INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
MAESTRÍA EN SISTEMAS COMPUTACIONALES

CONTROL DE CALIDAD DE PERFILES HIDROGRÁFICOS BASADO EN REDES NEURONALES ARTIFICIALES

QUE PARA OBTENER EL GRADO DE
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:
MARCOS DANIEL FONG GONZÁLEZ

DIRECTORES DE TESIS:
MARCO ANTONIO CASTRO LIERA
LEONARDO TENORIO FERNÁNDEZ

LA PAZ, BAJA CALIFORNIA SUR, MÉXICO, AGOSTO 2024.

La Paz, B.C.S., 19/ AGOSTO /2024

DEPI/152/2024

ASUNTO: Autorización de impresión

C. MARCOS DANIEL FONG GONZÁLEZ,
ESTUDIANTE DE LA MAESTRÍA EN
SISTEMAS COMPUTACIONALES,
P R E S E N T E .

Con base en el dictamen de aprobación emitido por el Comité Tutorial de la Tesis denominada: **"CONTROL DE CALIDAD DE PERFILES HIDROGRÁFICOS BASADO EN REDES NEURONALES ARTIFICIALES"**, mediante la opción de tesis (Proyectos de Investigación), entregado por usted para su análisis, le informamos que se **AUTORIZA** la impresión.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®

JUDITH GUADALUPE MARTÍNEZ TIRADO,
JEFA DE LA DIV. DE ESTUDIOS DE POSGRADO E INV.

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA
TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

c.c.p. Depto. de Servicios Escolares
c.c.p. Archivo.

*JGMT/icl**

DICTAMEN DEL COMITÉ TUTORIAL

La Paz, B.C.S., **19/AGOSTO/ 2024**

JUDITH GUADALUPE MARTÍNEZ TIRADO,
JEFA DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN,
P R E S E N T E.

Por medio del presente, enviamos a usted dictamen del Comité Tutorial de tesis para la obtención del grado de Maestro, con los siguientes datos generales:

No. de Control	Nombre
M22310005	MARCOS DANIEL FONG GONZÁLEZ
Maestría en:	SISTEMAS COMPUTACIONALES
Título de la tesis: CONTROL DE CALIDAD DE PERFILES HIDROGRÁFICOS BASADO EN REDES NEURONALES ARTIFICIALES	
DICTAMEN: Se autoriza el trabajo de investigación, en virtud de que realizó las correcciones correspondientes conforme a las observaciones plantcadas por este Comité Tutorial.	

A t e n t a m e n t e.
El Comité Tutorial

MSC. ILIANA CASTRO LIERA

MSC. EMMANUEL ROMERO QUINTERO

DR. LEONARDO TENORIO FERNÁNDEZ

DR. MARCO ANTONIO CASTRO LIERA

c.c.p. Coordinador de la Maestría.
c.c.p. Departamento de Servicios Escolares.
c.c.p. Estudiante.

Boulevard Forjadores de B.C.S. #4720, Col. 8 de octubre 1ra Sección, C.P. 23080, La Paz, B.C.S. Tel. (612) 12 1-04-24 e-mail: depi_paz@tecnm.mx | www.lapaz.tecnm.mx

Dedicatoria

A mis padres Luz Berenice y Marcos Kakuei, y a mi hermana Marianne Abril, quienes me dieron las fuerzas necesarias para salir adelante cuando sentía que ya no podía dar más.

Son la razón de todos mis logros.

Agradecimientos

A mis padres y a mi hermana, por haberme apoyado incondicionalmente durante la maestría.

A los miembros de mi comité tutorial. Al Dr. Marco Antonio Castro Liera, por haber estado conmigo en cada etapa del trayecto, ayudándome a superar cada obstáculo que se presentaba. A la M.S.C. Iliana Castro Liera, quien propició esta increíble colaboración entre el ITLP y CICIMAR. Al Dr. Leonardo Tenorio Fernández, por presentarme a un equipo de trabajo tan extraordinario, además de ofrecer su punto de vista desde el ámbito de la Oceanografía, haciéndome ver cosas que de otra forma no habría podido ver. Al M.S.C. Emmanuel Romero, por todos los consejos que me dio y el conocimiento que me prestó para sacar adelante la tesis.

Al Consejo Nacional de Humanidades Ciencia y Tecnología (CONAHCyT) por la beca que me otorgaron, la cual me ayudó a sacar adelante este estudio.

Al Instituto Tecnológico de La Paz (ITLP) y al Centro Interdisciplinario de Ciencias Marinas (CICIMAR) por haber hecho posible la realización de este estudio.

Resumen

El grupo internacional Argo Floats realiza mediciones de los océanos empleando perfiladores hidrográficos autónomos. Los datos, producto de dichas mediciones, son analizados por un sistema automatizado que les asigna etiquetas de acuerdo a su calidad y posteriormente, son puestos a disposición de la comunidad científica. Más tarde, son revisados por especialistas a través del control de calidad en modo diferido, donde se realizan correcciones. Analizar los perfiles en modo diferido resulta difícil debido al aumento de perfiles medidos mensualmente, por lo que este estudio propone una metodología que utiliza redes neuronales artificiales (RNA) para clasificar los datos hidrográficos de acuerdo con su nivel de calidad, para posteriormente estimar los valores ajustados de los mismos datos para mejorar su calidad, usando RNA de regresión. Dicha metodología permite realizar un balanceo de la distribución de los datos, para después desarrollar modelos de RNA capaces de clasificarlos en base a su calidad, logrando un nivel de exactitud promedio de 90 % para temperatura y presión, y 80 % para salinidad. Adicionalmente, la metodología hace posible crear RNA de regresión que generan predicciones de los valores ajustados de los datos hidrográficos con un valor de correlación R^2 superior a 0.99 para las tres variables, y un valor de raíz de error cuadrático medio de 0.0005, 0.0136 y 0.1298 para temperatura, salinidad y presión respectivamente. Gracias a estos resultados, se propone la metodología como una alternativa viable, cuando no se cuenta con datos procesados por el control de calidad en modo diferido.

Abstract

The international Argo Floats group makes ocean measurements using autonomous hydrographic profilers. The data, the product of these measurements, is analyzed by an automated system that assigns labels according to their quality levels, and is subsequently made available to the scientific community. Later, it is reviewed by specialists through quality control in delayed mode, where corrections are made. Analyzing the profiles in delayed mode is difficult due to the increase in profiles measured monthly, so this study proposes a methodology that uses artificial neural networks (ANN) to classify hydrographic data according their quality levels, to subsequently estimate the adjusted values of said data using regression ANN. This methodology allows for balancing data distribution, and then developing ANN models capable of classifying them based on their quality level, achieving an average accuracy level of 90 % for temperature and pressure, and 80 % for salinity. Additionally, the methodology makes it possible to create regression ANNs that generate predictions of the fitted values of the hydrographic data with an R^2 correlation value greater than 0.99 for the three variables, and a root mean squared error value of 0.0005, 0.0136 and 0.1298 for temperature, salinity and pressure respectively. Thanks to these results, the methodology is proposed as a viable alternative, when there is no data processed by quality control in delayed mode.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Antecedentes

Abarcando alrededor del 70 % de la superficie del planeta, los océanos resultan ser fuente de una variedad de recursos, así como ser reguladores climáticos, lo que los vuelve vitales para la supervivencia de los seres vivos. La comprensión de los cambios que ocurren en el océano resulta indispensable para abordar los efectos inducidos por el cambio climático [1].

Argo Floats es un grupo internacional que realiza mediciones *in situ* de presión, temperatura y conductividad (salinidad) en la columna de agua de los océanos, empleando perfiladores hidrográficos autónomos (PHA). Los datos recolectados por estos perfiladores son puestos a disposición de la comunidad científica para su libre uso. Los datos son procesados por el control de calidad en tiempo real, el cual es un sistema automatizado que los clasifica de acuerdo con su nivel de calidad, asignándoles sus banderas correspondientes. Más tarde estos datos son analizados por el control de calidad en modo diferido por especialistas ubicados en centros de asimilación de datos, que se encargan de realizar ajustes y/o correcciones sobre los datos [2]. Este control de calidad tiene como objetivo aplicarse dentro de los primeros 12 meses después de su medición.

En la actualidad, se han realizado esfuerzos que han hecho uso de técnicas de aprendizaje automático para trabajar con estos datos hidrográficos, que buscan reducir los tiempos en que se efectúan los controles de calidad, así como reducir el trabajo humano requerido para ello. Tal es el caso de [3], quienes usaron mapas autoorganizados para identificar patrones de error en los perfiles hidrográficos. En [4] también se enfocaron en encontrar anomalías en los datos, con

la diferencia de que su trabajo se enfoca en datos de temperatura *in situ*, empleando el modelo probabilístico *Gaussian Mixture Model*. Por su parte, en [5] demostraron que es posible estimar la calidad de datos desconocidos con buenos resultados al representar la forma de los perfiles hidrográficos como integrales iteradas, para luego realizar aprendizaje supervisado usando el método de firma (signature method). En [6] utilizaron la técnica de análisis de grupos, K-medias, para filtrar derivas de salinidad dentro de polígonos establecidos. Finalmente, en [7] se desarrolló un modelo de aprendizaje automático con redes neuronales artificiales (RNA) para completar series temporales de datos hidrográficos en un polígono dado.

Algunos de los trabajos mencionados anteriormente cuentan con enfoques específicos que limitan su alcance para trabajar con datos hidrográficos. Tal es el caso de [3], quienes se enfocan solamente detectar errores en los datos para mejorar el proceso de clasificación, mas no realizan ajuste de los datos para mejorar su calidad, similar a [4], quienes cuentan con un alcance aún más limitado al enfocarse solamente a los datos de temperatura.

En otros trabajos, como el de [5], es necesario transformar la información de los datos para poder ser procesados, además de que no abordan el tema de mejorar el proceso de clasificación.

1.2. Descripción del problema

Actualmente resulta difícil alcanzar el objetivo de aplicar el control de calidad en modo diferido dentro de los primeros 12 meses después de su medición, debido al aumento de perfiles medidos que ocurre mes tras mes, lo que en consecuencia aumenta la carga de trabajo sobre los recursos humanos de los centros de asimilación de datos [8][9].

