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

DISEÑO DE RED NEURONAL ARTIFICIAL PARA PREDECIR LA DISTRIBUCIÓN GEOGRÁFICA DE LA SARDINA DEL PACÍFICO (*SARDINOPS SAGAX*)

TESIS

QUE PARA OBTENER EL GRADO DE
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:
JONATHAN ALEJANDRO REYES GRACIA

DIRECTORES DE TESIS:
DR. MARCO ANTONIO CASTRO LIERA
DRA. LAURA SÁNCHEZ VELASCO

LA PAZ, BAJA CALIFORNIA SUR, MÉXICO, DICIEMBRE 2021.

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Instituto Tecnológico de La Paz

La Paz, B.C.S., 1/agosto/2016

DEPI/345/2016

ASUNTO: Autorización de impresión

[redacted]
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: "MODELO DE DATOS PARA EL [redacted], 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
"CIENCIA ES VERDAD, TÉCNICA ES LIBERTAD"

M.C. MANUEL E. CASILLAS BROOK,
SUBDIRECTOR ACADÉMICO.

INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO
E INVESTIGACIÓN

c.c.p. Archivo.
MACB/LACF/icl

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Tecnológico Nacional de México
Instituto Tecnológico de La Paz

DICTAMEN DEL COMITÉ TUTORIAL

SUBDIRECCIÓN ACADÉMICA
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN.

La Paz, B.C.S., 01 /agosto/ 2016

C. MC. MANUEL E. CASILLAS BROOK,
SUBDIRECTOR ACADÉMICO,
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 M14310019	Nombre [redacted]
Maestría en:	SISTEMAS COMPUTACIONALES
Título de la tesis: [redacted]	
DICTAMEN: Se autoriza el trabajo de investigación, en virtud de que realizó las correcciones correspondientes conforme a las observaciones planteadas por este Comité Tutorial.	

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

[redacted]

[redacted]

[redacted]

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

Dedicatoria

Este trabajo está dedicado a principalmente a mis padres, a mi prometida y a todas aquellas personas que han creído en mí y me han apoyado de alguna manera, tanto a los que todavía están conmigo como a los que ya han partido.

Agradecimientos

Agradezco a mis padres Jorge y América que siempre han creído en mí desde que era un niño y hasta el último momento.

A mi amada prometida Mafer, que siempre está para escucharme por muchas horas y para apoyarme incondicionalmente.

A mi hermano Cristian, que incluso de lejos siempre confía en que lograré mis propósitos.

A mis maestros, que me han brindado sus conocimientos y guía para este trabajo.

A mi comité de tesis, Dr. Marco Antonio Castro Liera, Dra. Laura Sánchez Velasco y Dr. Jesus Alberto Sandoval Galarza que en cada una de mis evaluaciones me corrigieron, me cuestionaron y asesoraron para lograr mi objetivo.

A mi amigo y compañero ajedrecista, Joel Artemio Morales Viscaya que me apoyó en mis debilidades con sus conocimientos de matemáticas para crear soluciones indispensables para este trabajo.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada durante los dos años de trabajo.

Resumen

En ecología, los modelos de distribución de especies (SDM) son ampliamente usados dada la importancia que tiene poder predecir la distribución geográfica de las especies a partir de un conjunto de ocurrencias, por intereses económicos y principalmente con propósitos de preservación de las especies ante el cambio climático. Usualmente se recurre a modelados de nicho ambiental para crear una representación del entorno de la especie a partir de un conjunto limitado de variables climáticas, lo cual supone una limitante. Este trabajo presenta una alternativa a los modelos de distribución de especies populares, basada en redes neuronales artificiales, con mejores resultados de predicción y cuya arquitectura permite añadir patrones de entrada que en los modelos populares no son considerados.

Abstract

In ecology, species distribution models (SDM) are widely used given the importance of being able to predict the geographical distribution of species from a set of occurrences, for economic interests and mainly for the purposes of species preservation against climate change. Environmental niche modeling is usually used to create a representation of the environment of the species from a limited set of climatic variables, which is a limitation. This work presents an alternative to popular species distribution models, based on artificial neural networks, with better prediction results and with an architecture that allows adding input patterns wich are not considered in popular models.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Los modelos de distribución de especies son ampliamente usados en ecología con propósitos de económicos y de conservación de las especies [1]. En esencia estos modelos intentan predecir la distribución geográfica de la especie a partir de registros de presencia de la misma. Este trabajo presenta un modelo de red neuronal artificial capaz de predecir la distribución geográfica de la Sardina del Pacífico a partir de los registros de presencia y con mejores resultados respecto a una de las técnicas más empleadas en este tipo de problemas. La importancia de estudiar este tema y diseñar modelos con mayor exactitud de predicción radica en la relevancia que ha cobrado el cambio climático en años recientes y la necesidad de crear estrategias para preservar los ecosistemas y por lo tanto también mitigar el impacto que tiene en las economías mundiales. En el capítulo tres de este trabajo se describe la solución propuesta, el procesamiento de los datos y; el diseño y configuración del modelo seleccionado. Por otro lado, en el capítulo cuatro se muestran los resultados, su contraste con el estado del arte y se evalúan dichos resultados, de igual forma se somete a discusión la importancia de usar múltiples métricas para evaluar este tipo de modelos con el fin de interpretar correctamente la información obtenida.

1.1. Antecedentes

El calentamiento global antropogénico ha influenciado significativamente procesos físicos y biológicos a escalas regionales y globales. Los cambios observados y anticipados en el clima global presentan oportunidades y desafíos para las sociedades y economías[2].

El impacto del ser humano en el medio ambiente se hace más notorio año con año, impac-

tando especies en peligro de extinción hasta fracciones de ecosistemas erradicados. Esto afecta el balance ecológico global y con ello la economía de las sociedades que dependen de actividades primarias como la pesca, por lo que es una prioridad desarrollar herramientas que permitan estudiar el comportamiento de las especies marinas y los parámetros oceanográficos que constituyen el respectivo hábitat idóneo con el fin de predecir la distribución geográfica de las especies.

Uno de los trabajos más recientes en esta área corresponde a Petetán-Ramirez[3] con el trabajo *Potential changes in the distribution of suitable habitat for Pacific sardine (Sardinops Sagax) under climate changes scenarios*, quien describe el estado de la pesquería de la Sardina del Pacífico en México. En ese trabajo se utiliza una técnica de *Machine Learning* llamada Modelado de Máxima Entropía, también conocido como MaxEnt para estudiar el comportamiento de la especie bajo cambios en el entorno.

1.2. Descripción del problema

Las pesquerías mundiales proporcionan a más de 2,600 millones de personas al menos el 20 % de su ingesta anual promedio de proteínas per cápita[2]. La Sardina por su volumen se encuentra posicionada en el lugar número uno de la producción pesquera en México[4], llegando a una producción total en el año 2018 de 587,433 toneladas de peso vivo y siendo los principales productores nacionales los estados de Sonora, Baja California Sur y Sinaloa. En la figura 1.1 se puede observar con más detalle la producción de Sardina en México por estado.

ENTIDAD	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
TOTAL	872,640	629,811	684,132	721,735	727,816	562,872	443,787	441,608	721,571	587,433
SONORA	623,184	404,684	382,016	409,767	441,371	257,971	253,228	228,309	385,248	320,114
BAJA CALIFORNIA SUR	64,001	69,062	80,414	90,829	91,577	94,703	69,830	81,594	107,818	110,564
SINALOA	120,522	90,069	139,437	152,522	137,309	107,438	78,585	56,511	81,274	82,378
BAJA CALIFORNIA	64,856	65,912	81,590	68,520	57,515	102,448	41,997	74,244	146,246	73,973
YUCATÁN	0	4	-	0	-	223		215	233	296
VERACRUZ	67	46	37	48	17	19	79	236	141	75
COLIMA	8	21	9	23	16	61	67	46	39	30
CAMPECHE	1	1	-	14	-	2	1	1	1	2
MICHOACÁN	-	0	-	-	-	-	-	-	-	1
OTRAS*	1	12	630	12	11	7	1	452	573	-

*CHIAPAS,GUERRERO,JALISCO,NAYARIT,OAXACA,QUINTANA ROO,TABASCO.

