
25 %	19.653951
50 % (Mediana)	21.291940
75 %	24.774826
Max	27.462077

Tabla 3.8: Descripción de los datos de temperatura promedio para el año 2020.

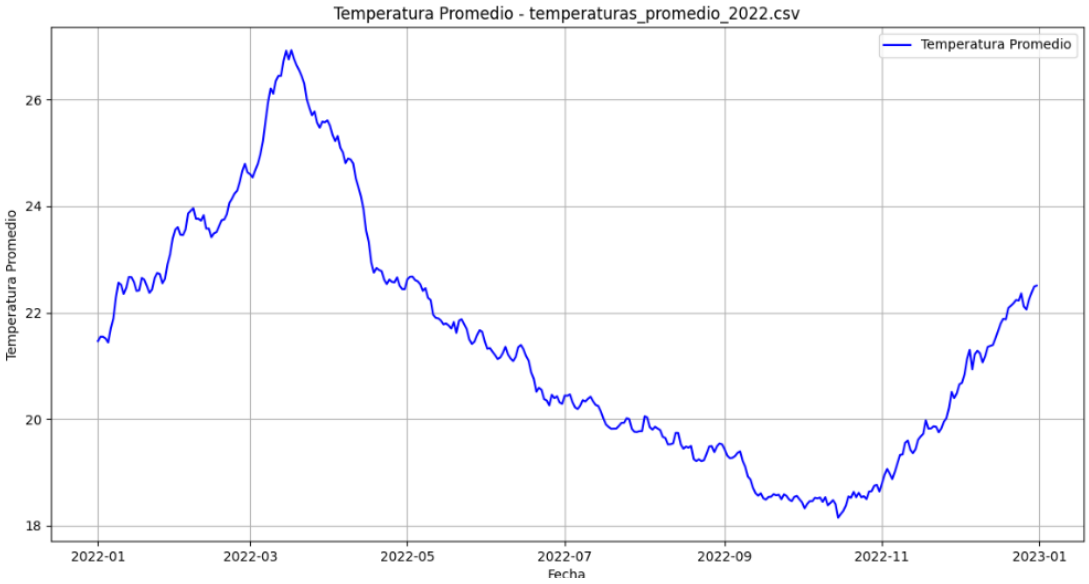

Figura 3.5: Temperatura Superficial del Mar - 2022.

Mes	Temperatura Mínima	Temperatura Máxima	Temperatura Promedio
Enero	21.82	24.54	22.99
Febrero	24.26	25.39	24.92
Marzo	24.75	27.18	26.12
Abril	23.31	25.05	24.03
Mayo	22.41	23.63	22.94
Junio	22.17	22.85	22.59
Julio	21.26	22.18	21.65
Agosto	19.49	21.15	20.34
Septiembre	19.26	20.03	19.72
Octubre	19.20	19.91	19.55
Noviembre	19.54	20.29	19.88
Diciembre	20.14	21.51	20.66

Tabla 3.9: Temperaturas mensuales (mínima, máxima, promedio) para el año 2021.

Descripción	Valor
Count	365.000000
Mean	22.097863
Std Dev	2.125606
Min	19.199008
25 %	20.029472
50 % (Mediana)	22.022613
75 %	23.531962
Max	27.183427

Tabla 3.10: Descripción de los datos de temperatura promedio para el año 2021.

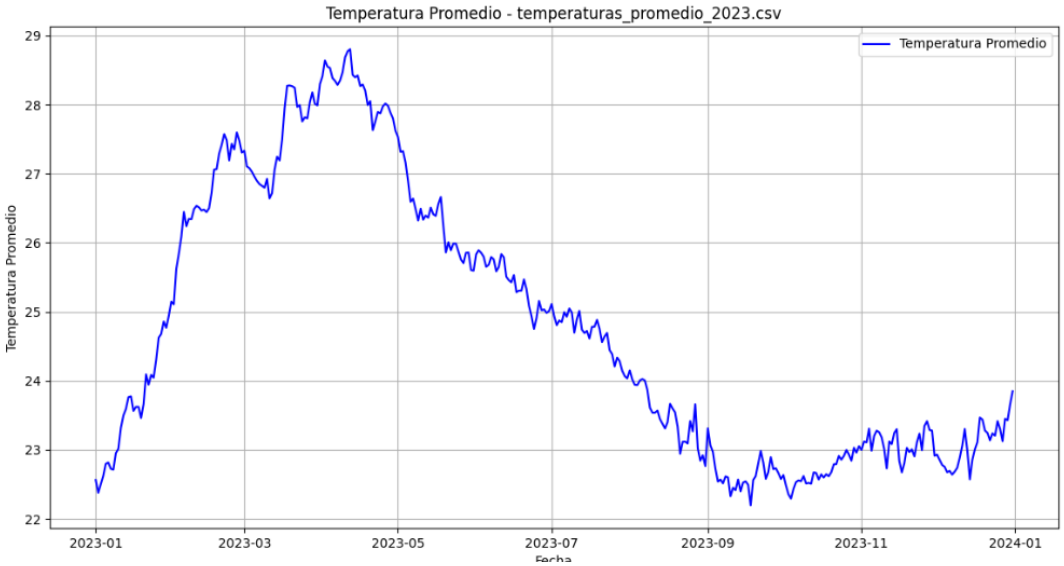

Figura 3.6: Temperatura Superficial del Mar - 2023.

Mes	Temperatura Mínima	Temperatura Máxima	Temperatura Promedio
Enero	21.44	23.56	22.41
Febrero	23.42	24.80	23.88
Marzo	24.54	26.93	25.91
Abril	22.44	25.51	23.73
Mayo	21.41	22.68	21.97
Junio	20.26	21.39	20.90
Julio	19.76	20.46	20.09
Agosto	19.21	20.05	19.56
Septiembre	18.46	19.43	18.80
Octubre	18.15	18.77	18.49
Noviembre	18.79	20.51	19.61
Diciembre	20.65	22.51	21.66

Tabla 3.11: Temperaturas mensuales (mínima, máxima, promedio) para el año 2022.

Descripción	Valor
Count	365.000000
Mean	21.404506
Std Dev	2.245725
Min	18.147245
25 %	19.540776
50 % (Mediana)	21.218947
75 %	22.652250
Max	26.929729

Tabla 3.12: Descripción de los datos de temperatura promedio para el año 2022.

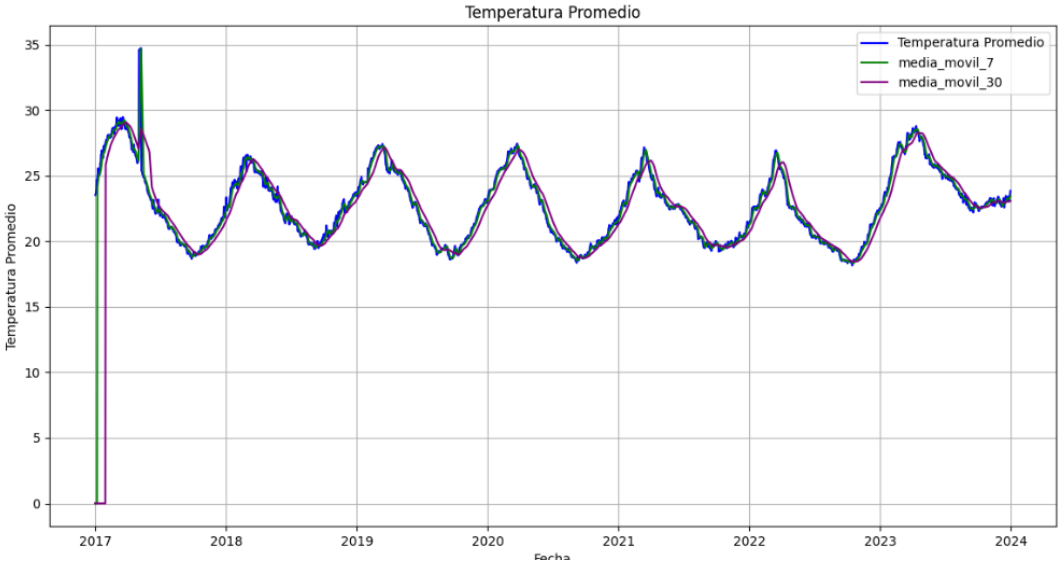

Figura 3.7: TSM y Medias Móviles desde 2017 hasta 2023.

