Herramienta operativa para el pronóstico de valores de clorofila en zonas de mar abierto utilizando series temporales MODIS y técnicas de aprendizaje máquina.

Carlos Mario Nasayó Valverde

Sebastián Camilo Medina González

Corporación Universitaria Comfacauca - Unicomfacauca

Facultad de Ingenierías

Ingeniería de Sistemas

Popayán, Colombia

2023

Herramienta operativa para el pronóstico de valores de clorofila en zonas de mar abierto utilizando series temporales MODIS y técnicas de aprendizaje máquina.

Carlos Mario Nasayó Valverde

Sebastián Camilo Medina González

Proyecto de grado para optar al título de Ingeniero de Sistemas

Director: MSc. Julián Fernando Muñoz Ordóñez

Codirector: Ing. Cristian Muñoz Ordóñez

Corporación Universitaria Comfacauca - Unicomfacauca

Facultad de Ingenierías

Ingeniería de Sistemas

Popayán, Colombia

2023

DATOS GENERALES	
Facultad:	Ingeniería
Programa académico:	Ingeniería de Sistemas
Título del proyecto:	Herramienta operativa para el pronóstico de valores de clorofila en zonas de mar abierto utilizando series temporales MODIS y técnicas de aprendizaje máquina.
Grupo de investigación:	Mind computación e informática aplicada Unicomfacauca
Línea de investigación:	Inteligencia Artificial & Big Data

Proyecto institucional asociado:	Vigilancia de pacientes para prevenir eventos adversos relacionados con caídas en ambientes hospitalarios utilizando inteligencia artificial. (cambiar)
Fecha de aprobación del perfil:	
Fecha de elaboración de la propuesta:	15/11/2020
Fecha de radicación:	

Nota de aceptación

Aprobado por el comité de grado en cumplimiento de los requisitos exigidos por la Corporación Universitaria Comfacauca - Unicomfacauca para optar al título de Ingeniero de Sistemas.