Lo anterior deriva en que la comunidad científica cuente con una cantidad disminuida de perfiles hidrográficos con datos válidos disponibles para sus investigaciones.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar una metodología que emplee modelos de RNA tanto para clasificar datos de temperatura, salinidad y presión de perfiles hidrográficos de acuerdo con su calidad, como para estimar sus valores ajustados para mejorar su calidad.

1.3.2. Objetivos específicos

- Delimitar un polígono sobre aguas nacionales dentro del cual se consultarán perfiles hidrográficos medidos y revisados en modo diferido, con los que se trabajarán sus datos de PHA.
- Descargar y analizar los datos de los perfiles de temperatura, salinidad y presión de la base de datos global de PHA de Argo Floats del polígono delimitado.
- Construir e implementar modelos de RNA que clasifiquen la calidad de los perfiles de temperatura, salinidad y presión.
- Construir e implementar modelos de RNA para corregir los datos clasificados como erróneos, dentro de los perfiles de temperatura, salinidad y presión.

1.4. Justificación

La metodología se presenta como una alternativa viable para clasificar datos hidrográficos de temperatura, salinidad y presión de acuerdo con su nivel de calidad, así como estimar los valores ajustados de los mismos datos para mejorar su calidad. Gracias a ello, los investigadores podrán hacer uso de ella para obtener datos hidrográficos con un nivel de calidad altamente confiable y en menor tiempo.

1.5. Limitaciones y alcance

- En caso de querer analizar la calidad de una nueva variable, se deberá construir una nueva RNA.
- La información será recibida a través de archivos separados por comas (CSV).
- El tiempo que le tomará a la RNA procesar los datos estará delimitado por los recursos y características con los que cuente el equipo en el que esté alojada, así como la cantidad de datos de los perfiles hidrográficos con los que cuente.

1.6. Marco teórico

1.6.1. Perfiladores hidrográficos autónomos

El grupo Argo Floats administra el monitoreo que realizan perfiladores hidrográficos autónomos sumergidos en los océanos, los cuales recolectan datos (como salinidad, temperatura y presión). Estos son transmitidos via satélite, para después procesarlos, generar información y distribuirla.

En Argo Floats se monitorean cerca de 4,000 flotadores (Figura 1.1). Su objetivo es establecer un estándar para la recolección de datos al cual puedan apegarse sistemas de observación oceanográficos, permitiendo así brindar información confiable que puede ser usada para investigaciones científicas.

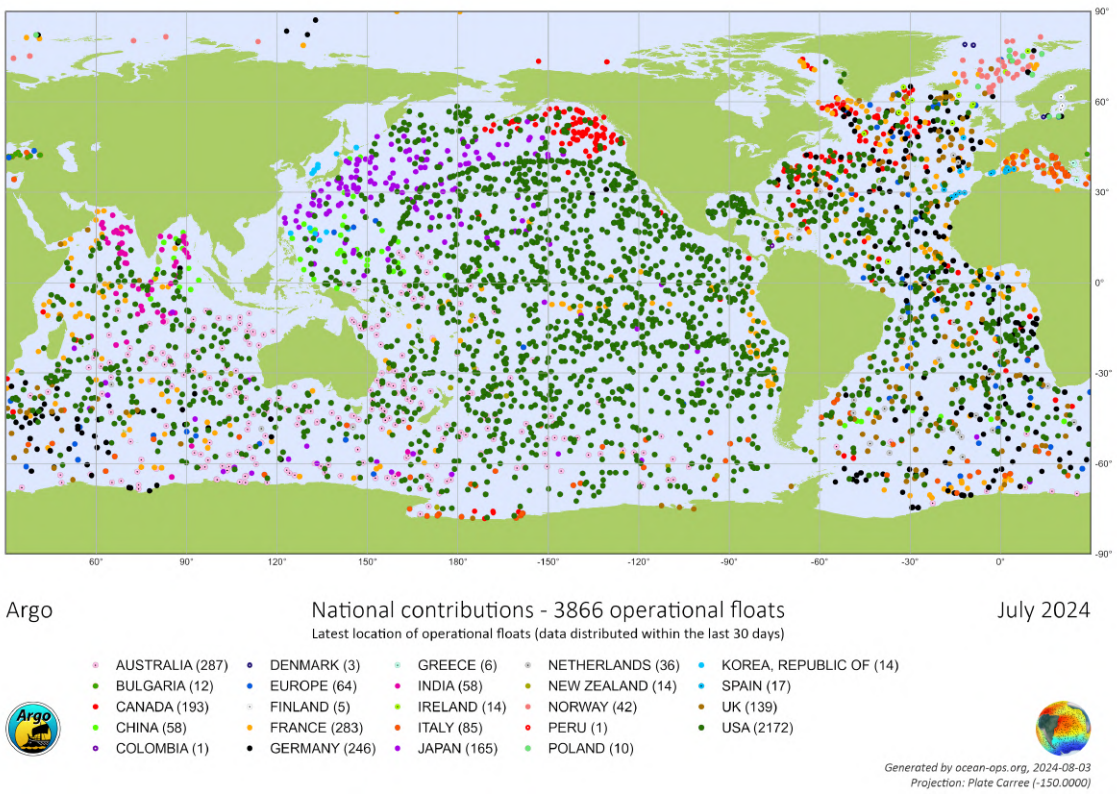


Figura 1.1: Perfiladores hidrográficos autónomos desplegados en el mundo, generado en [10].

El programa es dirigido y gestionado por científicos pertenecientes a los países que forman parte del mismo. Cada país despliega y financia sus perfiladores. Lo anterior es monitoreado por un coordinador técnico de Argo Floats.

Los datos de los perfiladores son utilizados por centros de oceanografía operacional y centros climáticos en diversas aplicaciones como por ejemplo: pronóstico del clima, seguridad naval, transporte marítimo, gestión pesquera, entre otros.

La misión estándar de un PHA consiste en un ciclo de aproximadamente diez días, en el cual la mayoría del tiempo el dispositivo se encuentra flotando a la deriva con las corrientes oceánicas a una profundidad de 1,000 metros durante nueve días. En el décimo día, desciende hasta los 2,000 metros para comenzar su ascenso hacia la superficie, periodo en el cual realiza una serie de mediciones (perfiles) de la columna de agua. Una vez en la superficie, obtiene su posición (por lo general a través de un GPS) y se comunica con un satélite para enviar sus datos y recibir cualquier nueva instrucción de misión. En la mayoría de los casos, el tiempo en la superficie oscila entre 15 minutos y 1 hora. Después de eso, el perfilador vuelve a sumergirse. El ciclo se repite hasta que la batería del perfilador se agota, usualmente cuatro o cinco años después [2].

Una vez recibidos los datos por parte de Argo Floats, sus sistemas automatizados les aplican el control de calidad en tiempo real, donde son clasificados de acuerdo con su nivel de calidad, asignándoles sus etiquetas correspondientes. Tiempo después, los especialistas ubicados en los centros de asimilación de datos, realizan el control de calidad en modo diferido, donde ajustan los datos para mejorar su calidad, así como cambiar las etiquetas asignadas por el control de calidad en tiempo real.

1.6.2. Reconocimiento de patrones

En este estudio se abordaron dos problemas de reconocimiento de patrones en el ámbito de ingeniería, computación y matemáticas, para mejorar el proceso de control de calidad de los PHA.

El primero de ellos fue de **clasificación**, la cual se dividió a su vez en binaria y multiclase. El desempeño de los modelos de clasificación fue evaluado usando las siguientes métricas:

- **Matriz de confusión:** Permite ver los aciertos y errores del modelo (desplegados como verdaderos y falsos positivos y negativos) durante el proceso de aprendizaje automático (Figura 1.2).
- **Exactitud (Accuracy):** Es la cercanía que se encuentra un resultado del valor verda-

dero. Se representa como la proporción de resultados verdaderos (tanto positivos como negativos) entre el total de casos examinados.

- **Precisión:** Dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Es representada como la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos como falsos).

- **Sensibilidad:** Tasa de verdaderos positivos. Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

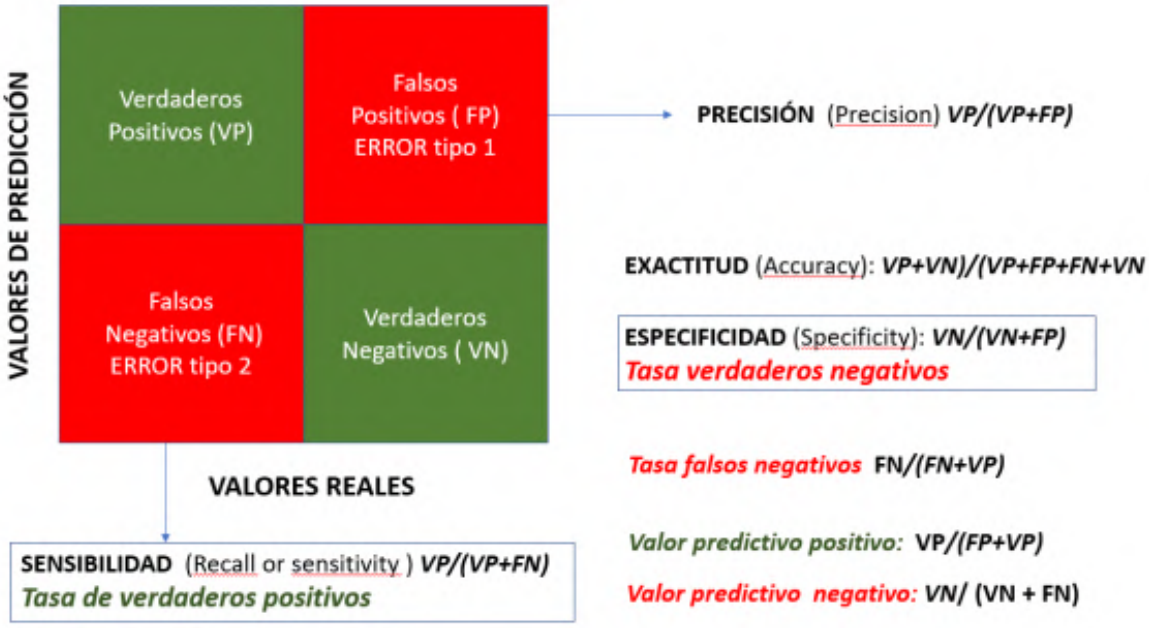


Figura 1.2: Matriz de confusión binaria y sus métricas, consultada en [11].