Mes	Temperatura Mínima	Temperatura Máxima	Temperatura Promedio
Enero	22.38	25.15	23.63
Febrero	25.11	27.60	26.73
Marzo	26.64	28.30	27.52
Abril	27.62	28.80	28.22
Mayo	25.60	27.53	26.34
Junio	24.75	25.89	25.42
Julio	24.04	25.11	24.67
Agosto	22.77	24.15	23.49
Septiembre	22.20	23.31	22.65
Octubre	22.29	23.05	22.68
Noviembre	22.67	23.42	23.09
Diciembre	22.57	23.85	23.09

Tabla 3.13: Temperaturas mensuales (mínima, máxima, promedio) para el año 2023.

Descripción	Valor
Count	365.000000
Mean	24.778198
Std Dev	1.954640
Min	22.197413
25 %	22.998755
50 % (Mediana)	24.304194
75 %	26.466520
Max	28.802305

Tabla 3.14: Descripción de los datos de temperatura promedio para el año 2023.

Descripción	Valor
Count	2556.000000
Mean	22.748397
Std Dev	2.774521
Min	18.147245
25 %	20.303690
50 % (Mediana)	22.582634
75 %	24.784888
Max	34.735517

Tabla 3.15: Descripción de los datos combinados de temperatura promedio (2017-2023).

3.4.1. Modelamiento

División de Datos

Se utilizan los datos del 2017 hasta el 2022 para el entrenamiento de los modelos, y los datos del 2023 se utilizan para la validación. Todos los datos son utilizados tanto para el entrenamiento como para la validación, lo que es particularmente importante ya que se dispone de conjuntos de datos limitados.

Modelos utilizados

Se utilizan diversos modelos de machine learning para la predicción de las temperaturas superficiales del mar, seleccionando cada uno por sus características particulares y por los antecedentes revisados en la literatura.

1. **Red Neuronal Convolucional** Este modelo combina capas de convolución 1D con capas LSTM para el procesamiento de secuencias temporales, lo que es optimo para capturar patrones espaciales y dependencias temporales.

   Detalle de las Capas

   - Capa Conv1D: Esta capa aplica convoluciones unidimensionales sobre la secuencia de entrada [8]. Los parámetros que se especifican son:
     - filters=64: Número de filtros convolucionales que se aplican. Cada filtro aprenderá diferentes características de las secuencias de entrada.
     - kernel size=2: Tamaño del kernel de convolución, que especifica cuántas secuencias de tiempo consecutivas se consideran en cada paso de la convolución.
     - activation='relu': Función de activación ReLU (Rectified Linear Unit) que se aplica después de la convolución para introducir no linealidad.

- Capa LSTM: Después de la capa convolucional, se aplica una capa LSTM para modelar dependencias temporales más largas en los datos [7]. Los parámetros especificados son:
  - units=50: Número de unidades LSTM o celdas de memoria en esta capa. Cada unidad tiene conexiones internas que le permiten recordar información de secuencias pasadas.
  - activation='relu': Función de activación ReLU aplicada dentro de la LSTM para introducir no linealidad.
- Capa Dense: Finalmente, una capa densa con una sola unidad (neurona) se utiliza para predecir la temperatura siguiente en la secuencia de tiempo [6]. No se aplica función de activación en esta capa porque estamos realizando una regresión lineal.

2. **Red Neuronal Recurrente**

   Se ha implementado específicamente utilizando LSTM (Long Short-Term Memory), pues son especialmente adecuadas para modelar datos secuenciales y capturar dependencias a largo plazo en las series temporales. En este caso ademas se usa MinMaxScaler de sklearn [5] para normalizar los datos de temperatura promedio.

   Detalle de las Capas

   - Primera Capa LSTM: tiene 50 unidades y espera secuencias de entrada de longitud `sequence_length`. `return_sequences=True` indica que la capa LSTM devuelve secuencias completas [40].
   - Capa de Dropout después de cada capa LSTM con una tasa del 20 %. Dropout ayuda a prevenir el sobreajuste al apagar aleatoriamente un porcentaje de unidades durante el entrenamiento [39].
   - Segunda Capa LSTM también tiene 50 unidades, pero `return_sequences=False`, lo que significa que solo devuelve la salida en el último paso de tiempo, esto permite permite que la red neuronal optimice el aprendizaje y la predicción manteniendo al

mismo tiempo la capacidad de capturar dependencias temporales a largo plazo en las secuencias de datos [40].

- Capa Densa con una sola unidad predice la temperatura promedio siguiente [38].

3. **Arboles** Son fáciles de interpretar y pueden manejar tanto datos categóricos como continuos y sirven para identificar las variables más importantes y las interacciones entre ellas.

   En este caso, se ha utilizado un árbol de decisión para la regresión, optimizando los hiperparámetros con GridSearchCV, en función del error cuadrático medio. Tambien se realiza una validación cruzada para evaluar la capacidad de generalización del modelo.

4. **XGboost**

   Tiene la capacidad para manejar características complejas como el día del año y el mes. La optimización de hiperparámetros a través de GridSearchCV permite encontrar configuraciones óptimas, maximizando así el rendimiento predictivo del modelo. Además, XGBoost es eficiente en términos de rendimiento y manejo de grandes volúmenes de datos, lo cual es crucial para trabajar con series temporales extensas y datos climáticos [53].

   Para este modelo adicionalmente se hizo un featuring de caracteristicas como agregar el día del año, el mes y el día, para capturar mejor los patrones estacionales. Tambien se hizo un escalamiento de las características usando StandardScaler, lo cual es importante para muchos modelos de machine learning, incluyendo XGBoost.

5. **Random Forest**

   El modelo puede manejar datos no lineales y capturar relaciones complejas entre variables, que son esenciales par este caso de estudios, que son multifactoriales y tienen variaciones estacionales. Además, los bosques aleatorios evitan el sobreajuste porque tienen una estructura basada en múltiples árboles de decisión, cada uno de los cuales se entrena utilizando muestras y características aleatorias del conjunto

de datos. Esto permite una buena generalización incluso de datos no vistos, como lo demuestra la capacidad del modelo para hacer predicciones precisas sobre datos de prueba [16]. También puede realizar automáticamente la selección de características y manejar datos con valores faltantes sin requerir un procesamiento previo extenso, lo que lo hace ideal para el análisis de series temporales de temperatura donde los datos pueden ser ruidosos e incompletos. Finalmente, GridSearchCV combinado con validación cruzada optimiza los hiperparámetros del modelo y proporciona una evaluación rigurosa de su rendimiento predictivo.

3.4.2. Catalogación del ENSO

Definimos las temperaturas normales estacionales para cada mes utilizando un diccionario, lo cual nos permitió calcular las anomalías estacionales como la diferencia entre las temperaturas previstas y las temperaturas normales. Posteriormente, aplicamos una media móvil de tres meses a estas anomalías para suavizar las fluctuaciones diarias. Para clasificar las anomalías según el Índice Costero El Niño (ICEN) utilizamos la tabla 2.2. Finalmente, filtramos los eventos significativos y representamos la evolución de la clasificación ICEN a lo largo del tiempo mediante una gráfica, facilitando la visualización de los patrones climáticos estacionales.