_____

Ingeniero

_____

Ingeniero

Tabla de contenido

1 INTRODUCCION....10

1.1 PLANTEAMIENTO PROBLEMA 10

1.2 JUSTIFICACIÓN 12

1.3 OBJETIVOS 13

1.3.1 Objetivo General 13

1.3.2 Objetivo Específicos 13

2 MARCO TEÓRICO Y ESTADO DEL ARTE....15

2.1 Variable estudio-Clorofila 15

2.2 Descripción de las Imágenes satelitales 16

2.2.1 Definición del sensor MODIS 17

2.3 Proceso de transformación Corregistro de Imágenes 17

2.4 Técnicas de aprendizaje automático Algoritmos de predicción 18

2.4.1 Algoritmos de predicción Redes neuronales artificiales 19

2.4.2 Modelo de regresión lineal BAYESIANRIDGE 20

2.5 Conjunto de datos (Dataset) 21

2.6 Algoritmo de aprendizaje automático (Redes Neuronales) 22

2.7 Métrica de valor cuadrático medio (coeficiente de correlación) 23

2.8 Librería para manejo imágenes (NetCdf4) 23

2.9 Corregistro de imágenes satelitales 24

2.10 Técnica de aprendizaje automático Deep Learning 25

2.11 Entrenamiento de imágenes satelitales 26

2.12 Librería para el procesamiento de imágenes georreferenciadas (GDAL) 27

2.13 Librería par visión por computadora (OpenCV) 28

2.14 Librería de aprendizaje automático Scikit Learn 28

2.15 ESTADO DEL ARTE 29

3 METODOLOGIA PARA LA PUESTA PUNTO DE IMÁGENES SATELITALES MODIS 38

3.1 Metodología desarrollada del proyecto 38

3.2 Metodología de descarga imágenes satelitales 40

3.3 Corregistro de las imágenes satelitales 43

3.4 Verificación de las imágenes 46

4 IMPLEMENTACION SOFTWARE DE LA HERRAMIENTA OPERATIVA 48

4.1 Módulo para el procesamiento de NetCDFmodule. 48

4.2 Módulo para el procesamiento de ImageProcessing 49

4.3 Módulo para el procesamiento de builderDataset 50

4.4 Módulo para el procesamiento de Trainer 52

4.5 Módulo para el procesamiento de Tester 53

4.6 Módulo para el procesamiento de Predictor 55

5 RESULTADOS Y ANALISIS 57

5.1 Descripción del modelo 57

5.2 Descripción de la herramienta 58

5.3 Resultados 63

6 CONCLUSIONES Y TRABAJOS A FUTURO 66

6.1 Conclusiones 66

6.2 Trabajos a futuro 67

7 REFERENCIA 68

8 ANEXO 73

LISTA DE FIGURAS

Figura 1 Selección de característica del sitio web ocean color 41
Figura 2 Confirmación de petición para la descarga de imágenes satelitales 42
Figura 3 Imagen de clorofila descargada y visualizada en el software SEADAS 43
Figura 4 Ejecución interfaz Corregistro de imágenes satelitales MODIS 45
Figura 5 Línea ejecutable en la consola 45
Figura 6 Ejecución de interfaz gráfica, verificación de contenido imágenes 46
Figura 7 Metadatos de verificación de imagen satelital 47
Figura 8 Módulo NetCDF 49
Figura 9 Módulo ImageProcessing 50
Figura 10 Módulo BuilderDataset 52
Figura 11 Módulo Trainer 53
Figura 12 Módulo Tester 54
Figura 13 Módulo Predictor 56
Figura 14 Comparación de modelos 58
Figura 15 Imagen binaria con Vecindad Variable nxn 59
Figura 16 Posibles combinaciones de vecindad 2 60
Figura 17 Conformación conjunto de datos clorofila con vecindad 61
Figura 18 Posibles modelos con diferentes conectividades 62
Figura 19 Comparación de valores perdidos- valores pronosticados 64
Figura 20 Formula que se implementó para la Confianza del modelo 65

RESUMEN

Para mejorar la pesca en Colombia, es esencial contar con información precisa sobre la concentración de peces en los mares. Esto permitirá un mayor control sobre la pesca y evitará el uso de métodos inapropiados, ayudando así al medio ambiente marítimo para la reproducción de los peces. Una forma de obtener esta información es mediante el uso de imágenes generadas por el satélite MODIS, que ha recolectado información desde 2001, tomando fotos tanto en el día como en la noche, observando desastres naturales, cambios climáticos y anomalías. Además, permite observar los niveles de clorofila en el agua, los cuales se relacionan con la presencia de bancos de peces. Sin embargo, en los mares de Colombia, la presencia de nubes durante todo el año dificulta la detección de la clorofila en las imágenes satelitales, lo que impide observar dónde se encuentran los peces. Para solucionar este problema, se propone realizar un estudio con todo el conjunto de dos que el satélite MODIS ha estado recolectando durante más de 20 años para aplicar técnicas de procesamiento de imágenes y generar un conjunto de datos precisos, con el cual se construyó un modelo de aprendizaje automático para

el pronóstico de valores de clorofila perdidos debido a la alta cobertura nubosa. Asimismo, se pueden manejar métricas como el error cuadrático medio para evaluar la precisión del modelo. De esta manera, se podrá ayudar a las personas involucradas en la pesca y reducir el impacto ambiental.

ABSTRACT

To improve fisheries in Colombia, accurate information on the concentration of fish in the seas is essential. This will allow greater control over fishing and avoid the use of inappropriate methods, thus helping the marine environment for fish reproduction. One way to obtain this information is by using imagery generated by the MODIS satellite, which has collected information since 2001, taking photos both day and night, observing natural disasters, climate changes and anomalies. In addition, it allows to observe the levels of chlorophyll in the water, which are related to the presence of schools of fish. However, in the seas of Colombia, the presence of clouds throughout the year makes it difficult to detect chlorophyll in satellite images, which makes it impossible to observe where the fish are. To solve this problem, it is proposed to carry out a study with the whole set of two that the MODIS satellite has been collecting for more than 20 years to apply image processing techniques and generate an accurate

data set, with which a machine learning model was built for the forecast of chlorophyll values lost due to high cloud cover. Also, metrics such as mean square error can be handled to evaluate the accuracy of the model. In this way, it will be possible to help the people involved in fishing and reduce the environmental impact.