Figura 1.1: Serie histórica de producción de Sardina en México 2009-2018

Predecir con precisión el impacto que tendrá el cambio climático en los ecosistemas durante los próximos años con intereses económicos, de preservación de las especies y por lo tanto de la conservación del balance de las cadenas tróficas es una necesidad evidente; tal como lo expresa Allison2009[2] "... es difícil estimar o predecir los efectos más amplios o agregados del cambio climático a escalas nacional y regional. Además, se ha prestado poca atención a las consecuencias de los cambios en los ecosistemas pesqueros en las personas, particularmente para los millones de pescadores a pequeña escala (pescadores, procesadores de pescado, comerciantes y trabajadores auxiliares) en el mundo en desarrollo los cuales se encuentran entre los más vulnerables al cambio climático". El crecimiento del poder de cómputo que se tiene a disposición actualmente y el reciente auge del Machine Learning permite la implementación de un modelo de Red Neuronal Artificial que estime el peso que tiene cada variable en el entorno de la Sardina y así posteriormente utilizar el modelo entrenado para determinar con alta precisión la presencia o ausencia de la especie bajo determinados parámetros. Adicional al cambio climático antropogénico también hay otros fenómenos naturales conocidos como oscilaciones atmosféricas que provocan cambios en la presión atmosférica y en las temperaturas marinas en diferentes regiones del mundo. *El Niño Southern Oscillation* (ENSO por sus siglas en inglés) es uno de estos fenómenos que afecta al Pacífico Ecuatorial elevando la temperatura y hundiendo la termoclina en el océano y por consecuencia el hábitat de la Sardina del Pacífico y de otras muchas especies.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar un modelo de Red Neuronal Artificial de Clasificación entrenado con parámetros oceanográficos relevantes para la habitabilidad de la Sardina del Pacífico *(Sardinops Sagax)* con la capacidad de predecir su distribución geográfica ante los cambios climáticos.

1.3.2. Objetivos específicos

- Tener un conjunto de datos de entrenamiento válido para entrenar y validar un modelo de red con las características adecuadas para la especie bajo estudio.

- Contar con un modelo de red neuronal artificial de clasificación óptimo.
- Obtener un reporte de desempeño del modelo de clasificación útil para comparar resultados con el estado del arte.

1.4. Justificación

Implementar con éxito un modelo de Red Neuronal Artificial que proporcione una mayor confiabilidad de predicción que las aproximaciones hechas al día de hoy permitirá demostrar la eficiencia y la versatilidad de esta técnica de Machine Learning para usarse en otras especies de interés a través de mínimos cambios en el modelo y en el proceso de entrenamiento. Se conseguirá trazar un procedimiento claro y poco variable para el uso de esta técnica en especies en peligro de extinción, especies de importancia comercial y para los fines que los expertos convengan.

Por otro lado, tener un modelo que logre predecir la distribución geográfica de las especies en función de los cambios que sufre su entorno permitirá desarrollar estrategias a favor de la preservación de las especies al anticipar el impacto que tendrá el cambio climático ya previsto por fenómenos como El Niño.

1.5. Alcances y limitaciones

- La investigación está dedicada a la Sardina del Pacífico (*Sardinops Sagax*).
- El área geográfica de estudio está delimitado al norte con latitud $31.04°$, al sur con latitud $20.0°$, al oeste con longitud $-120.0°$ y al este con longitud $-105.04°$.
- El periodo de estudio está comprendido del 1 de enero del 2000 hasta el 31 de diciembre del 2012.

1.6. Hipótesis

Es posible generar un modelo de clasificación, basado en Redes Neuronales Artificiales, para predecir la distribución geográfica de la Sardina del Pacífico (*Sardinops Sagax*) obteniendo un mejor desempeño de predicción que el encontrado en el estado del arte.

Capítulo 2

Marco teórico

2.1. Aprendizaje automático

También conocido como *Machine Learning* es la ciencia (y arte) de programar computadoras para que puedan aprender de datos. Una definición más orientada a ingeniería sería: Un programa de computadora se dice que aprende de la experiencia E con respecto a alguna tarea T y con alguna medida de desempeño P, si su desempeño en T, mejora con la experiencia E[5].

2.2. Red neuronal artificial

El campo de las Redes neuronales artificiales (RNA) se concentra en la investigación de modelos computacionales inspirados por la observación de la estructura y funcionamiento de las redes biológicas de células neuronales en el cerebro. Generalmente se diseñan como modelos para abordar problemas matemáticos, computacionales y problemas de ingeniería. Una RNA está compuesta de un grupo de neuronas artificiales interconectadas entre si para desempeñar ciertos cálculos a partir de patrones de entrada y generar patrones de salida. Son sistemas adaptativos capaces de modificar su estructura interna, usualmente los pesos entre los nodos (neuronas) de la red, de tal forma que pueden usarse para una gran variedad de problemas de aproximación de funciones tales como clasificación, regresión, extracción de características y memoria de contenido direccionable. Un ejemplo de arquitectura de red neuronal se muestra en la figura 2.1.

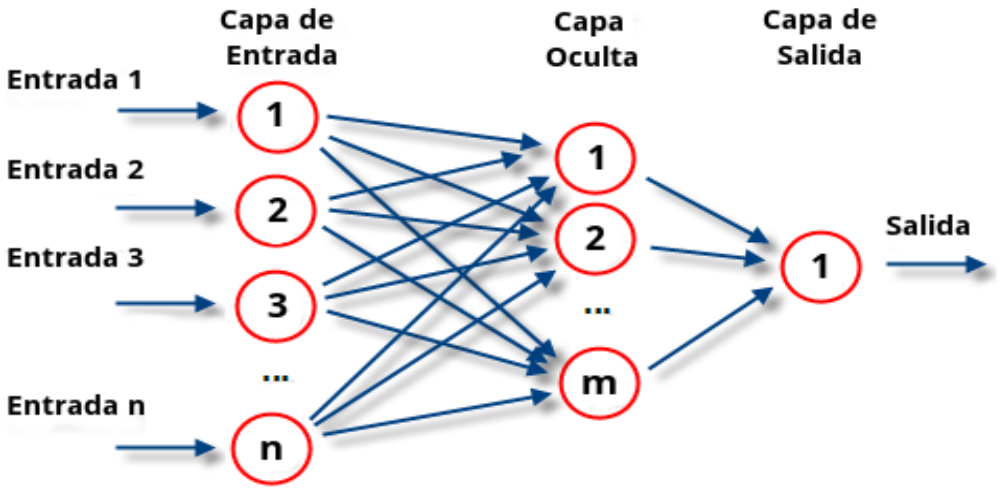

Figura 2.1: Esquema de Red Neuronal Artificial

2.3. Perceptrón

El perceptrón es una de las arquitecturas de RNA más simples, inventado en 1975 por Frank Rosenblatt. Se basa en una neurona artificial ligeramente diferente llamada unidad de umbral lineal (LTU): las entradas y salidas ahora son números (en lugar de valores binarios de encendido / apagado) y cada conexión de entrada está asociada con un peso. La LTU calcula una suma ponderada de sus entradas y luego aplica una función escalonada a esa suma y genera el resultado tal como se muestra en la figura 2.2.

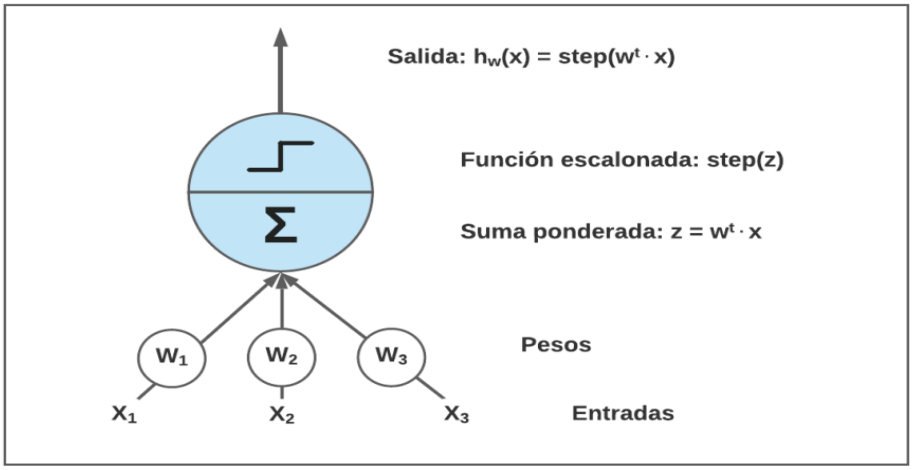

Figura 2.2: Unidad de umbral lineal

2.4. Perceptrón multicapa

Un perceptrón multicapa (MLP) se compone de una capa de entrada (de paso), una o más capas de LTUs llamadas capas ocultas, y una capa final de LTUs llamada capa de salida, véase figura 2.3.