Capítulo 4

Resultados

4.1. Resultados de los Modelos

Se muestran los siguientes resultados:

- Las principales métricas de los modelos tales como el MSE, RMSE, MAE y MAPE. Ver tabla 4.1.
- Las graficas de las temperaturas predichas de los modelos vs. la temperatura observada, asi como las metricas que obtenemos en cada modelo. Ver figuras 4.5, 4.3, 4.9, 4.7, 4.1.
- Se muestra en ICEN de las proyecciones de los modelos para el año 2023. Ver figuras 4.6, 4.2, 4.8, 4.10, 4.4
- Se muestran las proyecciones de la temperatura superficial del mar (TSM) de los años 2024 y 2025 con los modelos de Redes Neuronales Convolucionales(Fig.4.11) y Redes Neuronales Recurrentes(Fig.4.12).
- Se muestran las proyecciones para El Índice Costero El Niño de los años 2024 y 2025 con los modelos de Redes Neuronales Convolucionales(Fig.**??**) y Redes Neuronales Recurrentes(Fig.**??**).

Modelo	**MSE**	**RMSE**	**MAE**	**MAPE**
Árboles de regresión	7.43	2.725	2.43	9.94 %
RNN	0.0395	0.198	0.152	0.62 %
CNN	0.03	0.18	0.14	0.58 %
Random Forest	0.04	0.20	0.15	0.62 %
XGBoost	7.37	2.72	2.43	9.92 %

Tabla 4.1: Comparación de métricas de modelos.

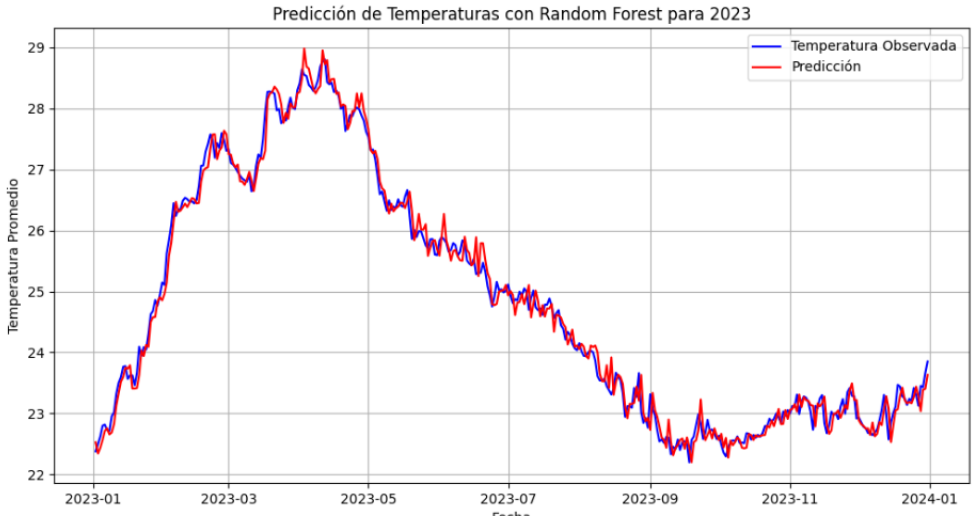

Figura 4.1: TSM predicha por el modelo Random Forest vs. mediciones reales.

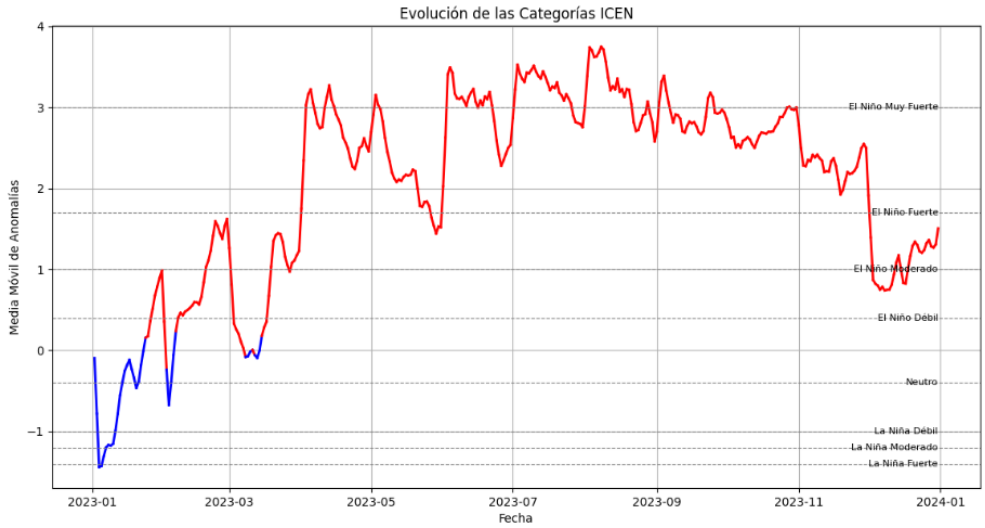

Figura 4.2: ICEN predicho por el modelo Random Forest.

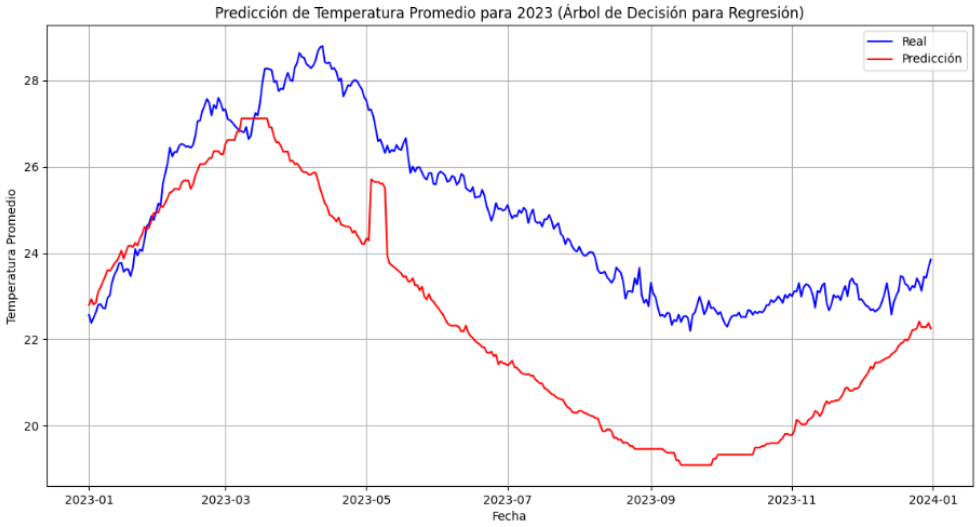

Figura 4.3: TSM predicha por el modelo Árbol de Regresión vs. mediciones reales.