El segundo problema de reconocimiento de patrones abordado fue el de **regresión**, en el cual la columna objetivo es de tipo numérico.

Las métricas usadas para evaluar los modelos de regresión desarrollados para este estudio son las siguientes [12][13]:

- **Error absoluto medio (MAE):** Permite conocer la diferencia entre los valores predichos y los valores reales u observados.

  $MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$

- **Error cuadrado medio (MSE):** Se utiliza para conocer la cercanía de la línea de ajuste de regresión con las observaciones.

  $MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}$

- **Raíz del error medio cuadrado (RMSE):** Le da más peso a errores grandes debido a la operación de raíz cuadrada. Las unidades de RMSE son las mismas que los datos originales, volviéndola fácilmente interpretable.

  $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}}$

- **Coeficiente de determinación/R cuadrada:** Indica el grado de ajuste con el que cuenta el modelo respecto a las observaciones reales. El mejor valor posible que se puede obtener es 1, mientras que el peor es 0. Se calcula dividiendo la varianza residual con la varianza de la variable dependiente.

  $R^2 = 1 - \frac{\sum(y_i - x_i)^2}{\sum(y_i - \mu_y)^2}$

- **Error máximo (ME):** Se utiliza para conocer el peor caso posible, la diferencia más grande entre un dato observado y su estimación.

  $ME = max(|y_i - x_i|)$

1.6.3. Redes Neuronales Artificiales (RNA)

Una red neuronal artificial es una tecnología que consiste en un modelo creado en una computadora, cuyo objetivo es igualar o adoptar ciertas funciones básicas del cerebro humano [14].

Se constituye de unidades de procesamiento que intercambian datos o información. Una de sus características fundamentales es que tienen la capacidad de aprender y mejorar su funcionamiento. Dado que el propósito de una RNA es parecerse lo más posible a su homólogo biológico, esta presenta características propias del cerebro:

- **Aprender:** Adquiere conocimiento de algo a través de estudio, ejercicio o experiencia. El comportamiento de la RNA puede cambiar de acuerdo a su entorno.

- **Generalizar:** Dentro de cierto margen, la RNA puede ofrecer respuestas correctas a entradas que pueden presentar pequeñas variaciones.

- **Abstraer:** Se puede abstraer la esencia de una serie de entradas que a simple vista pueden no presentar aspectos en común.

En una RNA, la unidad homóloga a la neurona biológica es el perceptrón. Este cuenta con varias entradas y las combina (generalmente) con una suma básica. Dicha suma se modifica por una función de transferencia o activación y el valor resultante pasa a la salida del perceptrón (Figura 1.3). A la suma se le agrega la constante conocida como **bias**, la cual ayuda a los modelos a dirigir la función de activación hacia el lado positivo o negativo.

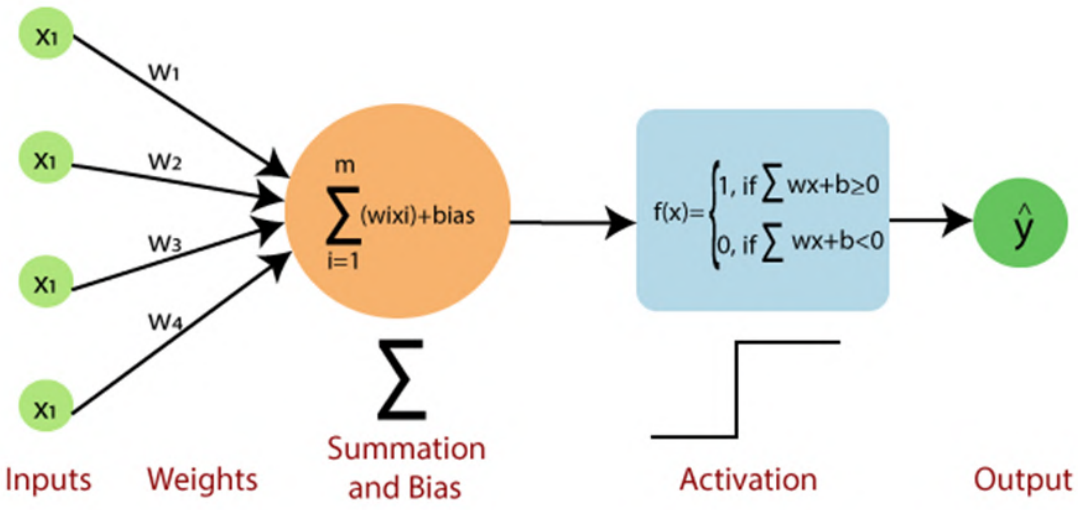


Figura 1.3: Estructura básica de un perceptrón, obtenida de [12].

Los datos de entrada (x) del perceptrón se multiplican por unos parámetros llamados "pesos"(w), antes de ser procesadas por la función sumatoria. La función de activación tiene como objetivo evaluar si el valor resultante traspasa cierto umbral, que a su vez sirve para determinar la salida ($\widehat{y}$) del perceptrón. Para los perceptrones de clasificación, la salida es un número entero positivo. Para los de regresión, el resultado es un valor numérico continuo. La distribución de neuronas dentro de una RNA se realiza formando niveles o capas de un número determinado cada una. (Figura 1.4). Se pueden distinguir tres tipos de capas [15]:

Entrada: Es la capa que recibe directamente la información proveniente de las fuentes externas de la red.

Ocultas: Son internas a la red. El número de niveles oculto puede ser de cero a un número entero positivo determinado. La neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina junto a su número, las distintas topologías.

Salida: Transfieren información de la red hacia el exterior, brindando la respuesta del sistema.

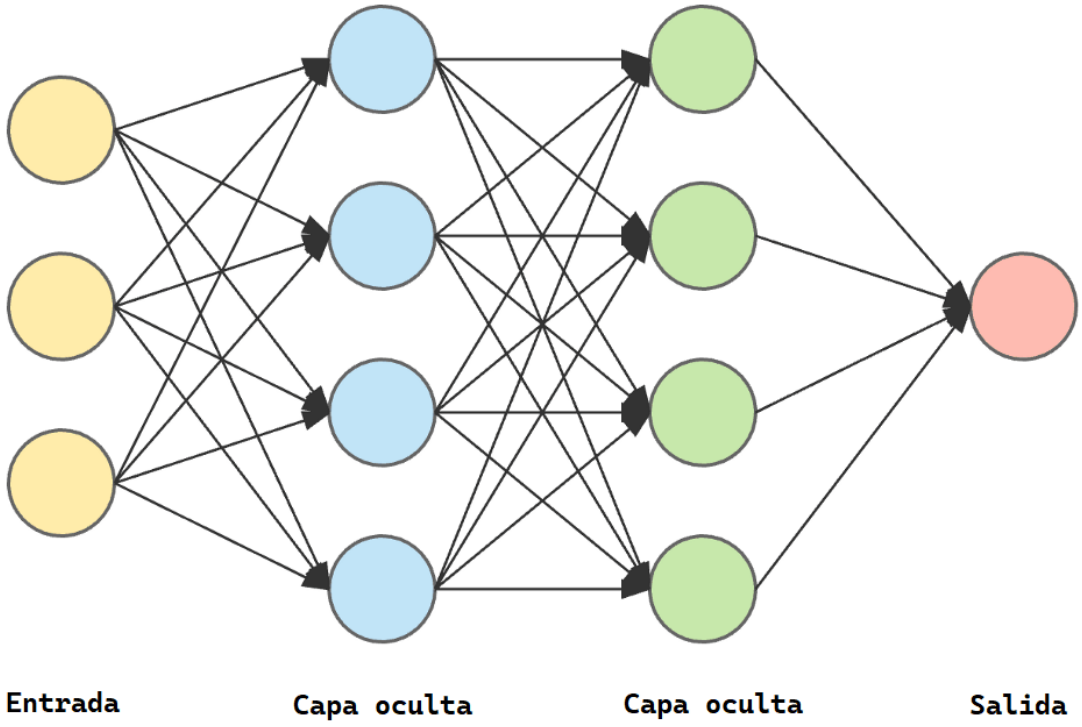


Figura 1.4: Composición de una red neuronal estándar, obtenida de [16].

1.6.3.1. Algoritmo backpropagation

El algoritmo backpropagation es un método usado para el entrenamiento de las RNA. Se destaca por su capacidad de entrenar las capas ocultas [17].

En cada instancia del entrenamiento, el algoritmo se lo retroalimenta la red y calcula la salida de cada neurona en cada capa consecutiva. El siguiente paso del algoritmo es medir el error en la salida de la red y se determina la contribución de cada neurona de la última capa oculta al error de cada neurona en la capa de salida. Luego se mide la cantidad de contribución de error que provino de cada neurona en las capas anteriores de forma repetitiva hasta alcanzar la capa de entrada (Figura 1.5). En esta regresión se mide el gradiente de error con los pesos en

las conexiones de la red, para posteriormente modificar dichos pesos con el objetivo de reducir el error de salida.