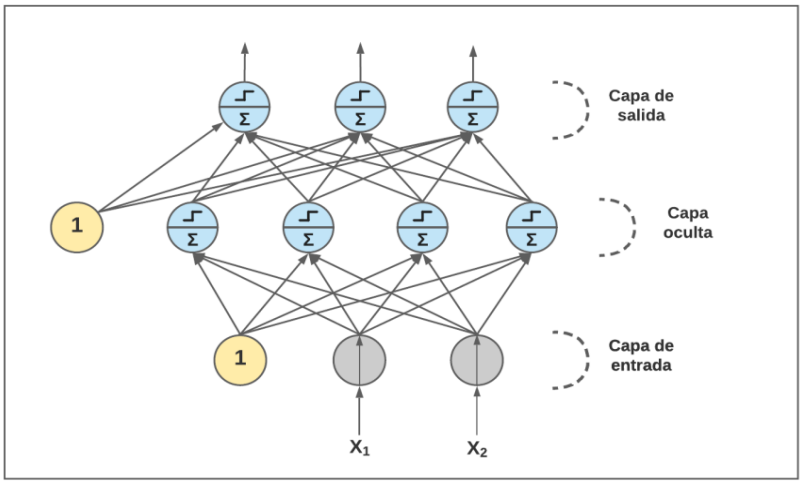

Figura 2.3: Modelo de perceptrón multicapa

2.5. *Backpropagation*

El algoritmo de *Backpropagation* es un método para entrenamiento de pesos en redes neuronales tipo *feed-forward* de múltiples capas. Como tal, requiere definir una estructura de red de una o más capas donde una capa está completamente conectada a la siguiente. Una estructura de red estándar es una capa de entrada, una capa oculta y una capa de salida. El método se ocupa principalmente de adaptar los pesos al error calculado en presencia de patrones de entrada, y el método se aplica hacia atrás desde la capa de salida de la red hasta la capa de entrada.

Para preparar la red usando el método de entrenamiento de *Backpropagation* se inicializa un peso para cada entrada más un peso adicional para una constante bias fija de entrada que casi siempre se establece en 1.0. La activación de una sola neurona dado un patrón de entrada

se calcula como se muestra en (2.1)

$$activation = \left(\sum_{k=1}^{n} w_k \cdot x_{ki} \right) + w_{bias} \cdot 1.0 \tag{2.1}$$

donde n es el número de pesos y entradas, x_{ki} es el *k-ésimo* atributo en el *i-ésimo* patrón de entrada, y w_{bias} es el peso del bias. Se utiliza una función de transferencia logística (sigmoidea) para calcular la salida de una neurona $\in [0, 1]$ y provee no linealidad entre las señales de salida y de entrada: $\frac{1}{1+exp(a)}$, donde a representa la activación de la neurona. La actualización de pesos utiliza la regla delta, específicamente una regla delta modificada donde el error es propagado hacia atrás a través de la red, empezando en la capa de salida y ponderado a través de las capas anteriores.

2.6. Modelado de Distribución de Especies (SDM)

Los modelos de distribución de especies son una herramienta ampliamente usada en estudios ecológicos. Estos modelos analizan los patrones espaciales de los organismos a través de procedimientos estadísticos y cartográficos basados en registros y datos reales. Buscan inferir la distribución espacial de una especie a partir de áreas potencialmente adecuadas según sus características ambientales. La idoneidad de un hábitat es una relación matemática entre la distribución real conocida y un conjunto de variables independientes que se usan como indicadores, dichas variables pueden ser geológicas, topográficas o climáticas, y se espera que con alguna combinación de ellas, se puedan definir los factores ambientales que delimiten las condiciones favorables para la presencia de la especie [6].

2.7. Métricas de clasificación

Si bien la preparación de los datos y el entrenamiento de un modelo de aprendizaje automático es un paso clave en el proceso, es igualmente importante medir el rendimiento de este modelo entrenado. Lo bien que el modelo generaliza sobre los datos no vistos es lo que define los modelos de aprendizaje automático adaptables frente a los no adaptables, es decir, aquellos que son capaces de modificar su arquitectura interna (generalmente pesos) según el problema lo requiera.

Al utilizar diferentes métricas para la evaluación del rendimiento, deberíamos estar en posición de mejorar el desempeño de predicción general de nuestro modelo antes de que lo pongamos en marcha para la predicción sobre datos no vistos antes. Si no se realiza una evaluación adecuada del modelo aprendizaje automático utilizando diferentes métricas, y se usa sólo la precisión, puede darse un problema cuando dicho modelo se despliega sobre datos no vistos y dar lugar a predicciones incorrectas.

2.7.1. Matriz de confusión

Una matriz de confusión, también conocida como matriz de error, es una tabla resumida que se utiliza para evaluar el rendimiento de un modelo de clasificación. El número de predicciones correctas e incorrectas se resumen con los valores de conteo y se desglosan por clase. A continuación se presenta en la figura 2.4 una matriz de confusión de 2x2 para ejemplificar.

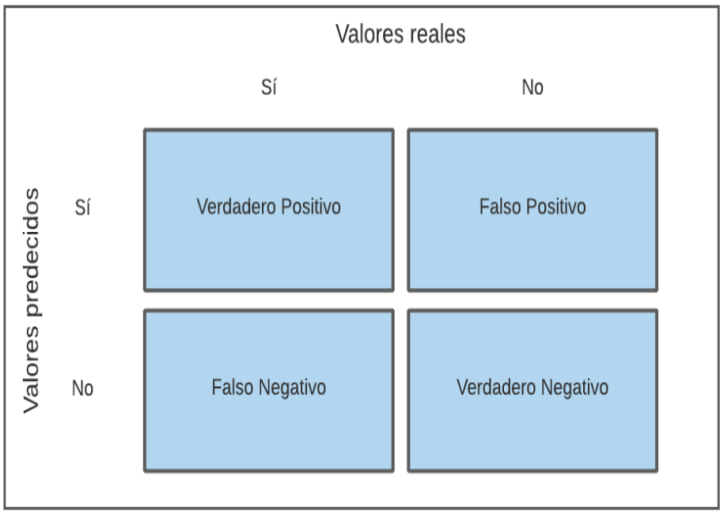

Figura 2.4: Estructura de una matriz de confusión de 2x2

Para entender mejor la matriz es necesario definir algunos conceptos:

- Positivo (P): La observación es positiva.
- Negativo (N): La observación es negativa.
- Verdadero Positivo (TP): Resultado en el que el modelo predice correctamente la clase positiva.

- Verdadero Negativo (TN): Resultado en el que el modelo predice correctamente la clase negativa.
- Falso Positivo (FP): También llamado error de tipo 1, resultado donde el modelo predice incorrectamente la clase positiva cuando en realidad es negativa.
- Falso Negativo (FN): También llamado error de tipo 2, resultado donde el modelo predice incorrectamente la clase negativa cuando en realidad es positiva.

2.7.2. *Accuracy* (Exactitud)

Es el porcentaje total de elementos clasificados correctamente. Se consigue sumando los verdaderos positivos más los verdaderos negativos y dividiendo el resultado entre el total de datos de la matriz de confusión, es decir, la suma de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos y se calcula mediante la expresión (2.2). Cuando las clases son aproximadamente iguales en tamaño, se puede usar esta métrica, de otro modo no se recomienda porque puede ser difícil interpretar correctamente los datos.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{2.2}$$

2.7.3. *Precision* (Precisión)

La precisión también se conoce como valor predictivo positivo y es la proporción de instancias relevantes entre las instancias recuperadas. En otras palabras, responde a la pregunta ¿Qué proporción de identificaciones positivas fue realmente correcta?. Se calcula como se muestra en (2.3).