1 INTRODUCCION

1.1 PLANTEAMIENTO PROBLEMA

Según cifras entregadas por la Dirección de Cadenas Pecuarias, Pesqueras y Acuícolas perteneciente al Ministerio de Agricultura, Colombia produce 41.356 toneladas de carne de pescado y camarón, esta producción representa el 0,3% del PIB del país y corresponde al 3,3% del PIB agropecuario. La industria de la acuicultura exporta 4.308 toneladas por un valor de 21,6 millones de dólares y se estima que en el año 2020 logró generar 53.805 empleos directos y 161.416 empleos indirectos [1] . La pesca y la acuicultura en Colombia se producen a lo largo de las costas Pacífico y Atlántica, también en aguas interiores correspondientes a los ríos Magdalena, Amazonas, Orinoco y Sinú, lo que demuestra que esta actividad juega un papel importante para Colombia tanto en la seguridad alimentaria como para las economías locales de las regiones donde se ejercen estas actividades. Sin embargo, la sobreexplotación del recurso ha condicionado que la conservación de los ecosistemas se vea beneficiada [2] , además,

sumado a la poca cantidad de información recolectada alrededor de las actividades de pesca artesanal e industrial por parte de Colombia, la implementación de planes para realizar una actividad de pesca sostenible se ha visto ralentizada, por tales razones, el Gobierno Nacional ha puesto su mirada en la recopilación y creación de datos que ayuden a la implementación de una pesca sostenible con el fin de diseñar políticas que protejan los ecosistemas y las poblaciones de peces beneficiando el sector social y económico de las regiones [3] .

La percepción remota satelital constituye un conjunto de técnicas que permiten la captura de información sobre extensas áreas (continentales y marítimas) con una disponibilidad de los datos espacial, radiométrico y temporal que facilitan estudios de distintos fenómenos o actividades que se desarrollen sobre alguna región en específico [4] . Diferentes proyectos relacionados con actividades de pesca han sido llevados a cabo demostrando como la percepción remota satelital favorece a la captura y generación de nuevos datos [5]–[8] . La clorofila medida sobre zonas marítimas es un factor altamente relacionado con la presencia de bancos de peces [9], [10] . Los satélites ópticos MODIS ponen a disposición series temporales relacionadas con la concentración de clorofila en zonas de mar abierto [11] , sin embargo, para que estos valores sean válidos, la ausencia de cobertura nubosa es un requisito. Colombia es un país que se caracteriza por su alta concentración de cobertura nubosa durante la mayor parte del año dificultando conformar series temporales lo que afecta la resolución temporal de las medidas y por ende constituyéndose en uno de los principales retos a superar en la aplicación de técnicas de percepción remota satelital [12]–[14] .

Actualmente, la Dirección General Marítima de Colombia a través de sus Centros de Investigaciones Oceanográficos e Hidrográficos del Caribe y Pacífico desarrollan una metodología para la detección de bancos de peces utilizando información asociada con la concentración de clorofila en zonas de mar abierto, sin embargo, la alta presencia de coberturas nubosas sobre las zonas de interés ha dificultado aplicar la metodología en áreas extensas, por tales motivos en este proyecto se plantea la siguiente pregunta de investigación:

¿Cómo pronosticar valores de concentración de clorofila afectados por la presencia de cobertura nubosa en zonas de mar abierto teniendo como principal insumo las series temporales del satélite MODIS?

1.2 JUSTIFICACIÓN

Actualmente el Gobierno Nacional a través de su Ministerio de Agricultura buscan generar políticas encaminadas a la generación de pesca sostenible, el objetivo es impactar de forma positiva a las comunidades pesqueras tanto social como económicamente debido a que el sector pesquero y la acuicultura representan un valor considerable en relación con el PIB del país y su seguridad alimentaria. Por otra parte, se requiere que el impacto de la sobreexplotación se reduzca y que los bancos de peces y sus ecosistemas no se vean afectados por las nuevas técnicas de pesca que se llevan a cabo en regiones como las costas Pacífico y Atlántico. La carencia de datos disponibles para generar información relacionado con el sector pesquero ha conllevado que la implementación de políticas en pro de una pesca sostenible se vea ralentizada, la herramienta operativa que en este proyecto se propone busca convertirse en un insumo que genere valores de concentración de clorofila en zonas de mar abierto tratando de mejorar la

resolución temporal y cantidad de pixeles disponibles para análisis en una metodología de pesca trabajada por la Dirección General Marítima a través de sus Centros de Investigaciones utilizando modelos de aprendizaje máquina y/o estadísticos.