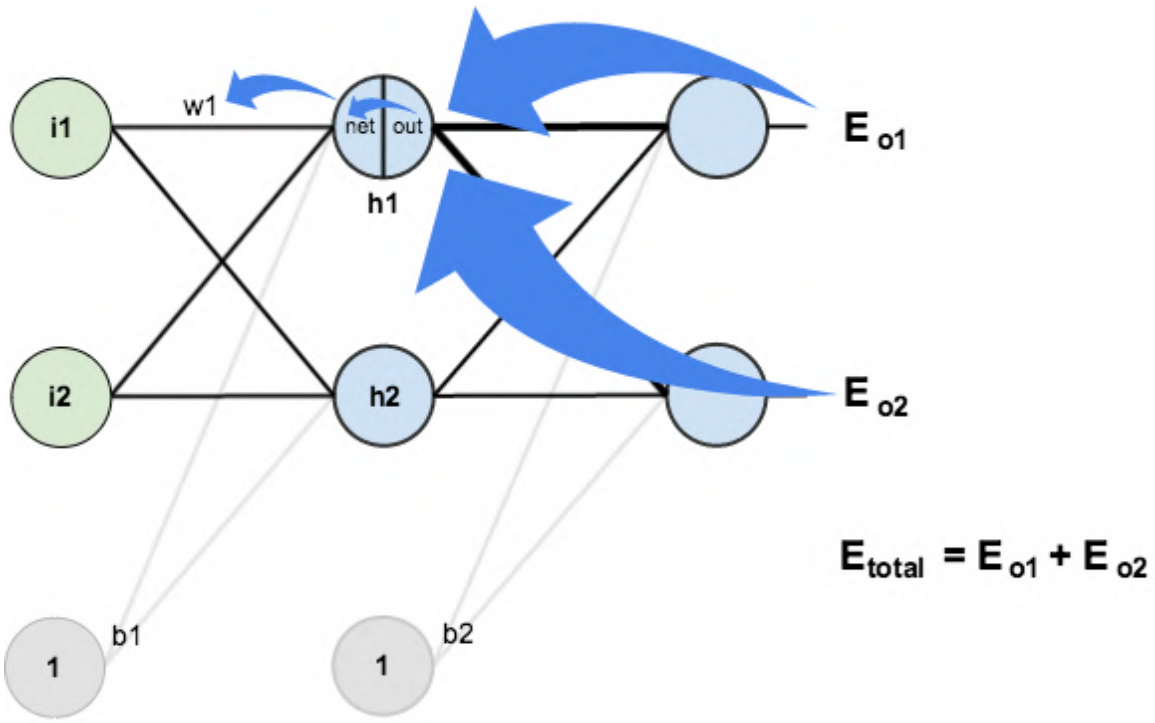


Figura 1.5: Retroalimentación de la RNA con el algoritmo backpropagation, obtenida de [18].

Capítulo 2

Metodología

2.1. Descarga de los perfiles hidrográficos

Se delimitó un polígono que cubre un área del Pacífico oriental frente a México (Figura 3.1), así como un rango de fechas que abarca desde el 1 de enero de 2015 al 31 de diciembre de 2020. Los datos de los perfiles hidrográficos se descargaron en formato netCDF desde el servidor FTP de Argo Floats (ftp://usgodae.org/pub/outgoing/argo, consultado el 30 de junio de 2023). A través del archivo de directorio de perfiles del Centro de ensamblaje de datos global de Argo Floats, "ar_index_global_prof.txt" se realizó la consulta de perfiles medidos dentro del polígono y del espacio temporal establecido, además de haber sido revisados en modo diferido. Después se compilaron los archivos descargados en un solo archivo con el formato separado por comas (CSV) para que la información fuese procesada por los modelos de RNA. Para la creación del polígono y la descarga de datos, se desarrolló un programa basado en la librería cluster_qc [19].

2.2. Identificación de clases

Argo Floats utiliza un sistema de banderas (Tabla 2.1) para etiquetar los datos de acuerdo a su nivel de calidad [2].

No. de bandera	Significado
0	No se realizó control de calidad.
1	Datos de buena calidad.
2	Datos probablemente buenos.
3	Datos de mala calidad potencialmente corregibles.
4	Datos de mala calidad.
5	Valor cambiado.
6	Bandera no usada.
7	Bandera no usada.
8	Valor estimado.
9	Valor faltante.

Tabla 2.1: Banderas de calidad empleadas por Argo.

Para simplificar el procesamiento de los datos, las banderas 1, 2 y 3, conservaron su propia clase, mientras que los datos etiquetados con el resto de banderas se agruparon en la clase 4. Esto con el objetivo de conservar los datos necesarios para realizar el ajuste de los mismos, y filtrar los datos que no pueden ser ajustados. Las nuevas clases se presentan en la Tabla 2.2.

No. de bandera	Significado
1	Datos buenos.
2	Datos probablemente buenos.
3	Datos de mala calidad potencialmente corregibles.
4	Datos no utilizables.

Tabla 2.2: Clases definidas para los modelos de RNA.

2.2.1. Balanceo de clases

Tras un análisis estadístico, se determinó que las clases resultantes se encontraban desbalanceadas (i.e., la clase 1 contenía una mayor cantidad de datos), por lo que fue necesario emplear un modelo de RNA de clasificación binario entrenado para reconocer los datos de la clase 1,

para posteriormente filtrar parte de ellos, y así balancear el conjunto de datos, de forma que se evitara que las RNA de clasificación centraran su atención en la clase mayoritaria.

En la Figura 2.1 se muestra la metodología propuesta, tomando en cuenta el filtrado de datos de la clase 1 mencionado anteriormente. El modelo RNA de clasificación binario en cuestión transformó las cuatro clases definidas para este estudio en dos. La clase 1 conservó su propia clase, mientras que la 2, 3 y 4 se agruparon en la segunda clase. Se usaron los datos originales para entrenar, validar y probar el modelo binario. Una vez probado, se aplicó el filtro 1, que consistió en eliminar parte de los datos que claramente son de buena calidad y no necesitan ajuste (i.e. clase 1) por el modelo binario de los datos originales, balanceando así el conjunto de datos.

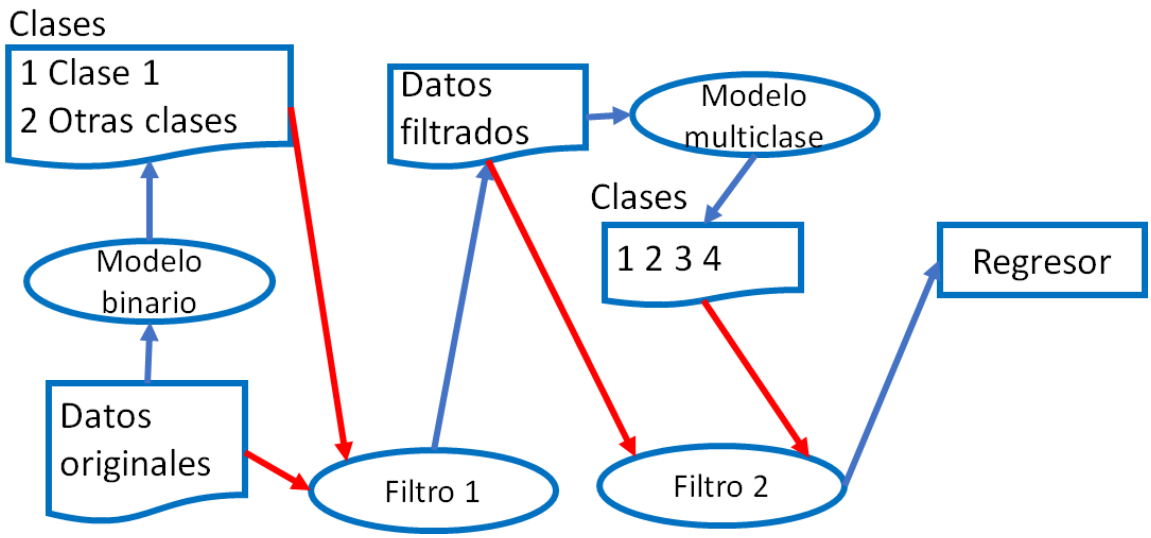


Figura 2.1: Metodología propuesta. Los elementos utilizados para los filtros se conectan con flechas rojas, mientras que las azules indican el proceso de obtención de los datos clasificados.

Los datos filtrados resultantes se usaron para entrenar el modelo de RNA multiclase, que los clasificó en las cuatro clases definidas previamente. Una vez hecha la clasificación, se aplicó el filtro 2, el cual consistió en eliminar los datos clasificados como la clase por defecto (i.e. clase 4) de los datos a los que ya se les había filtrado la clase 1 previamente; lo que resultó en un conjunto de datos balanceado y solamente con los datos necesarios para entrenar, validar y probar el modelo RNA de regresión que ajustó los datos.

2.3. Construcción de los modelos de RNA

Los datos usados por los modelos de RNA fueron divididos en tres conjuntos que fueron usados para entrenar, validar y probar los modelos (64 %, 16 % y 20 % de los datos, consecutivamente). Los datos que se reservaron para las pruebas consistieron en información que no se proporciona a la RNA durante el entrenamiento, por lo que los experimentos representaron el desempeño de los modelos, utilizando datos que no se les presentaron previamente en el entrenamiento.

Para todos los modelos de RNA usados en este estudio (tanto clasificación como regresión) se utilizaron las mismas variables de entrada (datos de las mediciones de los perfiles hidrográficos), seleccionadas por haber probado ser las más representativas y que mejor describían la naturaleza de los datos trabajados. Dichas variables fueron: año, mes, latitud, longitud, temperatura, salinidad y presión. En el caso de la variable de salida de las RNA, se utilizaron los valores ajustados de los tres tipos de datos para sus respectivas RNA. Para la capa de salida, todos los modelos de clasificación (binaria y multiclase) contaron con cuatro neuronas, correspondientes a las cuatro clases trabajadas, mientras que los modelos de regresión contaron con una sola neurona, la cual produjo la salida de toda la RNA, siendo esta la predicción de los valores ajustados de los datos hidrográficos.

Para definir las capas ocultas de la RNA, fue necesario experimentar con diferentes arquitecturas para encontrar la que tuviera el mejor desempeño en cada caso. Para ello se utilizó la herramienta Talos [20] (disponible para Python), que permitió realizar dichos experimentos. Se utilizó el rango de neuronas por capa de 10-60, en incrementos de 5, para todos los modelos de RNA.

2.3.1. Selección de las arquitecturas de RNA

Para los modelos de clasificación binarios, se seleccionaron las arquitecturas que contaron con el mayor grado de exactitud binaria de validación (*val_binary_accuracy*) como medida de desempeño, ya que esta es calculada tomando en cuenta los datos usados para validar los modelos, además de estar diseñada para funcionar con modelos binarios.