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

2.7.4. *Recall* (Sensibilidad)

La sensibilidad cuantifica la capacidad del clasificador de encontrar todas las muestras positivas. Se define como se muestra en la figura, siendo el mejor valor posible 1 (que corresponde

al 100 %) y el peor 0, este índice de desempeño se calcula mediante (2.4).

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

2.7.5. *F1-Score*

El *F1-Score* se puede interpretar como la media ponderada de la precisión y la sensibilidad, el *F1-Score* alcanza su mejor valor en 1 y la peor puntuación en 0. La contribución relativa de precisión y recuperación a la puntuación $F1$ es la misma, como se describe en (2.5).

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.5}$$

2.7.6. Curva AUC/ROC (Área bajo la curva/*Receiver operanting characteristics*)

La curva AUC - ROC es una medida de rendimiento para los problemas de clasificación en varios valores de umbral. ROC es una curva de probabilidad y AUC representa el grado o medida de separabilidad. Indica cuánto es capaz el modelo de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo para predecir 0 clases como 0 y 1 clases como 1.

La curva ROC se grafica con TPR (*True Positive Rate*) contra FPR (*False Positive Rate*), tal como se muestra en el ejemplo de la figura 2.5.

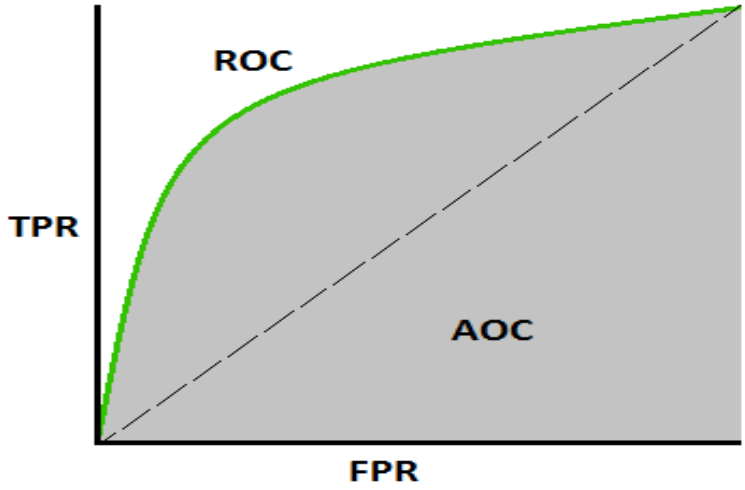

Figura 2.5: Ejemplo de curva ROC y AUC en el plano.

2.8. Python

Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. Sus estructuras de datos integradas de alto nivel, combinadas con tipado dinámico y enlace dinámico, lo hacen muy atractivo para el desarrollo rápido de aplicaciones, así como para su uso como lenguaje de *scripts* o pegamento para conectar componentes existentes.

La implementación de algoritmos de inteligencia artificial y aprendizaje automático puede ser complicada y requiere mucho tiempo. Es fundamental contar con un entorno bien estructurado y probado para que los desarrolladores puedan encontrar las mejores soluciones de codificación. Entre las múltiples librerías que Python ofrece, aquí se listan las más importantes que hacen de este lenguaje la mejor opción para software de inteligencia artificial:

- Keras, Scikit-learn y Tensorflow para aprendizaje automático.
- Numpy para análisis de datos y computo científico de alto rendimiento.
- SciPy para cómputo avanzado..
- Pandas para análisis de datos de propósito general.
- Seaborn y MatPlotLib para visualización de datos.

2.9. Scikit-learn

Scikit-learn es un módulo de Python que integra una amplia gama de algoritmos de aprendizaje automático de última generación para problemas supervisados y no supervisados de mediana escala. La gran variedad de algoritmos y utilidades de Scikit-learn la convierten en la herramienta básica para empezar a programar y estructurar los sistemas de análisis datos y modelado estadístico. Los algoritmos de Scikit-Learn se combinan y depuran con otras estructuras de datos y aplicaciones externas como Pandas o PyBrain.

2.10. Zona epipelágica

Comprende la región del océano donde la luz solar penetra con facilidad, ya que contempla desde la superficie hasta los 200 metros de profundidad, permitiendo la fotosíntesis de las plantas

y se caracteriza por su abundante vida marina.

2.11. Especie

El concepto biológico de especie más popular es el propuesto por Ernst Mayr en 1940, que concibe la especie biológica del siguiente modo: "Las especies son grupos de poblaciones naturales con cruzamiento entre sí que están aisladas reproductivamente de otros grupos"[7].

2.12. Especie epipelágica

Es aquella especie marina que habita en la zona epipelágica, misma que abarca desde la superficie hasta los 200 metros de profundidad.

2.13. Cadena trófica

Es el proceso de transferencia de energía alimenticia a través de una serie de organismos, en el que cada uno se alimenta del precedente y es alimento del siguiente. Cada cadena se inicia con un productor u organismo autótrofo, los demás integrantes de la cadena se denominan consumidores y el último eslabón de la cadena corresponde a los descomponedores, los cuales actúan sobre los organismos muertos, degradan la materia orgánica y la transforman nuevamente en inorgánica devolviéndola al suelo (nitratos, nitritos, agua) y a la atmósfera (dióxido de carbono)[8].

2.14. Batimetría

La batimetría, aplicada al medio marino, es la medición de las profundidades marinas para determinar la topografía del fondo del mar. Su medición implica la obtención de datos con los valores de profundidad y la posición de cada uno de los puntos muestreados. Estos puntos de posición al igual que en la altimetría, están conformados por coordenadas de puntos X, Y y Z.

2.15. Salinidad

La salinidad en el océano se define como los gramos de sal por 1000 gramos de agua, donde un gramo de sal por 1000 gramos de agua se define como una unidad práctica de salinidad o una PSU. Se necesitan observaciones de salinidad para calcular estimaciones de los transportes oceánicos de agua dulce y otras propiedades en la cuenca a escalas globales. La salinidad también proporciona un buen indicador de los cambios en el ciclo del agua, ya que indica el cambio en el agua dulce debido a la diferencia entre la precipitación y la evaporación. Junto con la temperatura, es un factor importante que contribuye a los cambios en la densidad del agua de mar y, por lo tanto, a la circulación del océano.

2.16. Temperatura de la superficie del mar

Conocida como SST, es la temperatura cercana a la superficie del mar, la cercanía de la medición a la superficie dependerá del método de medición empleado. La SST provee información fundamental del sistema climático global y es de especial importancia para el estudio de ecosistemas marinos.

Los datos de SST son especialmente útiles para identificar el inicio de los ciclos de El Niño y La Niña. Durante El Niño las temperaturas en el pacífico cerca del ecuador son más cálidas de lo normal, mientras que durante La Niña, la misma área experimenta temperaturas marinas más frías. Se pueden observar los resultados de las mediciones en simulaciones como la que se presenta en la figura 2.6.

2.17. Producción primaria

Se denomina así a la producción de materia orgánica que realizan los organismos autótrofos a través de los procesos de fotosíntesis y biosíntesis. Es el punto de partida de la circulación de energía y nutrientes en las cadenas tróficas.

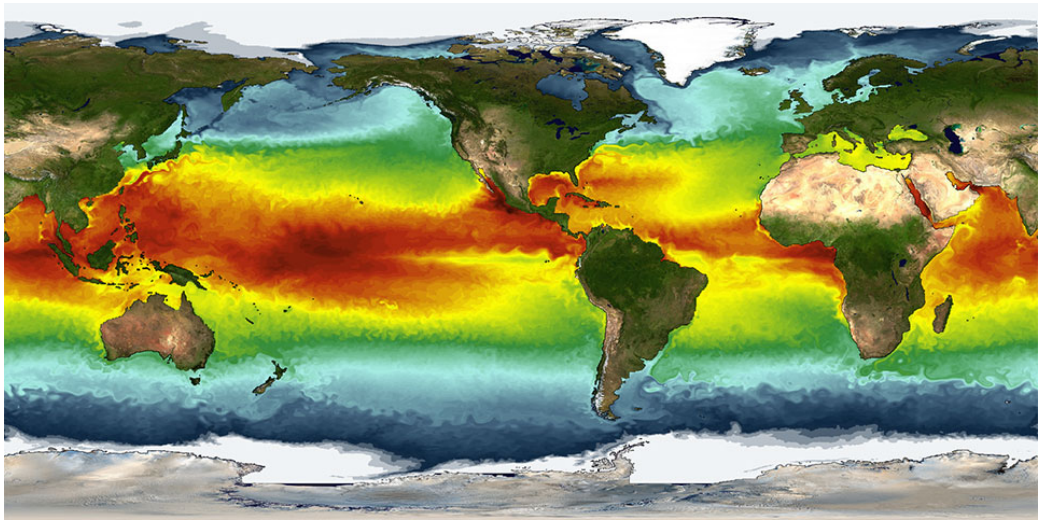

Figura 2.6: Simulación de las temperaturas marinas superficiales del *Geophysical Fluid Dynamics Laboratory* (GFDL).