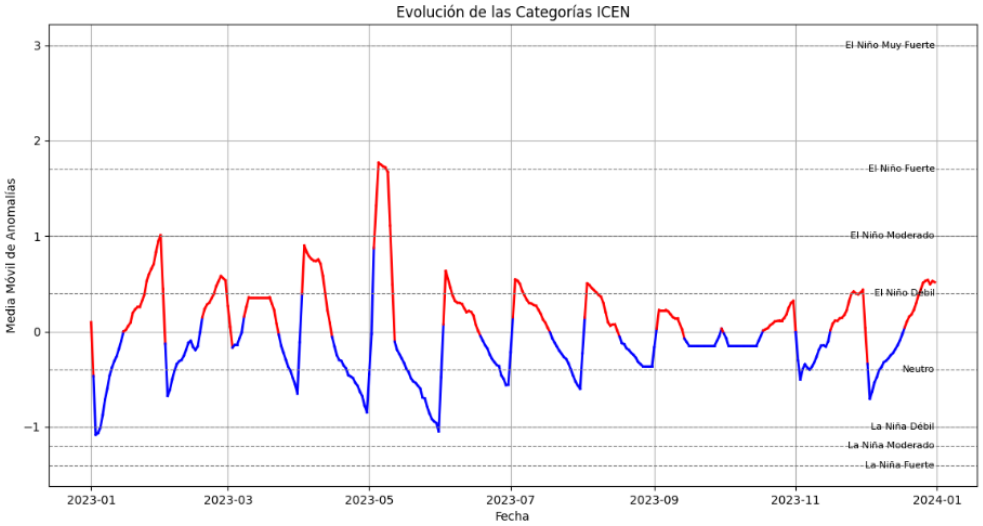

Figura 4.4: ICEN predicho por el modelo Árbol de Regresión.

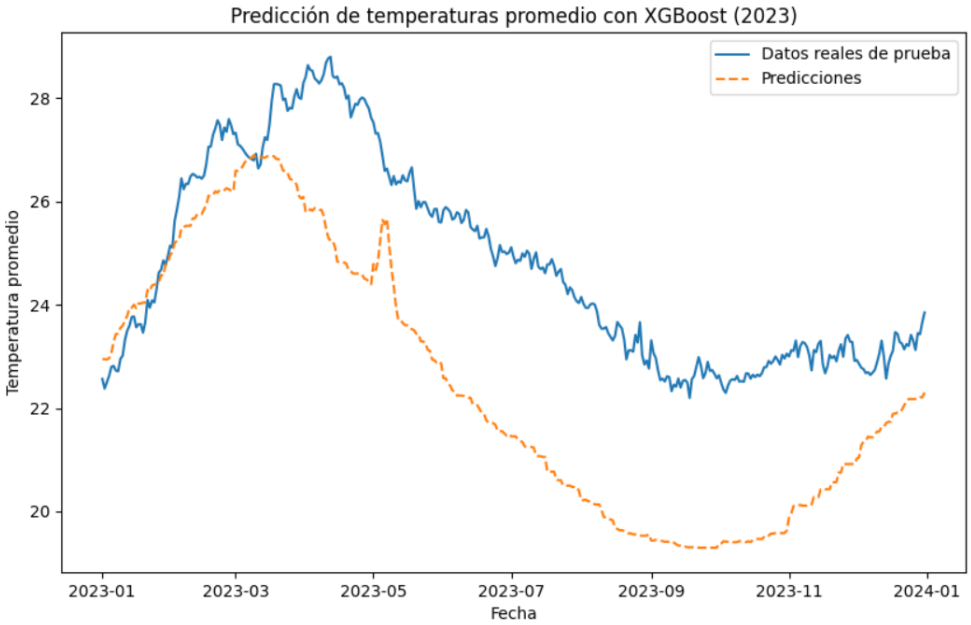

Figura 4.5: TSM predicha por el modelo XGBoost vs. mediciones reales.

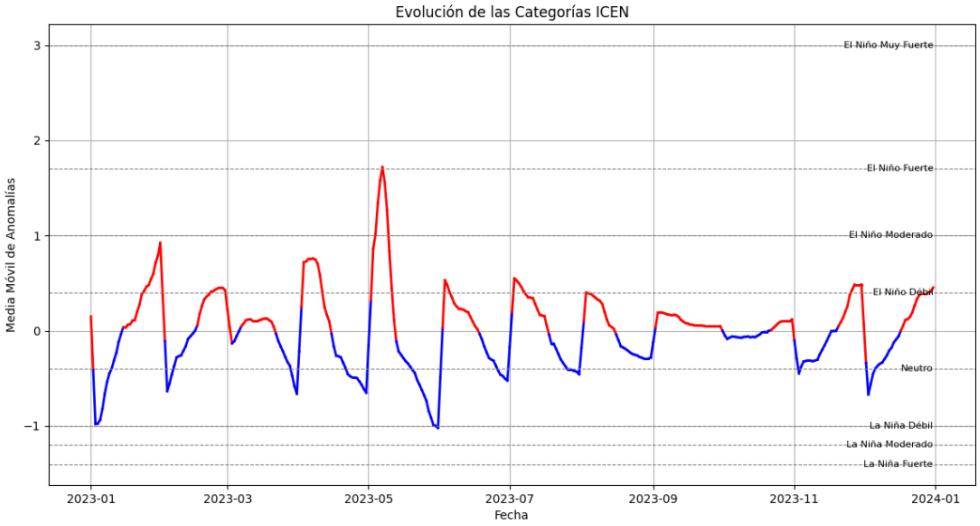

Figura 4.6: ICEN predicho por el modelo XGBoost.

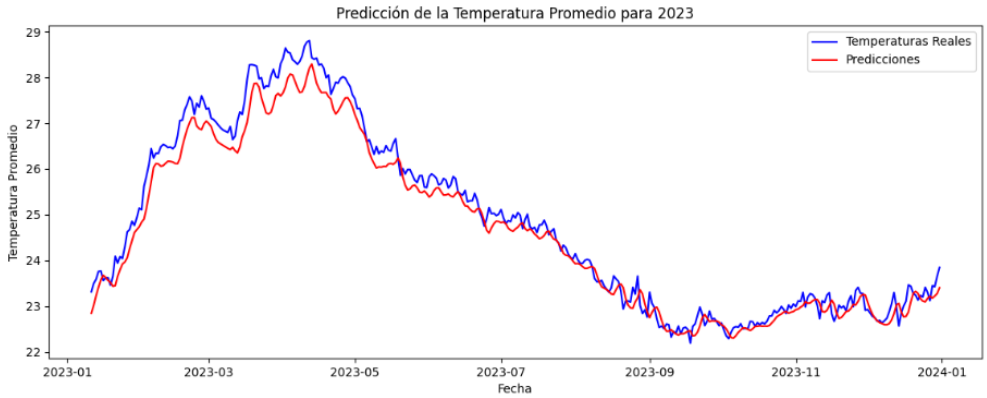

Figura 4.7: TSM predicha por el modelo Red Neuronal Recurrente vs. mediciones reales.

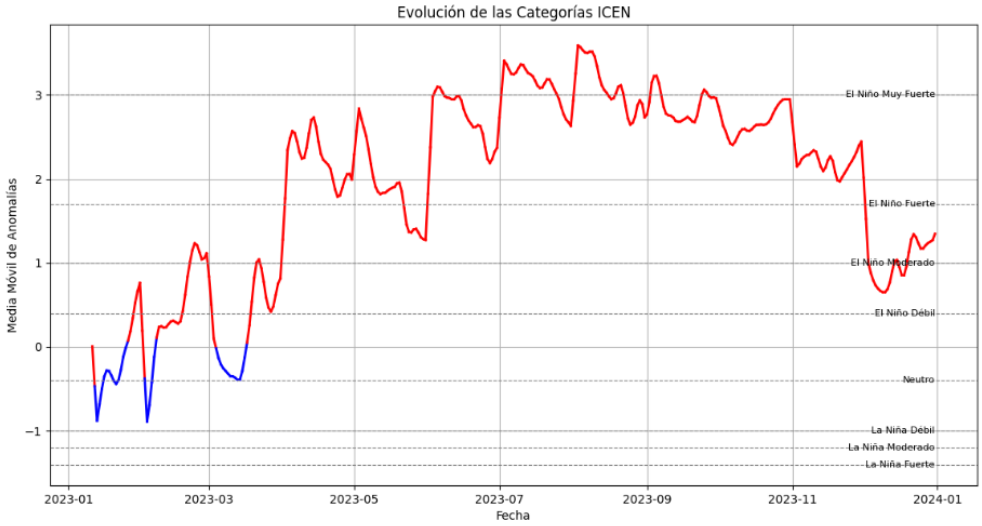

Figura 4.8: ICEN predicho por el modelo Red Neuronal Recurrente.