En el caso de los modelos de clasificación multiclase, se utilizó la métrica de exactitud de validación (*val_accuracy*) debido a que estos modelos trabajaron con más de dos clases.

Para los modelos de regresión, se eligieron las arquitecturas que presentaron el menor error absoluto medio. Después, estos modelos fueron entrenados utilizando los datos filtrados por los modelos de clasificación.

Finalmente, se utilizó el modelo de RNA de regresión con los datos reservados para pruebas para ajustar las mediciones de temperatura, salinidad y presión.

Capítulo 3

Resultados y discusión

3.1. Perfiles hidrográficos obtenidos

La consulta de perfiles hidrográficos medidos dentro del polígono (Figura 3.1) y del espacio temporal establecidos, reveló que 12, 874 perfiles de temperatura, salinidad y presión han sido medidos y revisados por el control de calidad en modo diferido dentro del área de estudio, los cuales fueron descargados para su posterior análisis.

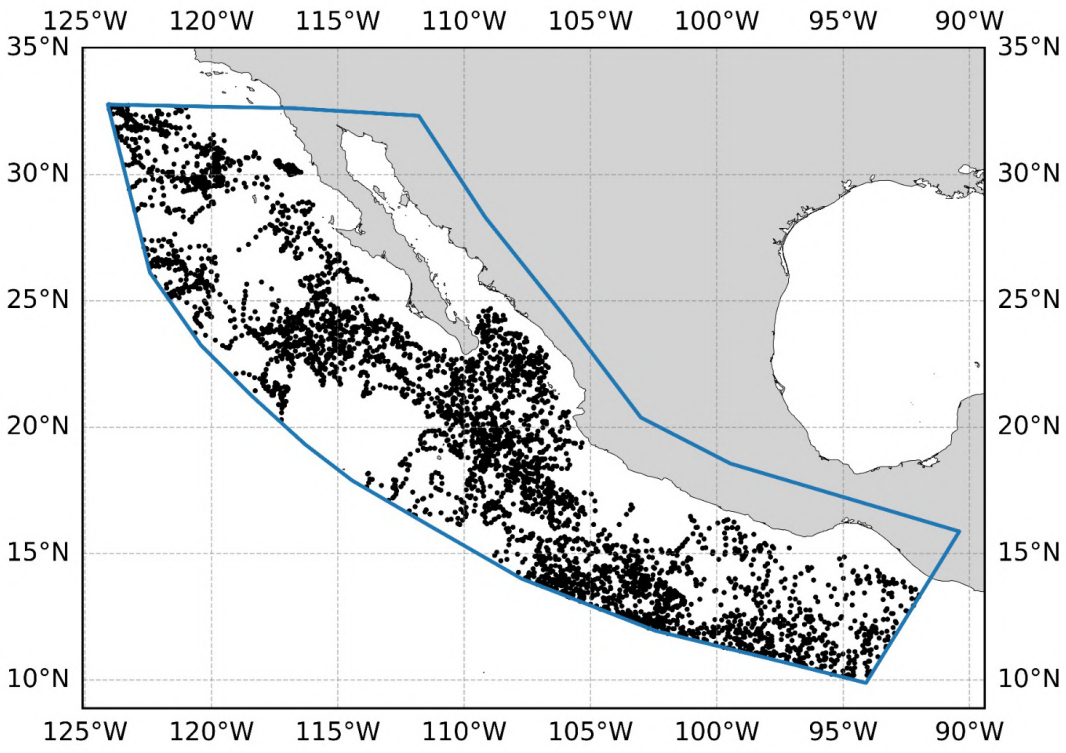


Figura 3.1: Polígono utilizado como área de estudio.

La proporción de la clasificación de estos perfiles se muestra en la Figura 3.2. Las clases se encuentran desbalanceadas, siendo la clase 1 la que contiene la mayor cantidad de datos (más del 90 %).

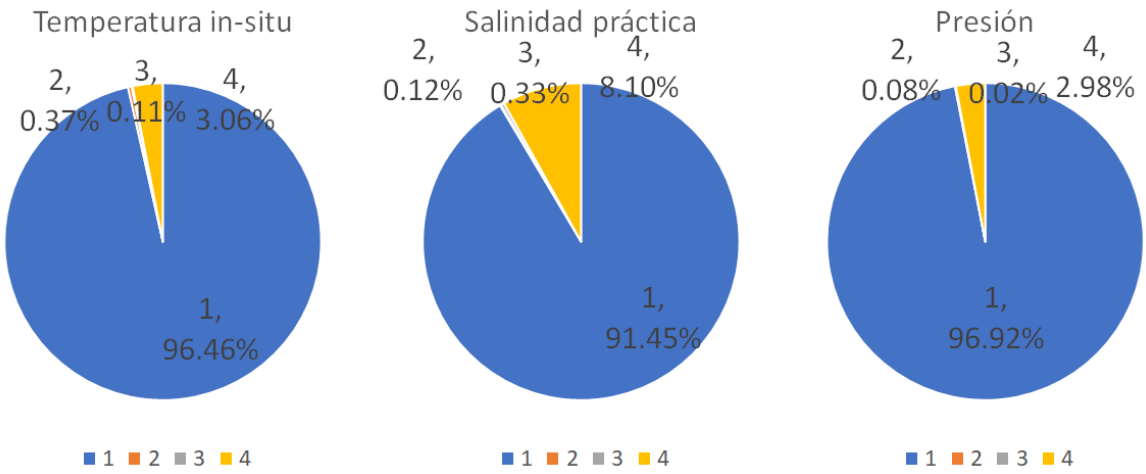


Figura 3.2: Proporción inicial de los datos de los perfiles hidrográficos.

3.2. Arquitecturas de RNA de clasificación

Las mejores arquitecturas de las RNA de clasificación encontradas por Talos se muestran a detalle en la Tabla 3.1.

	Binario		Multiclase	
	Capa oculta 1	Capa oculta 2	Capa oculta 1	Capa oculta 2
Temperatura	30	30	55	55
Salinidad	60	60	60	60
Presión	30	30	55	55

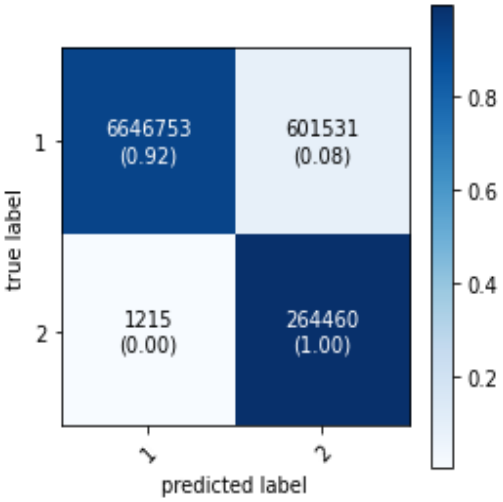
Tabla 3.1: Cantidad de neuronas por capa para cada RNA de clasificación.

Estas arquitecturas fueron las que presentaron mayor grado de exactitud. Como se observa en la tabla, para todos los modelos de RNA de clasificación, fueron necesarias sólo dos capas ocultas, y en el caso de la salinidad, una mayor cantidad de neuronas.

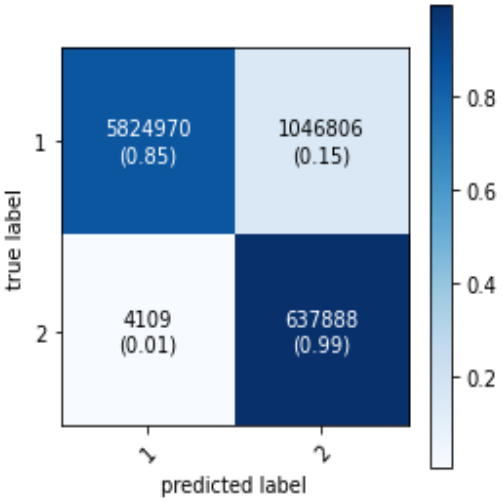
3.3. Pruebas de RNA de clasificación binaria

Los resultados de las pruebas de clasificación binaria se presentan en las Figuras 3.3a, 3.3b y 3.3c. Se puede apreciar que una gran cantidad de datos de la clase 1 de perfiles de temperatura,

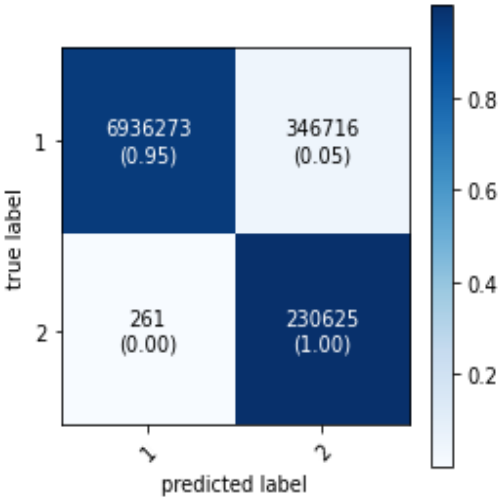
salinidad y presión fueron clasificados correctamente, con un porcentaje de verdaderos positivos de 91.7 %, 84.76 % y 95.23 % respectivamente.



(a) Temperatura.



(b) Salinidad.



(c) Presión.

Figura 3.3: Matrices de confusión de clasificación binaria.

Contando con los reportes de clasificación de las Tablas 3.2 para temperatura, 3.3 para salinidad y 3.4 para presión, destacando que los datos de la clase 1 tienen valores altos de **precisión**. Dicha métrica resultó ser la más relevante durante esta fase del proceso, ya que demostró que la clase 1 pudo ser identificada de forma satisfactoria, con un valor de precisión cercano a 1.0 para las tres medidas.

Clase	Precisión	Sensibilidad
1	**1.00**	0.92
2	0.31	1.00
Exactitud	0.92	

Tabla 3.2: Reporte de clasificación binario de temperatura in-situ.

Clase	Precisión	Sensibilidad
1	**1.00**	0.85
2	0.38	0.99
Exactitud	0.86	

Tabla 3.3: Reporte de clasificación binario de salinidad práctica.