2.18. Producción primaria bruta (GPP)

La GPP está definida como la tasa a la cual la vegetación captura el dióxido de carbono en un tiempo dado a través de la fotosíntesis, lo que resulta en productos primarios translocables y almacenables.

2.19. Producción primaria neta (NPP)

La NPP es la ganancia neta de carbono de la vegetación en un tiempo dado y está definida como la diferencia entre GPP y la respiración autotrófica como se muestra en (2.6).

$$NPP = GPP - R_a \tag{2.6}$$

2.20. Clorofila

La clorofila es un pigmento verde existente en las plantas, algunas algas y bacterias que permite llevar a cabo el proceso de fotosintesis que es la conversion de energia luminosa en energia quimica. Proviene del vocablo *chloros* que significa “verde” y *fylon* que significa “hoja”. Existen diferentes tipos de clorofila, A que se encuentra presente en la mayoria de los vegetales

y es la encargada de absorber la luz durante la fotosintesis; la B que se encuentra presente en los cloroplastos, se encarga de absorber la luz de otra longitud y transfiere la energía a la clorofila A; la C está presente en los cloroplastos de las algas pardas, las diatomeas y, por último, la D se halla únicamente en las algas rojas [9]. En la figura 2.7 se puede observar un cloroplasto, es decir, el orgánulo de las células vegetales y de las algas que contiene la clorofila.

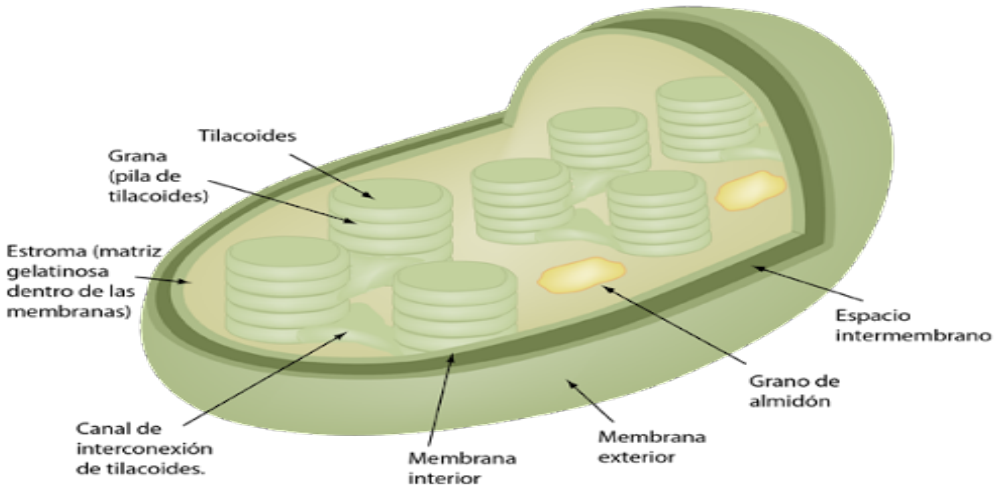

Figura 2.7: Estructura de un cloroplasto

2.21. *Vertically generalized production model* (VGPM)

Es un algoritmo comúnmente usado para estimar productividad primaria neta (NPP) "basado en clorofila", la base de este tipo de algoritmos es que la NPP varía de forma predecible con la concentración de clorofila. Debido a que la NPP es una tasa y la clorofila una reserva permanente, la derivación de la primera a partir de la última requiere un término de tasa específicamente una eficiencia de asimilación específica de la clorofila para la fijación de carbono. La descripción de este término de tasa es la incertidumbre más importante de todos los modelos basados en clorofila. El VPGM emplea una variable denominada Pb_{opt} que es la producción primaria neta máxima encontrada dentro de una columna de agua dada y expresada en unidades de miligramos de carbono fijado por mg de clorofila por hora [10].

2.22. Cambio climático

De acuerdo con el glosario del *Intergovernmental Panel on Climate Change* (IPCC), el cambio climático hace referencia a una variación del estado del clima identificable (p.ej. mediante pruebas estadísticas) en las variaciones del valor medio o en la variabilidad de sus propiedades, que persiste durante periodos prolongados, generalmente décadas o periodos más largos. El cambio climático puede deberse a procesos internos naturales o a forzamientos externos, tales como modulaciones de ciclos solares, erupciones volcánicas y cambios antropogénicos persistentes de la composición de la atmósfera o del uso de la tierra [11].

2.23. Cambio climático antropogénico

Por Cambio Climático Antropogénico se entiende la alteración que experimentan los diversos climas terrestres por el sobrecalentamiento causado al acumularse en la atmósfera ciertos gases emitidos por la quema de combustibles fósiles (carbón, petróleo y gas).

2.24. *El Niño Southern Oscillation* (ENSO)

El Niño Southern Oscillation, también conocido como ENSO, es una fluctuación periódica en la temperatura superficial marina (El Niño) y la presión del aire de la atmósfera suprayacente (Oscilación del sur) a través del Océano Pacífico ecuatorial.

Aunque las causas exactas de inicio de un evento ENSO cálido o frío no se comprenden completamente, los dos componentes de ENSO: la temperatura de la superficie del mar y la presión atmosférica están estrechamente relacionados. Durante un evento de El Niño, los vientos alisios del este que convergen a través del Pacífico ecuatorial se debilitan. Esto, a su vez, ralentiza la corriente oceánica que aleja el agua superficial de la costa occidental de América del Sur y reduce el afloramiento de agua fría rica en nutrientes del océano más profundo, aplanando la termoclina y permitiendo que el agua superficial cálida se acumule en la parte oriental de la cuenca [12].

Capítulo 3

Metodología

3.1. Propuesta de solución

Dos de los principales problemas cuando se trabaja con problemas SDM son la limitada cantidad de registros de presencia, y el sesgo del esfuerzo de muestreo respecto a la distribución real subyacente. Los algoritmos más populares para problemas de SDM como MAXENT resuelven el problema del sesgo con una técnica conocida como *environmental niche modeling* (Modelado de nicho ambiental) que consiste en predecir la distribución geográfica a partir de una representación del espacio ambiental que en la mayoría de los casos está representado como condiciones climáticas. Un problema de ese enfoque es que restringe fuertemente las variables sin una base teórica clara, ya que los patrones que afectan a cada especie son tan complejos y diversos que pueden estar vinculados a las condiciones del clima pero también a otro tipo de factores como los abióticos. En este último punto es donde el enfoque del aprendizaje automático es una excelente opción dado que las arquitecturas profundas proveen flexibilidad y favorecen los efectos de interacciones de alto orden entre las variables de entrada sin restringir la funcionalidad [1]. Este problema será trabajado como un problema de clasificación donde a partir de patrones de entrada el modelo (previamente entrenado) determinará si ese conjunto de entrada pertenece a la clase 1 o a la clase 2, es decir, será capaz de predecir la presencia o ausencia respectivamente de la especie bajo estudio a partir de variables climáticas pero sin estar restringido a estas, es decir, la flexibilidad propia de este tipo de arquitecturas hará viable la adición de otro tipo de patrones de entrada sin afectar el desempeño de predicción. Existen diferentes arquitecturas de redes neuronales (convolucionales, perceptrón multicapa, de base

radial, entre otras), se propone utilizar el perceptrón multicapa que ha demostrado ser útil en la resolución de problemas de clasificación y predicción tal como el que se presenta [13]. Una vez obtenido un conjunto de datos válido, se usará una mayor parte para entrenar el modelo y el resto del conjunto para validación.