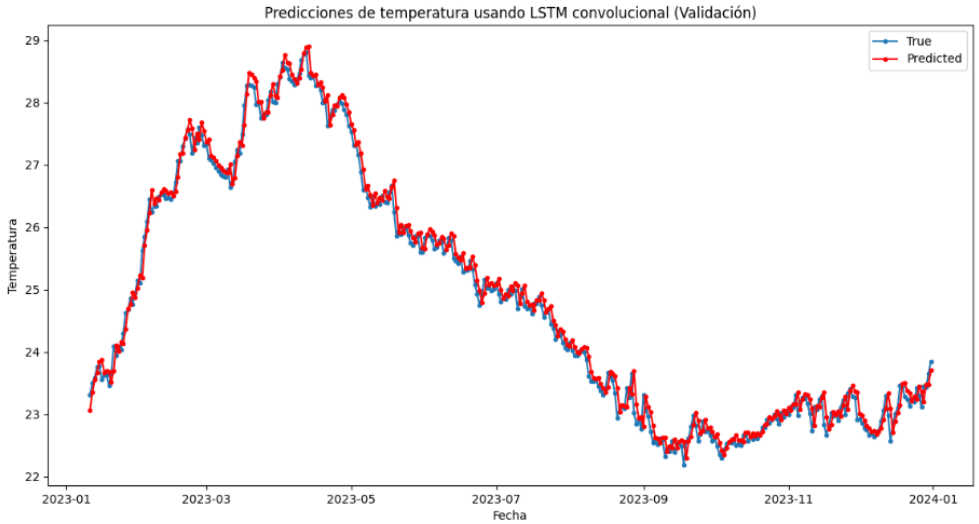

Figura 4.9: TSM predicha por el modelo Red Convolucional vs. mediciones reales.

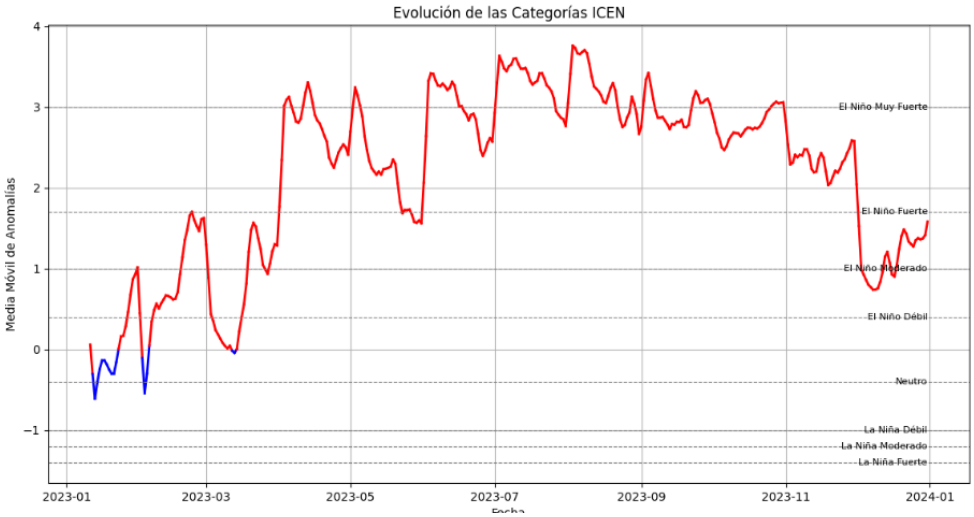

Figura 4.10: ICEN predicho por el modelo Red Convolucional.

4.2. Proyecciones 2024 y 2025

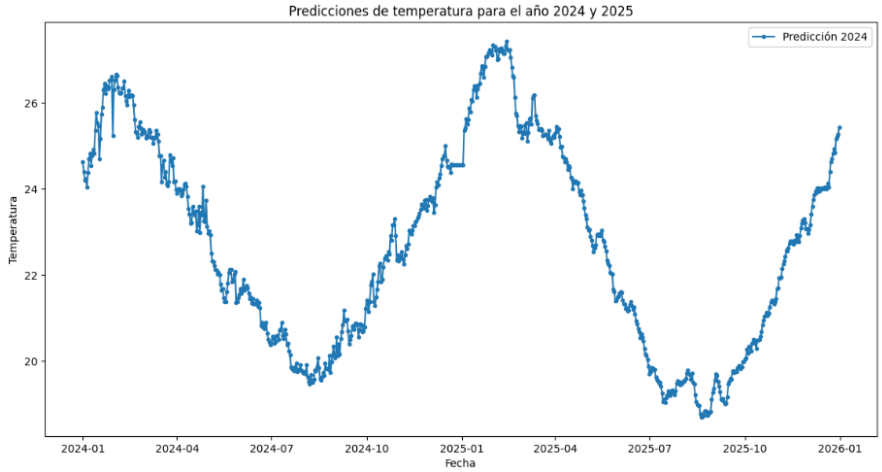

Figura 4.11: Proyecciones de la Temperatura Superficial del Mar del 2024 y 2025 con CNN.

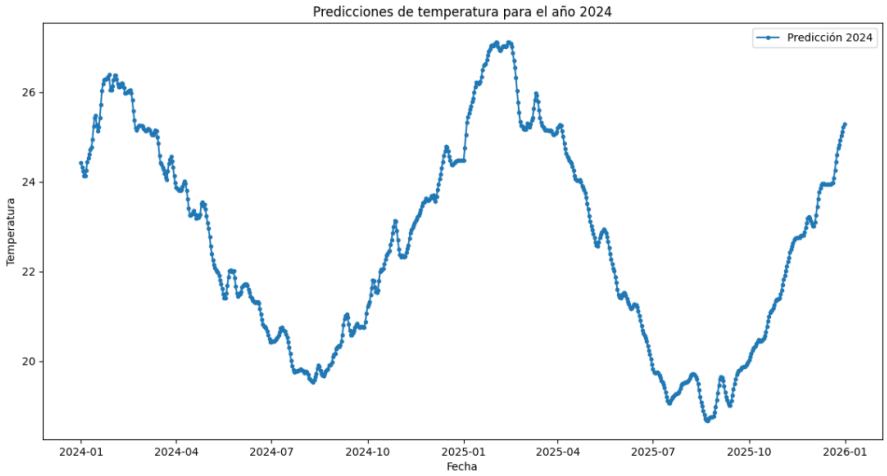

Figura 4.12: Proyecciones de la Temperatura Superficial del Mar del 2024 y 2025 con RNN.

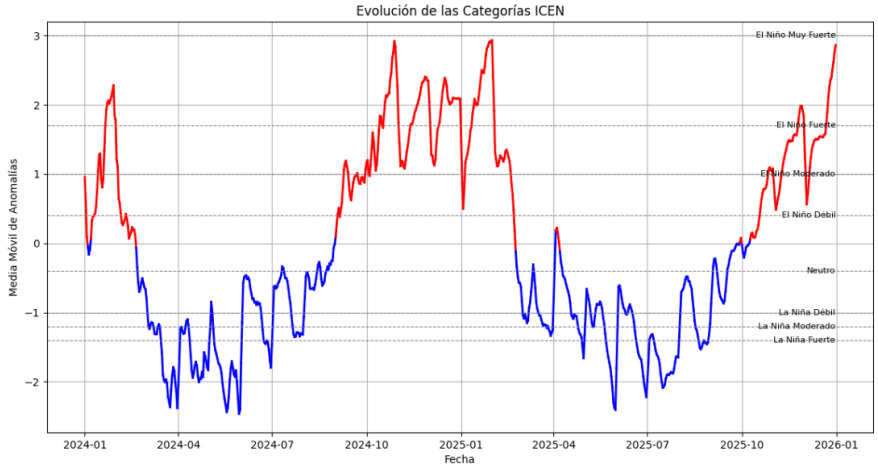

Figura 4.13: Proyecciones del ICEN del 2024 y 2025con CNN.