Clase	Precisión	Sensibilidad
1	**1.00**	0.95
2	0.40	1.00
Exactitud	0.95	

Tabla 3.4: Reporte de clasificación binario de presión.

La nueva proporción de las clases (con la clase 1 filtrada) se detalla en la Figura 3.4, mostrando un conjunto de datos con un mejor balance de las clases que el que se tenía al principio.

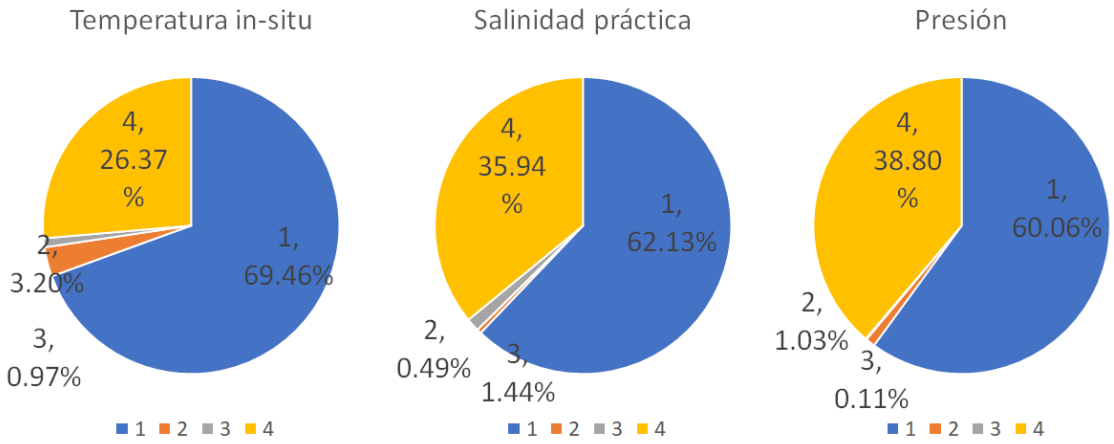
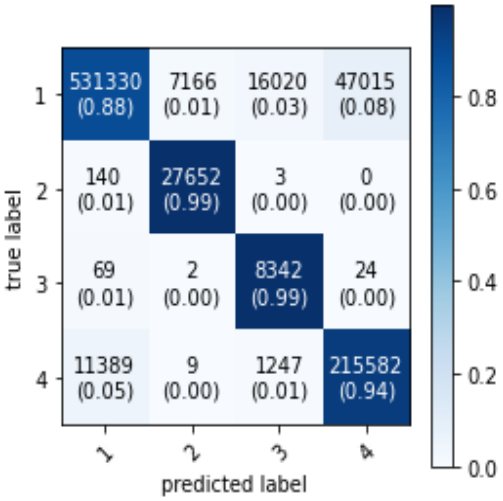


Figura 3.4: Clases de los perfiles hidrográficos balanceadas.

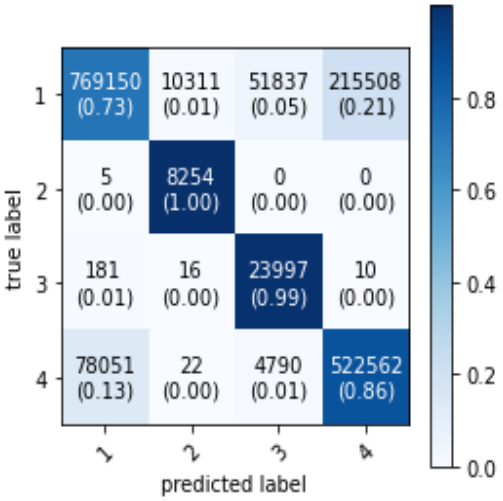
3.4. Pruebas de RNA de clasificación multiclase

Los resultados de las pruebas de clasificación multiclase se presentan en las Figuras 3.5a, 3.5b y 3.5c para temperatura, salinidad y presión respectivamente, mostrando en los tres casos que una gran parte de los datos clasificados se mantuvieron en la diagonal de las matrices, siendo los datos erróneamente clasificados menos que los correctamente clasificados.

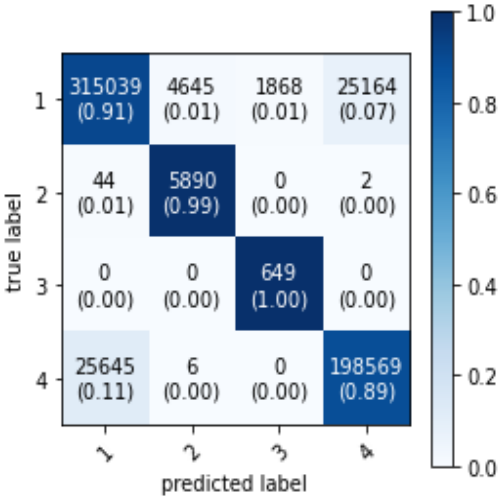
Contando con los reportes de clasificación de las Tablas 3.5, 3.6 y 3.7, destacando buenos niveles tanto de sensibilidad como de precisión, lo que demostró que los modelos tuvieron un buen desempeño para clasificar las cuatro clases de acuerdo con su calidad.



(a) Temperatura.



(b) Salinidad.



(c) Presión.

Figura 3.5: Matrices de confusión de clasificación multiclase.

Clase	Precisión	Sensibilidad
1	0.98	0.88
2	0.79	0.99
3	0.33	0.99
4	0.82	0.94
Exactitud	0.90	

Tabla 3.5: Reporte de clasificación multiclase de temperatura in-situ.

Clase	Precisión	Sensibilidad
1	0.91	0.73
2	0.44	1.00
3	0.30	0.99
4	0.71	0.86
Exactitud	0.79	

Tabla 3.6: Reporte de clasificación multiclase de salinidad práctica.

Clase	Precisión	Sensibilidad
1	0.93	0.91
2	0.56	0.99
3	0.26	0.99
4	0.89	0.89
Exactitud	0.90	

Tabla 3.7: Reporte de clasificación multiclase de presión.

Con los modelos de clasificación multiclase debidamente entrenados, se filtraron los datos que dichos modelos identificaron que pertenecían a la clase 4, de forma que solamente quedaran los datos de las otras clases (los que sí se pueden ajustar), para ser trabajados por los modelos de regresión.

3.5. Arquitecturas de RNA de regresión

Los resultados de los experimentos con Talos arrojaron las arquitecturas de RNA (Tabla 3.8) que presentaron el menor grado de error absoluto medio.

	Capa oculta 1	Capa oculta 2
Temperatura	30	30
Salinidad	30	30
Presión	30	30

Tabla 3.8: Arquitectura de RNA de regresión.

Al igual que para las RNA de clasificación, solamente fueron necesarias dos capas ocultas, y esta vez la cantidad de neuronas en cada capa oculta fue la misma para los tres modelos.

3.6. Pruebas de RNA de regresión

A continuación, en la Tabla 3.9 se muestran los resultados de las pruebas de estimación de los datos ajustados con respecto a los datos reales para el conjunto de datos de validación.

	R^2	RMSE	MAE	ME
Temperatura	0.9999	0.0005	0.0001	0.1211
Salinidad	0.9994	0.0136	0.0096	0.1559
Presión	0.9999	0.1298	0.0622	5.4696

Tabla 3.9: Resultados de RNA de regresión con datos de prueba.

Los tres modelos de RNA de regresión obtuvieron un valor de R^2 mayor a 0.99, demostrando que pudieron estimar con gran precisión los valores ajustados de los perfiles de temperatura, salinidad y presión. Dichos valores de R^2 se muestran en las gráficas de correlación lineal contra datos de validación, en las Figuras 3.6, 3.7 y 3.8, siendo la gráfica correspondiente a la salinidad la que mostró más inestabilidad, pero que de igual manera contó con un alto rendimiento, ayudando a resaltar los buenos resultados que pueden obtenerse al emplear modelos de RNA.

Además, las métricas RMSE y MAE dejaron ver que los modelos presentaron niveles de error extremadamente bajos, siendo el error promedio menor a la unidad, para cada uno de los tres datos hidrográficos, reforzando el potencial que tienen las RNA para obtener los valores correctos de los datos hidrográficos con un grado de error bastante aceptable.

Adicionalmente, la Figura 3.9 muestra algunos perfiles de temperatura cuyos datos fueron clasificados como clase 3 (datos de mala calidad potencialmente corregibles) siendo graficados junto a los valores ajustados de estos, así como los valores estimados por las RNA, mostrando la proximidad de estos últimos con los datos ajustados de buena calidad, además de que las estimaciones fueron calculadas en cuestión de segundos. Los datos medidos por los PHA son representados en color azul, mientras que los datos ajustados son de color naranja, y los predichos por las RNA de color verde.

Se puede observar que los valores estimados de temperatura coinciden casi en su totalidad con los valores ajustados por los especialistas del grupo Argo Floats, dejando ver que las RNA pudieron aproximar muy bien los datos del perfil de temperatura por su cuenta.

De igual manera, la Figura 3.10 muestra perfiles de salinidad con datos clasificados como clase 3, graficados junto a sus valores ajustados y valores predichos por la RNA. A diferencia de los perfiles de temperatura, en la salinidad se pueden distinguir ligeramente las líneas de diferentes colores que corresponden a cada tipo de dato. Esto puede deberse a la susceptibilidad de los sensores necesarios para calcular la salinidad de verse afectados por contaminación o sufrir cambios físicos que alteren su precisión. No obstante, la diferencia entre los valores ajustados y los predichos continúa siendo mínima, demostrando que la RNA puede estimar los valores correctos de salinidad con una alta precisión, un grado de error bajo y en un lapso de tiempo corto.