3.2. Selección de variables de entrada

La Sardina del Pacífico es del grupo de especies de los pelágicos menores, y como tal, es altamente sensible a cambios en el entorno, por lo que se le considera un indicador de respuesta del ecosistema a cambios en las condiciones del ambiente[3]. Para que el modelo sea capaz de predecir la presencia o la ausencia de la especie bajo un conjunto de características primero se tienen que seleccionar los parámetros del ecosistema que tienen más peso para la Sardina, los cuales fungirán como las variables de entrada en el modelo, el cual durante el entrenamiento asignará los pesos adecuados a cada variable para determinar la influencia que tienen estas variables en la presencia o ausencia de la especie. Se seleccionaron las siguientes variables de entrada:

- Temperatura marina superficial (SST). Al ser una especie epipelágica, la sardina se ubica entre la superficie y los 200 metros de profundidad por lo que la temperatura es la principal variable que afecta no solo su distribución si no también otros aspectos como su crecimiento y desove.
- Productividad primaria neta (NPP): La productividad primaria neta es un indicador altamente fiable de la cantidad de alimento disponible para la Sardina ya que es un consumidor primario en la cadena trófica.
- Salinidiad marina superficial (SSS): Muchos autores sugieren esta variable como indicador de osmoregulación y mobilidad de especies pelágicas menores[3].

3.3. Obtención de datos

En primera instancia se recopilaron registros de presencia de Sardina del Pacífico en el Océano Pacífico y Golfo de California dentro del polígono especificado en alcances y limitaciones

tal como se ilustra en el mapa de la figura 3.1 así como el periodo mencionado dentro de la misma sección.

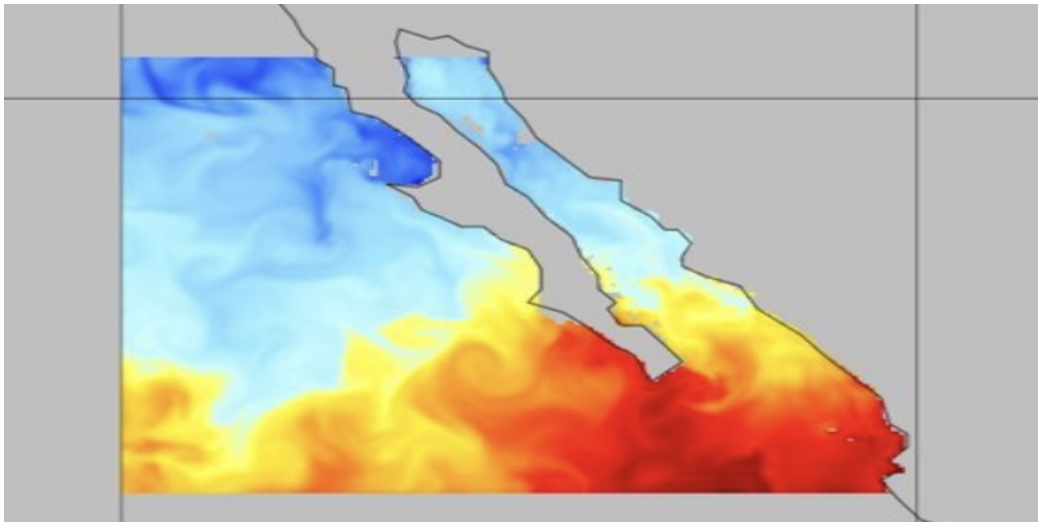

Figura 3.1: Mapa de temperatura de la zona marina de estudio

Una vez obtenidos los registros, se descargó un archivo de batimetría con el objetivo de verificar que las coordenadas en cada registro se encontrasen en mar, es decir, si el valor de la batimetría en las coordenadas asociadas a un registro es negativo, eso implica que ese punto se encuentra sobre el nivel del mar.

Posteriormente se descargaron archivos de las 3 variables de entrada (SST, SSS, NPP) para asociar el punto coordenado de cada registro a sus respectivos valores de temperatura, salinidad, productividad primaria neta y batimetría. En el caso de los datos de NPP, se utilizaron los datos generados a través de el algoritmo VGPM, que al ser un modelo basado en clorofila, tiene un alto grado de confianza pues la productividad primaria neta varía de forma predecible con la concentración de clorofila del mar.

Las fuentes de los datos, los formatos de descarga de los archivos y los detalles se desglosan en la tabla 3.1.

3.4. Creación del conjunto de datos

El primer paso para la generación de un conjunto de datos válido es la extracción de los registros de presencia y su validación, para ello se utilizaron los datos descargados de OBIS, GBIF y de la Colección Ictiológica CICIMAR-IPN que vienen en formato CSV y el archivo de

Variable	Fuente	Tipo de archivo
Registros de presencia	Ocean Biodiversity Information System (OBIS), Colección Ictiológica CICIMAR-IPN y Global Biodiversity Information System (GBIF).	Archivos tipo CSV
Temperatura (SST)	HYCOM	Archivos NetCDF4
Salinidad (SSS)	HYCOM	Archivos NetCDF4
Productividad primaria neta (NPP)	Oregon State University (Ocean Productivity)	Archivos HDF
Batimetría	General Bathymetric Chart of the Oceans (GEBCO)	Archivos NetCDF

Tabla 3.1: Datos obtenidos y sus respectivas fuentes

batimetría que viene en formato NetCDF, el objetivo es verificar que las coordenadas de los registros obtenidos sean, de hecho, coordenadas oceánicas, esto es de particular importancia porque más adelante en el proceso se usarán esas mismas coordenadas para obtener el resto de parámetros de ese punto específico.

Se creó un algoritmo en Python cuyo objetivo en primera instancia fue verificar el valor de la batimetría de cada registro, los puntos con batimetría negativa fueron descartados, en segunda instancia la resolución espacial de los archivos de temperatura, salinidad y productividad primaria es de 0.08°, por lo tanto el algoritmo elimina registros adyacentes en la misma celda para evitar sesgos en los datos. Después de la primer parte del filtro se redujo el número de registros de cerca de 500 a 323, en la proyección de la figura 3.2 se puede apreciar la distribución espacial antes del segundo filtro. Sin embargo tras observar la concentración de registros en zonas particulares, se implementó el segundo filtro en función de la resolución de los datos.

Hasta esta parte del proceso se tiene un conjunto de datos con 69 registros validados, cada uno de ellos cuenta con: latitud, longitud, fecha y batimetría. En este punto se desarrolló un segundo algoritmo cuyo propósito es recorrer los 13 archivos tipo NetCDF4 que contienen la información de temperatura y salinidad de cada día a lo largo del año correspondiente, en busca de las coordenadas de los registros y así extraer los valores necesarios para añadirlos al conjunto.

Figura 3.2: Proyección de registros con coordenadas validadas

Sin embargo, tras añadir los nuevos valores al conjunto, surgió un problema interesante, algunos de los valores extraídos eran nulos, es decir, se desconocía el valor de la temperatura o salinidad de ese punto particular, se especula que esos valores pueden faltar debido a errores en el proceso de medición o dada la cercanía a la costa de los puntos de interés es posible que se haya truncado el valor ya que hay porciones de tierra dentro de la celda que se está tratando de medir.