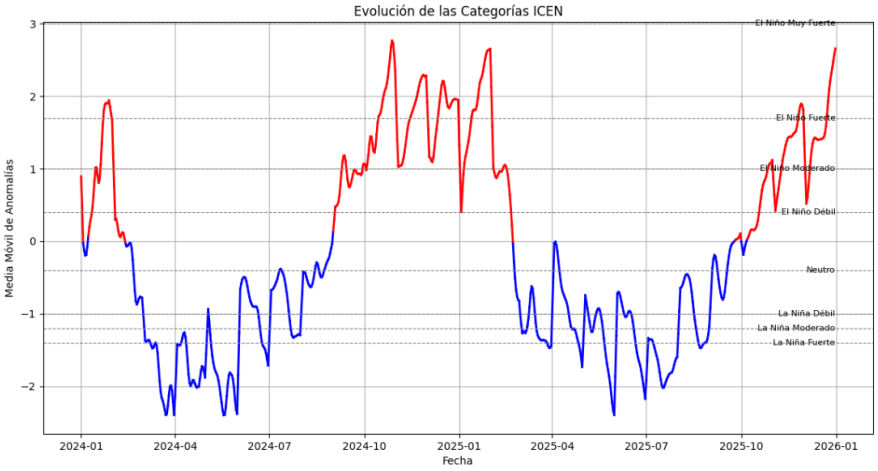

Figura 4.14: Proyecciones del ICEN del 2024 y 2025 con RNN.

Capítulo 5

Análisis de Resultados

5.1. Rendimiento y precisión de los Modelos

Basándonos en las métricas obtenidas según la tabla 4.1 y los gráficos de cada modelo, podemos concluir lo siguiente:

1. **Árbol de regresión**

    - El RMSE en el conjunto de prueba es de 2.725, indicando una discrepancia considerable entre las predicciones del modelo y los datos reales, lo cual podría afectar la precisión del pronóstico, especialmente en condiciones extremas de variabilidad climática como se describe en la clasificación ICEN de la tabla 2.2.
    - Como se observa en el gráfico 4.3, el modelo predice de manera aceptable en los primeros 2 meses, pero luego se aleja gradualmente de los datos reales, sugiriendo que este modelo puede ser más adecuado para pronósticos a corto plazo.

2. **XGBoost**

    - El RMSE es alto, aproximadamente 2.715, lo que sugiere un rendimiento inferior del modelo XGBoost en comparación con otros modelos en términos de precisión de predicción.
    - Dado que XGBoost requiere ajustes meticulosos de hiperparámetros para evitar sobreajuste y mejorar su rendimiento, sería beneficioso dedicar más tiempo a este proceso para obtener mejores resultados, dado que no se esta consiguiendo los resultados esperados a pesar

que este modelo es óptimo para captar patrones temporales como las mostradas en este estudio.

- La figura 4.5 muestra que el modelo predice razonablemente bien en los primeros 2 meses, pero se desvía de los datos reales posteriormente.

3. **Random Forest**

   - El RMSE en el conjunto de prueba es muy bajo, alrededor de 0.198 y un MSE de 0.0395, indicando que el modelo generaliza bien en datos no vistos.
   - La figura 4.1 muestra que las predicciones son muy precisas para todo el año 2023.

4. **Red Neuronal Convolucional**

   - El RMSE es bastante bajo, aproximadamente 0.18 asi como un MSE de 0.03, lo que indica un buen rendimiento del modelo LSTM en términos de precisión de predicción.
   - La figura 4.7 muestra que el modelo sigue de cerca los datos reales, demostrando alta eficiencia en la predicción.

5. **Red Neuronal Recurrente**

   - El RMSE es 0.198, mostrando que el modelo generaliza bien en datos no vistos durante la validación.
   - La figura 4.9 muestra que el modelo también predice con alta precisión la evolución de los datos reales.

5.2. Pronóstico del ICEN y el ENSO

El IMARPE [20] ha calculado el ICEN histórico utilizando datos del NOAA ERRSR.v5 [51], como se muestra en la figura 5.1. Cabe aclarar que este calculo es para toda la region Niño 1+2 (Ecuador y Perú) y nuestro calculo es solo para la costa norte del país.

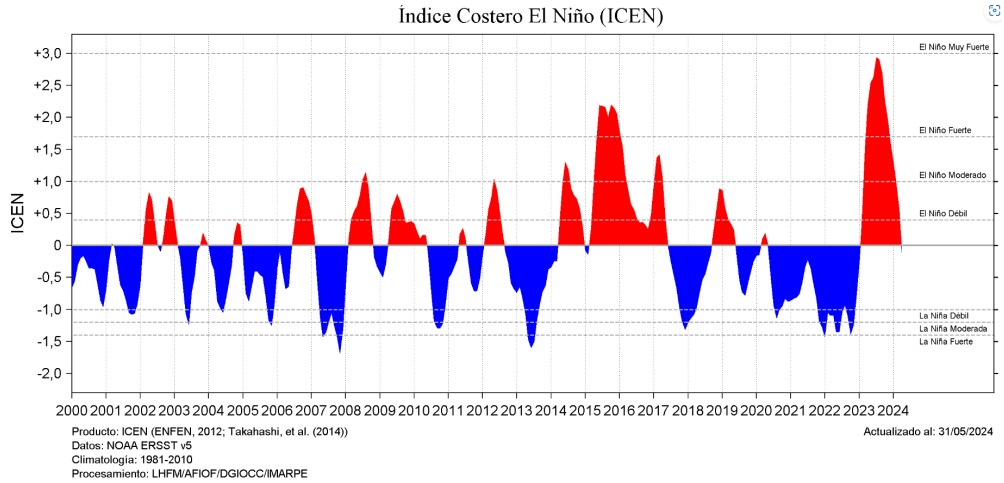

Figura 5.1: ICEN desde el 2000 hasta 2024 según IMARPE [20].

- El calculo del ICEN cos modelos que tienen un RMSE aceptable (ver figuras 4.2, 4.8, 4.10) muestran una aproximación al ICEN calculado con una mayor cantidad de datos históricos. Sin embargo, presentan picos de .El Niño muy fuerte", posiblemente debido a que las temperaturas promedio mensuales se calcularon con solo 7 años de datos, mientras que el gráfico de IMARPE(figura 5.1) utiliza más de 20 años de datos históricos, además que nosotros solo estamos utilizando datos de la Costa Norte del Perú mientras que el IMARPE utiliza datos promedios de la region 1+2 (Ecuador y Perú).

- Según el NOAA (National Oceanic and Atmospheric Administration) [50] y de la Comisión Multisectorial del Estudio Nacional del Fenómeno El Niño (ENFEN) en sus boletines 03-2023 [59] y 21-2023 [60], la climatología del ENSO ha variado entre Niño débil a principios del año 2023, pasando a Niño fuerte hacia finales de año con una alerta de vigilancia para principios de 2024. Esto tambien es mostrado por los modelos cuyo RSME sob bajos (Random Forest, CNN y RNN) y confirma la efectividad del modelo para el pronóstico del ciclo ENSO.