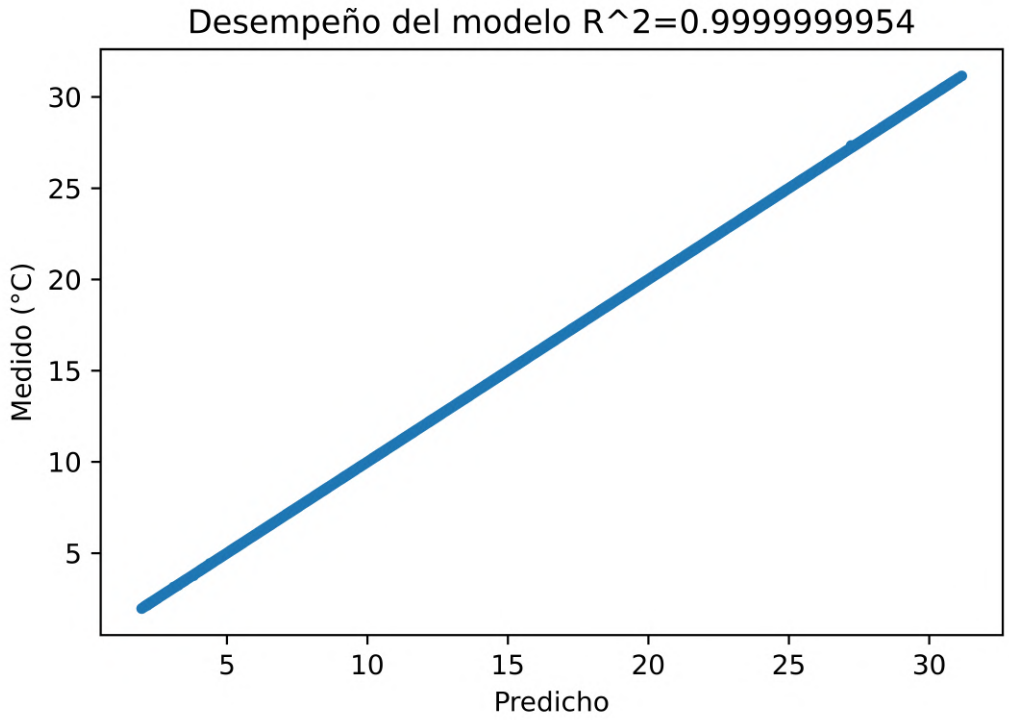


Figura 3.6: Correlación lineal entre los valores de temperatura predichos y los medidos.

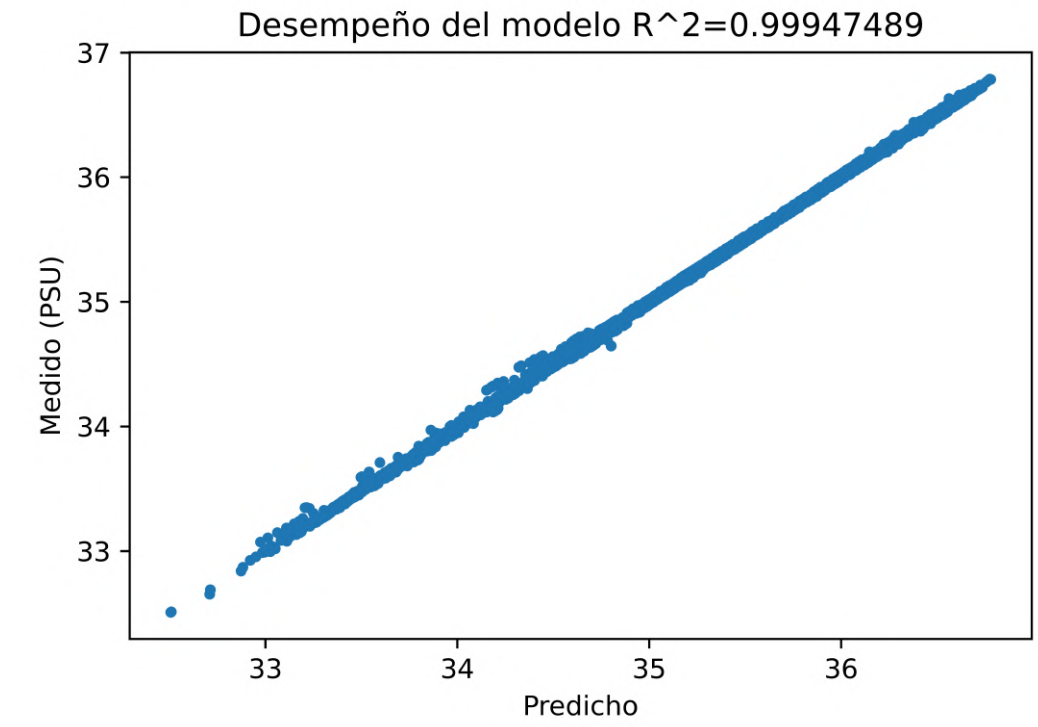


Figura 3.7: Correlación lineal entre los valores de salinidad predichos y los medidos.

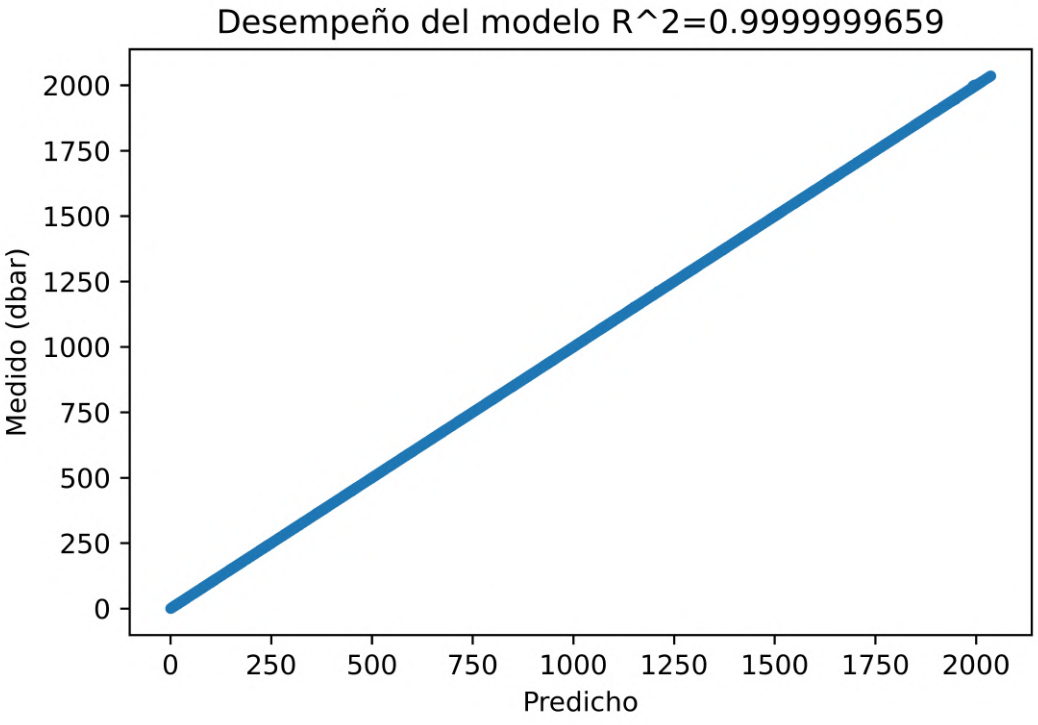
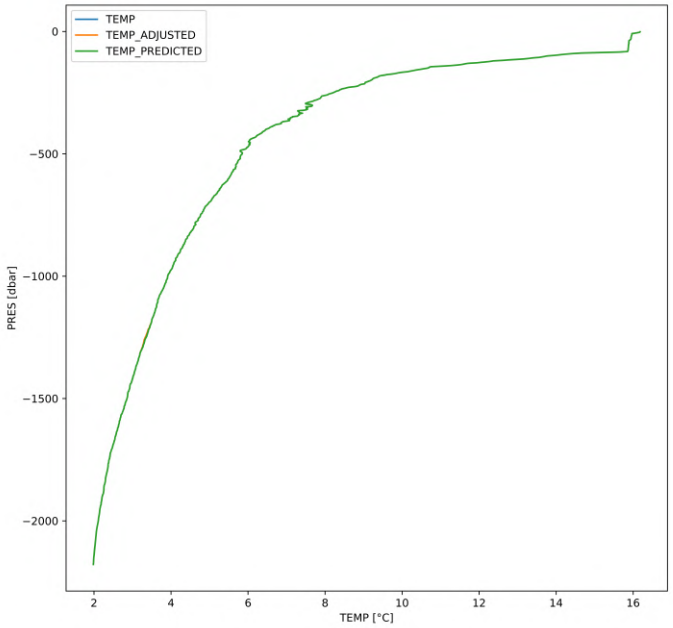
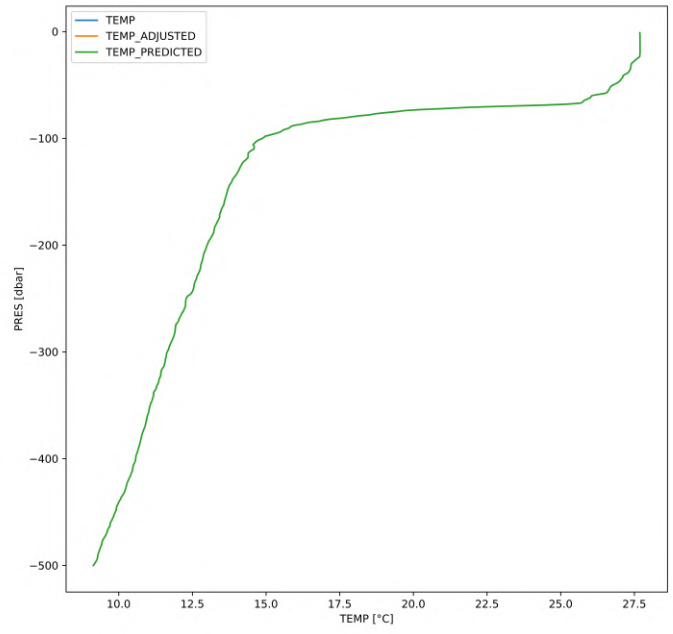


Figura 3.8: Correlación lineal entre los valores de presión predichos y los medidos.