Para dar solución a este problema se empleó un modelo bilineal por mínimos cuadrados para estimar los valores nulos de temperatura y salinidad, haciendo uso de los valores conocidos en el resto de coordenadas del plano de esa fecha en particular, tal como describe a continuación:

Dados conjuntos $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ y $\{z_1, z_2, \ldots, z_n\}$, donde x, y, z representan los valores conocidos de longitud, latitud y temperatura/salinidad respectivamente y donde n es el número de coordenadas con valores conocidos, se desea encontrar valores $\alpha_1, \alpha_2, \alpha_3$ tales

que $\forall i \in \{1, 2, \ldots, n\}$ y $f(x_i, y_i) = \alpha_1 + \alpha_2 x + \alpha_3 y \approx z$; Idealmente se quiere resolver:

$$\begin{aligned} \alpha_1 + \alpha_2 x_1 + \alpha_3 y_1 &= z_1 \\ \alpha_1 + \alpha_2 x_2 + \alpha_3 y_2 &= z_2 \\ &\vdots \\ \alpha_1 + \alpha_2 x_n + \alpha_3 y_n &= z_n \end{aligned} \tag{3.1}$$

que también se puede expresar como $A\alpha = b$ con:

$$A = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{pmatrix} \quad ; \quad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \quad ; \quad b = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \tag{3.2}$$

Para que el sistema de ecuaciones tenga solución exacta, se requiere que el número de variables sea igual al número de ecuaciones linealmente independientes entre sí, es decir, 3 ecuaciones y 3 variables, pero ya que $n > 3$, no es una opción realista, por lo que se busca un modelo aproximado, tal que $A\alpha \approx b$, es decir, $A\alpha - b \approx 0$, entonces el problema se puede replantear como el de minimizar la norma del vector $A\alpha - b$, lo que definimos como la función de error:

$$E(\alpha) = ||A\alpha - b||^2 \tag{3.3}$$

cuya solución se ha estudiado ampliamente en Álgebra lineal y su solución se puede encontrar el libro de Álgebra lineal de Grossman[14] como:

$$\left(A^T A\right) \alpha = A^T b \tag{3.4}$$

donde $A^T A$ es una matriz invertible siempre que A tenga rango 3, es decir, que los vectores columna $(1, 1, \ldots, 1)^T$, $(x_1, x_2, \ldots, x_n)^T$ y $(y_1, y_2, \ldots, y_n)^T$ que le componen sean linealmente independientes entre sí. Debido a lo anterior se puede inferir que se requieren circunstancias irreales para que este método falle en este problema particular, ya que requeriría, por ejemplo, que los vectores $(x_1, x_2, \ldots, x_n)^T$ y $(y_1, y_2, \ldots, y_n)^T$ tuviesen todos sus elementos en 0, que resulta absurdo ya que implica todas las coordenadas del archivo son $(0, 0)$.

A partir de (3.4) sustituimos las definiciones de (3.2):

$$\begin{pmatrix} n & \sum x & \sum y \\ \sum x & \sum x^2 & \sum xy \\ \sum y & \sum xy & \sum y^2 \end{pmatrix} \quad \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} \sum z \\ \sum xz \\ \sum yz \end{pmatrix} \tag{3.5}$$

y obtenemos un sistema de ecuaciones 3x3 como se expresa en (3.6):

$$\begin{aligned} \alpha_1 \sum n + \alpha_2 \sum x + \alpha_3 \sum y &= \sum z \\ \alpha_1 \sum x + \alpha_2 \sum x^2 + \alpha_3 \sum xy &= \sum xy \\ \alpha_1 \sum y + \alpha_2 \sum xy + \alpha_3 \sum y^2 &= \sum yz \end{aligned} \tag{3.6}$$

Una vez calculados los valores de α_1, α_2 y α_3 se sustituye en $z \approx \alpha_1 + \alpha_2 x + \alpha_3 y$ y se obtiene un valor estimado del valor real desconocido.

A partir de lo anterior se diseñó un algoritmo que es capaz de calcular α y usar este modelo para estimar z de forma automática y añadir esos datos al conjunto principal.

Por último se requiere añadir los valores de productividad primaria neta (NPP) al conjunto final, se procesó de forma independiente a las dos variables anteriores porque los datos descargados de Oregon State University vienen en formato HDF, el cual es un formato poco usado y requirió investigación adicional para acceder al contenido. Una vez determinadas las herramientas necesarias para procesar esta última variable se escribió el algoritmo adecuado para extraer los datos y se presentó el mismo inconveniente de datos ausentes en algunos puntos, por lo que se extrapoló y adaptó el modelo bilineal de la sección anterior. Al final del procesamiento de datos se obtuvo un conjunto de 69 registros, cada uno con latitud, longitud, fecha, batimetría, temperatura, salinidad, productividad primaria neta y una etiqueta llamada presencia y cuyo valor para todos los registros es 1.

3.5. Creación de pseudoausencias

Un modelo de red neuronal artificial de clasificación requiere que los datos de entrenamiento contengan muestras de todas las clases que se espera clasificar una vez entrenado, sin embargo, las bases de datos de especies contienen unicamente registros de presencia, por lo tanto, se requiere crear pseudoausencias que el modelo pueda utilizar para su entrenamiento.

Se propuso por lo tanto, crear pseudoausencias a partir del fundamento teórico de la especie y de los datos ya recolectados previamente, es decir, buscar coordenadas que contuvieran un valor de temperatura, salinidad o productividad primaria neta teóricamente adverso para la subsistencia de la Sardina y añadir el resto de variables de ese punto, formando así un registro de ausencia a partir de datos marinos reales. Utilizando este procedimiento se extrajo un total

de seis mil ochocientos sesenta y seis registros de pseudoausencias de los once años de estudio y cada uno estos con una etiqueta adicional de presencia con valor de 0.

3.6. Programación y configuración del modelo

La programación del modelo de red neuronal artificial se llevó a cabo en Python 3.6.9, con la asistencia de las librerías Numpy, Pandas y Scikit-Learn. El algoritmo consiste en cargar los archivos de presencia y pseudoausencia, ambos conjuntos se dividen en dos subconjuntos de 70 % y 30 %, los grupos de 70 % serán destinados a entrenamiento y los grupos de 30 % para validación, se replican los elementos de entrenamiento de presencia de forma uniforme hasta balancear los registros de las clases, se unen los dos subconjuntos de entrenamiento en uno solo, se crea instancia el perceptrón multicapa al cual se le entregan los datos de entrenamiento ordenados de forma aleatoria, se inicia el entrenamiento de la red neuronal y posteriormente se valida con un conjunto formado por la unión de los dos subconjuntos de validación.

Se importó de la librería Scikit-Learn la clase MLPClassifier que proporciona las funciones necesarias para crear un perceptrón multicapa de clasificación, en la tabla 3.2 se muestran los parámetros que esenciales para construir el modelo y una breve descripción de su función.

Parámetro	Descripción
Hidden_layer_sizes	Indica la cantidad de neuronas en la capa oculta.
Activation	Especifica la función de activación de la capa oculta de las cuatro disponibles: función sigmoidea logística, función identidad, tangente hiperbólica y función de unidad lineal rectificada (ReLU).
Solver	Es el solucionador interno para la optimización de pesos, las opciones disponibles son: lbfgs, sgd y adam.
Max_iter	Máximo número de iteraciones. El solucionador itera hasta converger o hasta llegar el número de iteración indicado por este parámetro.

Tabla 3.2: Parámetros necesarios para la instancia del perceptrón multicapa de clasificación

Capítulo 4

Resultados

4.1. Parámetros de instancia del modelo

Para encontrar un conjunto de parámetros de instancia que produjera un desempeño mejor que el encontrado en el estado del arte, se necesitó valorar los resultados obtenidos a partir de cambios en los parámetros del perceptrón multicapa de forma individual y posteriormente combinar esos resultados individuales para general una combinación que funcione de buena manera.

La selección de parámetros se analizó de la siguiente forma:

- Hidden_layer_sizes: Se probó este parámetro desde 10 de neuronas hasta 500, encontrando el mejor resultado en 100 neuronas.
- Activation: Se realizaron pruebas con las 4 funciones de activación disponibles, encontrando el mejor resultado con la función de activación sigmoidea logística.
- Solver: Dentro de los solucionadores para optimización de los pesos, se obtuvo mejores resultados con lbfgs que es un optimizador de la familia de métodos cuasi-newtonianos que ha demostrado converger más rápido y desempeñarse mejor para conjuntos de datos relativamente pequeños.
- Max_iter: La cantidad de épocas o iteraciones para entrenamiento se probó desde 100 hasta 50000, encontrando el mejor resultado con 29000 épocas, las suficientes para garantizar

un buen desempeño pero cuidando que no se sobreentrene, lo cual es perjudicial para el modelo.

La elección de una función de activación para la capa oculta puede estar sujeta al tipo de problema que se está trabajando o al modelo de red neuronal empleado, se sabe por ejemplo, que la función de activación ReLU es la opción que usualmente provee mejores resultados en modelos de perceptrón multicapa o redes neuronales convolucionales y también se sabe que en problemas de regresión, ReLU suele ser la mejor alternativa pero en problemas de clasificación no binarios la función sigmoidea logística es mejor. Sin embargo, cada problema es único y los comportamientos generados con cada función de activación pueden variar, por lo que, lo adecuado es sin duda realizar pruebas con las diferentes opciones disponibles y evaluar el desempeño.