5.3. Proyecciones para el 2024 y 2025

Según las figuras 4.13 y 4.14, se espera que la fase cálida del ENSO se mantenga hasta abril de 2024 y comience a disminuir en abril de 2024, dando

paso a la fase fría del ENSO conocida como La Niña a partir de junio hasta octubre del presente año.

Contrastando estos resultados con lo mencionado el informe del ENFEN [62], que indica que el evento niño costero en la zona 1+2 culminó entre finales de abril e inicios de mayo del 2024, pasando a una fase neutra del ENSO, nos indica que las proyecciones para el año 2024 han sido muy certeras. a esto hay que sumarle las proyecciones de la NOAA [52], en donde se indica que La Niña costera deberia comenzar en julio de este año, lo cual coincide a su vez con lo emitido por el SENAMHI como una alerta de La Niña costera [61], coincidiendo ambos organismos con las predicciones de los modelos aqui mostrados.

Para el año 2025, se proyectan temperaturas superficiales del mar más altas que las proyectadas para 2024 (Figuras 4.11 y 4.12), lo que podría indicar un potencial evento catalogado como "Niño fuerte "hasta una fase de "Niño extraordinario ", lo que podría llevar a un eventual fenómeno del niño costero, el cual podría traer implicancias desastrosas en nuestro país.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

- Rendimiento del modelo: los modelos evaluados, incluidos árboles de regresión, XGBoost, bosques aleatorios, redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN), mostraron diferencias significativas en la capacidad predictiva. Los modelos Random Forest, las redes neuronales convolucionales y recurrentes mostraron un ajuste más preciso y consistente en la predicción de los cambios ENSO, en específico, las redes convolucionales tuvieron un mejor desempeño, seguido de las redes recurrentes, posteriormente el modelo random forest y finalmente estan los modelos que no obtuvieron un desempeño óptimo como XGBoost y Árboles de Regresión, en ese orden.
- Proyecciones para 2024 y 2025: Se espera que ENSO pase a una fase fría a partir de mediados de 2024 e iniciando con una fase cálida fuerte a inicios del 2025, según las proyecciones del modelo.
- Aplicabilidad de los resultados: Los resultados obtenidos resaltan la utilidad de los modelos de inteligencia artificial en la predicción y gestión temprana del fenómeno El Niño ENSO. Estos hallazgos proporcionan una base sólida para la futura implementación de sistemas de alerta temprana y pueden facilitar una mejor planificación y respuesta a eventos climáticos extremos, ayudando así a reducir los impactos socioeconómicos en las comunidades costeras del Perú.

6.2. Trabajo Futuro

El trabajo que se recomienda hacer consiste en ampliar el estudio en los siguientes aspectos:

1. Ampliar la base de datos ya obtenida del Instituo Humbolt por la base de datos del NOAA ersst.v5 que tiene una data historia de los años 1990 aproximandamente.

2. Hacer un estudio por regiones de la costa norte del País (Tumbes, Piura, Lambayeque, entre otras) para estudiar el avance de las TSM anómalas, ademas que el promediar las temperaturas de toda la costa peruana puede incurrir a un sesgo estadistico debido a que en unas regiones del norte, la temperatura promedio suele ser mal altas que en otras regiones.

3. Hacer un mejor tratamiento a los hiperparametros de los diferentes modelos para ajustar la precisión que estos tienen para predecir el ENSO, especialmente en los modelos que no han presentado una eficiencia satisfactoria.

4. Hacer una comparación con respecto a los modelos de machine learning previamente encontrados en la literatura y metodos tradicionales usados para el pronostico de la temperatura superficial del mar, el ICEN y el OSI.

5. Añadir al analisis realizado las caracteristicas del mar peruano.

Bibliografía

[1] Bagging. https://www.ibm.com/topics/bagging#:~:text=Bagging%2C%20also%20known%20as%20bootstrap,be%20chosen%20more%20than%20once. Consultado en junio de 2024.

[2] Boosting. https://aws.amazon.com/es/what-is/boosting/#:~:text=Boosting%20creates%20an%20ensemble%20model,input%20to%20the%20next%20tree. Consultado en junio de 2024.

[3] Boosting. https://www.ibm.com/topics/boosting. Consultado en junio de 2024.

[4] Pycharm. https://datascientest.com/es/pycharm. Consultado el 29 de junio de 2024.

[5] Scikit-learn Documentation: sklearn.preprocessing.minmaxscaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html. Accessed: 2024-06-29.

[6] TensorFlow Documentation: tf.keras.layers.dense. https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dense. Accessed: 2024-06-29.

[7] TensorFlow Documentation: tf.keras.layers.lstm. https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM. Accessed: 2024-06-29.

[8] TensorFlow v2.16.1 documentation: tf.keras.layers.conv1d. https://www.tensorflow.org/api_docs/python/tf/keras/layers/Conv1D. Accessed: 2024-06-29.

[9] Sub-seasonal climate prediction in the upper colorado river basin using random forest classifier. Technical report, Center for Climate Prediction, 2022. Disponible en https://www.coaps.fsu.edu/bios/98-4.pdf.

[10] 4geeks. Boosting de algoritmos. https://4geeks.com/es/lesson/boosting-de-algoritmos. Consultado en junio de 2024.

[11] Físicas y Naturales Academia Colombiana de Ciencias Exactas. *Los fenómenos de El Niño y de La Niña, su efecto climático e impactos socioeconómicos*. Number 34 in Colección Jorge Álvarez Lleras. Academia Colombiana de Ciencias Exactas, Físicas y Naturales, 2017.

[12] Amazon Web Services. What is a neural network?, 2024. Accessed: 2024-06-28.

[13] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of xgboost. 2024.

[14] Xiaoqun Cao, Yanan Guo, Bainian Liu, Kecheng Peng, Guangjie Wang, and Mei Gao. Enso prediction based on long short-term memory (lstm). *IOP Conf. Series: Materials Science and Engineering*, 799:012035, 2020.

[15] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[16] Ciencia de Datos. Modelar tendencia en series temporales: modelos de árboles, sin fecha.

[17] Grupo Intergubernamental de Expertos sobre el Cambio Climático. Introducción a los modelos climáticos simples utilizados en el segundo informe de evaluación del ipcc. https://archive.ipcc.ch/pdf/technical-papers/paper-II-sp.pdf, 1997. OMM - ONU.

[18] Organización Mundial de la Salud. El niño-southern oscillation (enso). https://www.who.int/es/news-room/fact-sheets/detail/el-nino-southern-oscillation-(enso)#:~:text=El%20ENOS%

20se%20da%20en,Ni%C3%B1a%20y%20una%20fase%20neutra. Accedido: 2024-06-27.

[19] Ministerio del Ambiente de Perú. ¿qué es el niño y qué factores determinan su intensidad? evolución de la definición de el niño. https://www.minam.gob.pe/fenomenodelnino/que-es-el-nino-y-que-factores-determinan-su-intensidad/evolucion-de-la-definicion-de-el-nino/. Accedido: 2024-06-27.

[20] Instituto del Mar del Perú (IMARPE). El niño y la oscilación del sur, 2024. Accedido: 2024-06-28.

[21] Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40:139–157, 2000. Manufactured in The Netherlands.

[22] Esri. How xgboost works, 2024. Accessed: 2024-06-28.

[23] GeeksforGeeks. Decision tree, 2024. Accessed: 2024-06-28.

[24] H2O.ai. Weights and biases in neural networks, 2024. Accessed: 2024-06-28.

[25] Hacker Noon. Neural network layers: All you need is an inside comprehensive overview, 2024. Accessed: 2024-06-28.

[26] IBM. Decision trees, 2024. Accessed: 2024-06-28.

[27] IBM. Neural networks, 2024. Accessed: 2024-06-28.