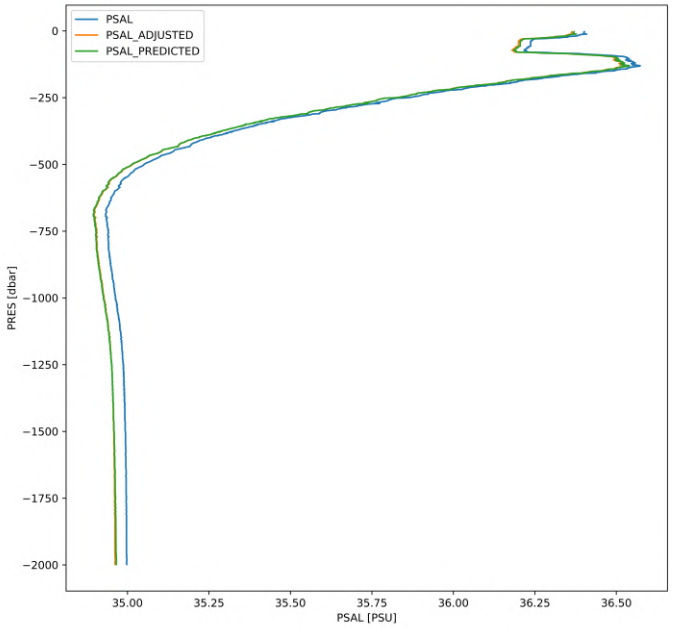


(a) Perfil 1.

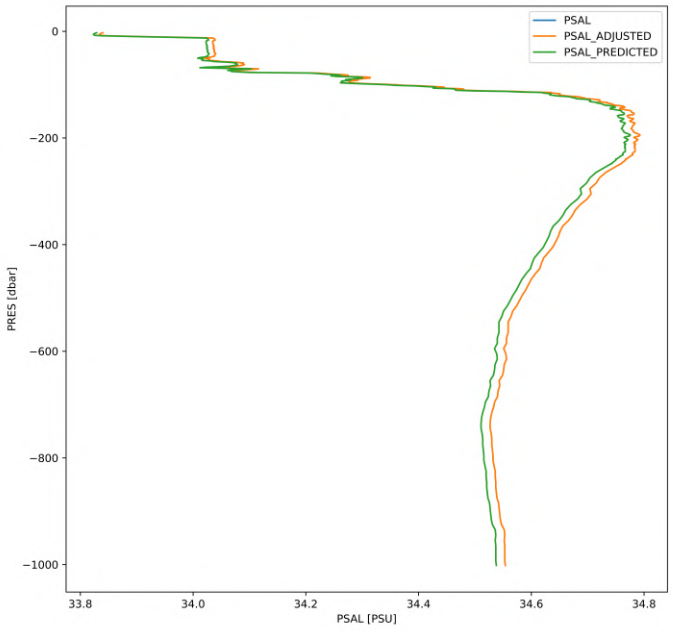


(b) Perfil 2.

Figura 3.9: Perfiles de temperatura clase 3 predichos por RNA de regresión.



(a) Perfil 1.



(b) Perfil 2.

Figura 3.10: Perfiles de salinidad clase 3 predichos por RNA de regresión.

3.6.1. Regresión lineal

Aunque los resultados de las pruebas de regresión con las RNA fueron satisfactorios, se probaron modelos de regresión lineal con datos de temperatura, salinidad y presión respectivamente, a modo de comparación con las RNA, dando un valor de R^2 similar al presentado por las RNA.

En la Tabla 3.10 se detallan los resultados de las pruebas de regresión lineal, con las métricas que se emplearon para evaluar los modelos de RNA de regresión.

	R^2	RMSE	MAE	ME
Temperatura	0.9999	$2.91 * 10^{-14}$	$2.87 * 10^{-14}$	$4.61 * 10^{-14}$
Salinidad	0.9981	0.0262	0.0156	0.2335
Presión	0.9999	0.1908	0.1426	2.4078

Tabla 3.10: Resultados de RNA de regresión lineal con datos de temperatura, salinidad y presión.

Con los resultados de los modelos de regresión lineal se comprobó que es posible obtener valores similares de R^2 con un modelo más sencillo. Sin embargo, en el caso de las métricas RMSE, MAE y Error máximo (ME), los valores varían dependiendo del tipo de dato que se esté trabajando. En el caso de la temperatura, estos presentan valores de error menores que los presentados por los modelos RNA, ya que no hubo ajuste de datos en los perfiles de temperatura que se descargaron, contribuyendo así al desempeño observado en el regresor lineal. No obstante, se pudo observar una diferencia en la salinidad y la presión, donde los niveles de error son más elevados que los presentados por las RNA.

3.7. Discusión

Los resultados obtenidos de las pruebas de clasificación demostraron que las RNA realizaron una mejor clasificación que otras técnicas revisadas en la literatura. Un ejemplo es el de [3], donde el modelo de campo aleatorio condicional que utilizaron les permitió identificar patrones de error en los perfiles hidrográficos con una exactitud de 97 %.

Para los modelos de RNA generados por la metodología propuesta, es posible identificar no solamente datos hidrográficos con errores (clase 3 o 4 por ejemplo), sino también datos

con niveles buenos de calidad, lo que permite clasificar mejor la información de los perfiles hidrográficos.

En [4], el enfoque fue similar al ejemplo anterior, con la diferencia de que únicamente trabajaron datos de temperatura, mientras que la metodología de este estudio permite construir modelos de RNA que pueden trabajar con datos de distintas variables.

En trabajos como el de [5], fue necesario transformar los datos a otro formato (integrales iteradas) para poder ser procesados, mientras que las RNA permitieron trabajar con los datos hidrográficos en su formato original.

En el caso de la regresión y la estimación de los valores ajustados de los datos hidrográficos, la comparación realizada contra los modelos de regresión lineal demostró que las RNA cuentan con un mayor rendimiento y menor grado de error para estimar los valores ajustados de los datos hidrográficos de salinidad y presión de los perfiles medidos dentro del polígono y espacio temporal establecidos.

Capítulo 4

Conclusiones

El filtro de clasificación binario propuesto en la metodología de este estudio para remover datos de la clase 1 (datos de buena calidad), permitió obtener un conjunto de datos mejor balanceado, lo cual a su vez derivó en un mejor desempeño para clasificar los datos de todas las clases trabajadas.

Se observó que los modelos de RNA obtenidos mediante la metodología propuesta cuentan con un adecuado rendimiento para realizar la clasificación de datos hidrográficos de acuerdo a su nivel de calidad, contando con valores altos de exactitud.

Por otra parte, se observó una gran precisión en la estimación de los valores ajustados por las RNA de regresión, con un valor de R^2 superior a 0.99, así como valores de error de RMSE y MAE bajos, por lo que el modelo es capaz de estimar los valores ajustados de los datos hidrográficos con una precisión alta y un error bajo.

Adicionalmente, es posible adaptar esta metodología de acuerdo a las necesidades de quien la utilice, ya sea para analizar variables diferentes, o para clasificar y ajustar datos de perfiles hidrográficos medidos dentro de un polígono y/o espacio temporal diferentes a los establecidos en este estudio.

Por lo anterior se concluye que la metodología propuesta sirve como alternativa viable cuando no se cuenta con datos procesados por el control de calidad en modo diferido.

Bibliografía

[1] OceanOPS. Argo: A window into the ocean, 2018.

[2] Argo Data Management Team. Argo user's manual version 3.3, 11 2019.

[3] Satoshi Ono, Haruki Matsuyama, Ken ichi Fukui, and Shigeki Hosoda. *A Preliminary Study on Quality Control of Oceanic Observation Data by Machine Learning Methods*, volume 1, pages 679–693. 2015.

[4] Qi Zhang, Chenyan Qian, and Changming Dong. A machine learning approach to quality-control argo temperature data. *Atmospheric and Oceanic Science Letters*, 16, 7 2023.

[5] Nozomi Sugiura and Shigeki Hosoda. Machine learning technique using the signature method for automated quality control of argo profiles. *Earth and Space Science*, 7, 9 2020.

[6] Emmanuel Romero, Leonardo Tenorio-Fernandez, Iliana Castro, and Marco Castro. Filtering method based on cluster analysis to avoid salinity drifts and recover argo data in less time. *Ocean Science*, 17:1273–1284, 9 2021.

[7] Marco Pérez. Modelo predictivo de series temporales de datos hidrográficos mediante aprendizaje automático. 2022.

[8] Dean Roemmich et al. On the future of argo: A global, full-depth, multi-disciplinary array, 1 2019.

[9] Annie P.S. Wong et al. Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats, 9 2020.

[10] OceanOPS. Argo national contributions. `https://www.ocean-ops.org/board/?t=argo#`, 2024.

[11] Juan Ignacio Barrios Arce. La matriz de confusión y sus métricas – inteligencia artificial. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/, 2019. [Accessed 12-06-2024].

[12] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. 1 edition, 1 2018.

[13] Nicolás Arrioja Landa Cosio. Métricas en regresión. https://medium.com/@nicolasarrioja/m%C3%A9tricas-en-regresi%C3%B3n-5e5d4259430b, 2021. [Accessed 14-06-2024].

[14] Damián Matich. Redes neuronales: Conceptos básicos y aplicaciones, 2001.

[15] Ramón Piloto Rodríguez. Redes neuronales artificiales. conceptos básicos y algunas aplicaciones en energía, 2017.

[16] Pablo Huet. Qué son las redes neuronales artificiales y sus aplicaciones, 2023.

[17] Juan Monroy, Adriana Ramírez, Roberto Alejo, and Erika López. Aspectos relevantes para mejorar el desempeño del algoritmo backpropagation, 12 2017.

[18] David Díaz Solis. Cómo funciona el algoritmo de backpropagation. https://medium.com/@ddiazsolis/blog-c%C3%B3mo-funciona-el-algoritmo-de-backpropagation-22575308f14b, 2020. [Accessed 14-06-2024].

[19] Emmanuel Romero, Leonardo Tenorio-Fernandez, Iliana Castro, and Marco Castro. romeroqe/cluster_qc: Filtering Methods based on cluster analysis for Argo Data, March 2021.

[20] Autonomio. Talos, 2024.