4.2. Evaluación del desempeño de clasificación

Dentro de las métricas existentes para evaluar el desempeño de un modelo de clasificación, se seleccionaron aquellas obtenibles a partir de la matriz de confusión y la métrica de área bajo la curva (AUC). Este conjunto de métricas permite observar e interpretar los resultados desde diferentes perspectivas, con el fin de hacer los ajustes más precisos al modelo; las métricas utilizadas son las siguientes:

- Exactitud: Indica el porcentaje de elementos clasificados correctamente.
- Precisión: Indica que porcentaje de las identificaciones positivas es correcto.
- *Recall*: Es la capacidad del clasificador de encontrar todos los positivos.
- Puntuación F1: Es una media ponderada entre precisión y Recall.
- Curva AUC/ROC: Representa el grado o medida de separabilidad, es decir, cuánto el modelo es capaz de distinguir entre las clases.

La teoría general sugiere que en un problema de clasificación binario, la función de activación ideal es la ReLU, sin embargo, las pruebas realizadas indican que en este problema particular se obtiene mejores resultados con la función sigmoidea e incluso con la tangente hiperbólica. Se puede dar el caso en que los resultados entre dos funciones sean tan parecidos que la primera

opción ofrezca valores ligeramente mejores en algunas métricas respecto a la segunda y valores ligeramente peores el resto, en ese caso lo ideal sería analizar qué métrica es más importante en base a la interpretación final que aporta a la investigación, ¿Es prioridad garantizar que todos los positivos sean identificados? ¿Es mejor identificar pocos positivos pero garantizar que haya pocos o ningún falso positivo? ¿Se busca la mayor precisión del modelo sin importar que clasifique mejor los positivos o negativos? Este tipo de preguntas son esenciales para valorar mejor los resultados.

En la tabla 4.1 se muestran los mejores resultados obtenidos por cada una de las cuatro funciones de activación disponibles para cada métrica de evaluación de interés.

	ReLU	Sigmoidea	Identidad	Tanh
Precisión	96.5485	99.28.05	87.849	99.2329
Recall	100	100	82.0772	100
Puntuación F1	98.2439	99.6389	84.8651	99.615
Exactitud	98.2082	99.6368	85.3268	99.6125
Curva AUC/ROC	98.2038	99.6359	85.3347	99.6116

Tabla 4.1: Comparación de desempeño por función de activación

Después de observar los resultados, se notó un desempeño muy parecido entre la función sigmoidea y tangente hiperbólica pero en todas las métricas la sigmoidea fue ligeramente superior.

Capítulo 5

Conclusiones

La métrica utilizada en el estado del arte para evaluar el desempeño con MAXENT fue la curva AUC/ROC con un valor de 94 % contra datos de validación, el valor de la misma métrica obtenido en este trabajo con el perceptrón multicapa es de 99.6359 %, lo cual muestra un desempeño significativamente mejorado, tal como se especulaba en la hipótesis de este trabajo. Por otro lado, resulta interesante someter a discusión la información que nos ofrece el resto de métricas de clasificación.

En primera instancia, el valor del *Recall* es consistentemente de 100 %, esto puede llamar la atención, ya que quiere decir que el modelo garantiza que todas las presencias de la especie son encontradas con seguridad, sin embargo, el valor que resulta más valioso con fines económicos y comerciales es la precisión, pues esta métrica indica que porcentaje de las identificaciones positivas es correcto, en otras palabras, si un modelo como este se utilizara para dirigir pescadores a las zonas con presencia de Sardina del Pacífico, entonces sería de gran importancia que el mayor porcentaje de predicciones positivas sea correcto, de otro modo implicaría pérdidas económicas en dicha actividad. Se sugiere encarecidamente incluir múltiples métricas de evaluación de este tipo de modelos que ofrezcan diversas perspectivas, ya que algunas métricas pueden parecer buenas pero su interpretación en el problema real no es de gran utilidad y otras si lo son.

Este trabajo demostró que la arquitectura de red neuronal artificial provee mejores resultados de clasificación en problemas de modelado de distribución de especies (SDM) que las técnicas más populares, además con la posibilidad de no limitar el conjunto de patrones de entrada a aquellos asociados a las condiciones climáticas si no incluir factores abióticos o de interacciones de otro tipo.

Bibliografía

[1] C. Botella, A. Joly, P. Bonnet, P. Monestiez, F. Munoz, A. Joly, P. Bonnet, P. Monestiez, and F. Munoz, "A Deep Learning Approach to Species Distribution Modelling." [Online]. Available: https://doi.org/10.1007/978-3-319-76445-0{_}10

[2] E. H. Allison, A. L. Perry, M.-C. Badjeck, W. Neil Adger, K. Brown, D. Conway, A. S. Halls, G. M. Pilling, J. D. Reynolds, N. L. Andrew, and N. K. Dulvy, "Vulnerability of national economies to the impacts of climate change on fisheries," *Fish and Fisheries*, vol. 10, no. 2, pp. 173–196, jun 2009. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-2979.2008.00310.x

[3] D. Petatán-Ramírez, M. Á. Ojeda-Ruiz, L. Sánchez-Velasco, D. Rivas, H. Reyes-Bonilla, G. Cruz-Piñón, H. N. Morzaria-Luna, A. M. Cisneros-Montemayor, W. Cheung, and C. Salvadeo, "Potential changes in the distribution of suitable habitat for Pacific sardine (Sardinops sagax) under climate change scenarios," *Deep-Sea Research Part II: Topical Studies in Oceanography*, vol. 169-170, no. July, p. 104632, 2019. [Online]. Available: https://doi.org/10.1016/j.dsr2.2019.07.020

[4] Conapesca, "Anuario estadístico de acuacultura y pesca 2018," Sinaloa, México, Tech. Rep., 2018. [Online]. Available: https://www.conapesca.gob.mx/work/sites/cona/dgppe/2018/ANUARIO{_}2018.pdf

[5] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 1st ed., Nicole Tache, Ed. O'Reilly Media, Inc., 2017.

[6] R. G. Mateo, Á. M. Felicísimo, and &. J. Muñoz, "REVISTA CHILENA DE HISTORIA NATURAL Species distributions models: A synthetic revision," *Revista Chilena de Historia Natural*, vol. 84, pp. 217–240, 2011.

[7] A. Melic, J. J. De Haro, M. Méndez, and I. Ribera, "Evolución y Filogenia de Arthropoda," 1999. [Online]. Available: http://sea-entomologia.org/PDF/BOLETIN{_}26/B26-000-000.pdf

[8] M. Spinelli, "Cadena alimentaria (= Cadena trófica)." [Online]. Available: https://www.mendoza.conicet.gov.ar/portal/enciclopedia/terminos/CadeAlim.htm

[9] R. Santiago, R. Velázquez, H. Becerra, G. Jiménez, and V. Villarreal, "Extracción y cuantificación de clorofila en hojas comestibles del estado de Tabasco RESUMEN," vol. 4, 2019.

[10] "Ocean Productivity: Custom Products." [Online]. Available: http://sites.science.oregonstate.edu/ocean.productivity/vgpm.model.php

[11] M. Babiker, H. De Coninck, S. Connors, R. Van Diemen, and Djalantem Riyanti, "Informe especial del IPCC sobre los impactos del calentamiento global de 1,5C con respecto a los niveles preindustriales y las trayectorias correspondientes que deberían seguir las emisiones mundiales de gases de efecto invernadero, en el contexto del r," Tech. Rep., 2018.

[12] NOAA, "El Nino/Southern Oscillation (ENSO) Technical Discussion — Teleconnections — National Centers for Environmental Information (NCEI)," 2020. [Online]. Available: https://www.ncdc.noaa.gov/teleconnections/enso/enso-tech.php

[13] G. Singh, A. Mishra, and D. Sagar, "An Overview Of Artificial Intelligence," *An overview of artificial intelligence*, vol. 2, no. January, p. 4, 2013.

[14] S. Grossman, *Álgebra lineal*, sexta edic ed. McGraw-Hill, 2008.