[28] IBM. Regression trees procedures, 2024. Accessed: 2024-06-28.

[29] IBM. What is machine learning?, 2024. Accessed: 2024-06-28.

[30] Domor Mienye Ibomoiye and Yanxia Sun. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa*, 2010.

[31] ATRIA Innovation. ¿qué son las redes neuronales y sus funciones?, 2024. Accessed: 2024-06-28.

[32] Instituto Humboldt de Investigación Marina y Acuícola (IHMA). Temperatura superficial del mar. https://ihma.org.pe/temperatura-superficial-del-mar/. Consultado en junio de 2024.

[33] Instituto Humboldt de Investigación Marina y Acuícola (IHMA). Temperatura superficial del mar - base de datos. https://ihma.org.pe/temperatura-superficial-del-mar/. Consultado el 29 de junio de 2024.

[34] International Research Institute for Climate and Society (IRI). Enso phases: Neutral, 2024. Accessed: 2024-06-28.

[35] Ferrer. F. J. El fenómeno de la oscilación del sur-el niño (enso). https://fjferrer.webs.ull.es/Apuntes3/Leccion05/1_el_fenmeno_de_la_oscilacin_del_surel_nio_enso.html. Accedido: 2024-06-27.

[36] JetBrains IDEs. Pycharm. https://www.jetbrains.com/es-es/pycharm/. Consultado el 29 de junio de 2024.

[37] Sunpreet Kaur and Sonika Jindal. A survey on machine learning algorithms. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, 3, 2016. Accessed: 2024-06-28.

[38] Keras. Keras documentation - dense layer, 2024.

[39] Keras. Keras documentation - dropout layer, 2024.

[40] Keras. Keras documentation - lstm layer, 2024.

[41] Dong-Hoon Kim, Il-Ju Moon, Chaewook Lim, and Seung-Buhm Woo. Improved prediction of extreme enso events using an artificial neural network with weighted loss functions. 2023.

[42] Will Koehrsen. Random forest: A simple explanation, 2018. Accessed: 2024-06-29.

[43] KW Foundation. Redes neuronales recurrentes, 2021. Accessed: 2024-06-28.

[44] Jenny Maturana, Mónica Bello, and Michelle Manley. Antecedentes históricos y descripción del fenómeno el niño, oscilación del sur. history and description of "el niño, southern oscillation" phenomenon, 2024.

[45] Glenn R. McGregor and Kristie Ebi. El niño southern oscillation (ENSO) and health: An overview for climate and health researchers. *Review*, 2020. Department of Geography, Durham University, Durham DH1 3LE, UK and Department of Global Health, Centre for Global Health and Environment, University of Washington, Seattle, WA 98105, USA; krisebi@uw.edu.

[46] Medium. Lstm diagram, 2024. Accessed: 2024-06-28.

[47] Ministerio del Ambiente, Senami. El fenómeno del niño en el perú. `https://www.minam.gob.pe/wp-content/uploads/2014/07/Dossier-El-Ni%C3%B1o-Final_web.pdf`, 2014. Páginas 24–30.

[48] National Centers for Environmental Information (NCEI). El niño/southern oscillation (enso) - sea surface temperature (sst) anomalies. `https://www.ncei.noaa.gov/access/monitoring/enso/sst`, 2024. Accessed: 2024-07-01.

[49] National Centers for Environmental Information (NCEI). El niño/southern oscillation (enso) - ssouthern oscillation index (soi). `https://www.ncei.noaa.gov/access/monitoring/enso/soi`, 2024. Accessed: 2024-07-01.

[50] National Weather Service. El niño, 2023. Accessed: 2024-06-30.

[51] NOAA PSL. Noaa extended reconstructed sea surface temperature (ersst), version 5, 2024. Accessed: 2024-06-30.

[52] NOAA's Climate Prediction Center. El niño/southern oscillation (enso) diagnostic discussion, 2024. Accessed: 2024-06-30.

[53] PyCon España. Pycon españa 2022 - charlas, 2022.

[54] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, 2016(1):Article 35, 2016.

[55] Lior Rokach. Decision forest: Twenty years of research. *Information Fusion*, 27:111–125, 2016.

[56] Lior Rokach and Oded Maimon. Decision trees. In *Data Mining and Knowledge Discovery Handbook*, chapter 9, pages 165–192. Springer, New York, NY, 2008. Department of Industrial Engineering, Tel-Aviv University, liorr@eng.tau.ac.il, maimon@eng.tau.ac.il.

[57] Lior Rokach and Oded Maimon. Decision trees, 2018. Accessed: 2024-06-28.

[58] Minu Rose and Hani Ragab Hassen. A survey of random forest pruning techniques. *Department of Mathematical and Computer Sciences, Heriot Watt University*, 2024.

[59] SENAMHI. Comunicado enfen n° 03-2023, 2023. Accessed: 2024-06-30.

[60] SENAMHI. Comunicado enfen n° 21-2023, 2023. Accessed: 2024-06-30.

[61] Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI). Fenómeno el niño, 2024. Accessed: 2024-06-30.

[62] Servicio Nacional de Meteorología e Hidrología del Perú (SENAMHI). Informe sobre el fenómeno el niño, 2024. Accessed: 2024-06-30.

[63] Cosma Rohilla Shalizi. Lecture 10: Maximum likelihood estimation, 2006. Accessed: 2024-06-28.

[64] Pascal Terray Sébastien Dominiak. Improvement of enso prediction using a linear regression model with a southern indian ocean sea surface temperature predictor. *Geophysical Research Letters*, 32:L18702, 2005. ffhal-00124054f.

[65] K. E. Trenberth. The definition of el niño. *Bulletin of the American Meteorological Society*, 78(12):2771–2777, 1997.

[66] Kevin E. Trenberth, David P. Stepaniak, and Julie M. Caron. The present and future of el niño and la niña. *Climate Dynamics*, 42(3-4):225–239, 2013.

[67] Kennedy Richard Gomez Tunque. Pronóstico enos en las regiones niño 3.4 y niño 1+2, utilizando redes neuronales profundas con secuencias espacio temporales. Tesis de maestría, Universidad Nacional Agraria La Molina, 2024. Tesis para optar el grado de Magister Scientiae en Recursos Hídricos.

[68] UC Berkeley School of Information. What is machine learning?, 2024. Accessed: 2024-06-28.

[69] Author(s) Unknown. Hybrid machine learning approach for construction cost estimation: An evaluation of extreme gradient boosting model. *ResearchGate*, 2024. Accessed: 2024-06-28.

[70] Uribe. Predictive model of the enso phenomenon based on regression trees. *MENDEL – Soft Computing Journal*, 29(1), June 2023.

[71] Cal W., Wu L., Lengaigne M., Li T., Mcgregor S., and Kug J. S. Interacciones climáticas pantropicales. *Ciencia*, 363(6430):eaav4236, 2019.

[72] J. M. Wallace and P. V. Hobbs. *Atmospheric Science: An Introductory Survey*. Academic Press, 2006.

[73] Y. Wang et al. Forecasting ENSO using convolutional LSTM network with improved attention mechanism and models recombined by genetic algorithm in CMIP5/6. *Information Sciences*, 642:119106, 2023.

[74] Wikipedia. El niño-oscilación del sur. `https://es.wikipedia.org/wiki/El_Ni%C3%B1o-Oscilaci%C3%B3n_del_Sur`, 2024. Accedido: 2024-06-27.

[75] Wikipedia contributors. Artificial neuron, 2024. Accessed: 2024-06-28.

[76] Jining Yan, Lin Mu, Lizhe Wang, Rajiv Ranjan, and Albert Y. Zomaya. Temporal convolutional networks for the advance prediction of ENSO. *Scientific Reports*, 2020.