La herramienta operativa tendrá como finalidad la actualización y mejoramiento continuo de las series temporales MODIS relacionados con los productos de concentración de clorofila y que se constituyen en pieza clave para la detección de bancos de peces en zonas de mar abierto. **Se denomina herramienta operativa porque realiza tareas o actividades especificas dentro de un proceso o sistema operativo y está diseñada para facilitar y mejorar la eficiencia de tareas concurrentes, en este caso recibir como entrada imágenes con valores perdidos de clorofila y dar como resultado una imagen pronosticada con dichos valores.**

El proyecto generará información entorno al sector pesquero brindando imágenes satelitales con una cantidad menor de cobertura nubosa facilitando mejorar procesos que requieran del insumo de imágenes MODIS para funcionar

1.3 OBJETIVOS

1.3.1 Objetivo General

Construir una herramienta operativa para el pronóstico de valores de clorofila en zonas de mar abierto a través de la aplicación de técnicas de aprendizaje máquina y/o modelos estadísticos utilizando productos derivados de series temporales del satélite MODIS.

1.3.2 Objetivo Específicos

· Construir una herramienta operativa para la conformación y actualización continua de una serie temporal del satélite MODIS relacionado con productos de clorofila en zonas marítimas utilizando algoritmos para la gestión y automatización de descargas desde servidores.

· Implementar un modelo de aprendizaje máquina y/o estadístico para pronosticar valores de clorofila perdidos a causa de las coberturas nubosas sobre las zonas de interés del proyecto a partir del conjunto de datos relacionado con las series temporales MODIS.

· Evaluar los modelos implementados sobre la herramienta operativa utilizando métricas de validación como el error cuadrático medio en los pronósticos de valores de clorofila sobre las zonas marítimas de interés del proyecto.

El resto del documento se encuentra distribuido de la siguiente manera partiendo del siguiente

Capítulo 1: Introducción: En este capítulo se presenta el planteamiento del problema, la justificación, la pregunta de investigación que originó el proyecto, el objetivo general y objetivos específicos definidos inicialmente en el anteproyecto y finalmente la organización del documento.

Capítulo 2: Marco de referencia: En este capítulo se presenta los trabajos más significativos e innovadores que se han hecho alrededor del pronóstico de valores de clorofila enfocado en interfaces graficas e inteligencia artificial

Capítulo 3: Metodología: En este capítulo se presenta el desarrollo desde la etapa inicial, se desarrolló una metodología para la descarga de las imágenes MODIS, y dos softwares que se

construyeron para realizar el corregistro de todas las imágenes satelitales y el otro para que nos de todos los metadatos que contiene dicha imagen.

Capitulo 4: En este capítulo se explica detalladamente toda la implementación del software, dando a entender para que funcionen los 6 módulos de esta herramienta operativa y entiendan cómo funciona el pronóstico de imágenes satelitales.

Capítulo 5: En este capítulo se presenta los resultados más importantes del proyecto, tanto las métricas de evaluación del sistema como los productos derivados del mismo.

Capítulo 6: Conclusiones y trabajos futuros: En este capítulo se presentan las conclusiones obtenidas al finalizar el proyecto e ideas que se esperan realizar en un trabajo a futuro.

Capítulo 7: Referencias: En este capítulo, se tiene las referencias bibliográficas de los artículos, libros, entre otros, consultados para la realización de la investigación

2 MARCO TEÓRICO Y ESTADO DEL ARTE

2.1 Variable estudio-Clorofila

Es el pigmento de color verde que se encuentra en las plantas, ya que estas absorben la energía lumínica y la transforman en energía química, a este proceso se le denota como fotosíntesis. Hay bastantes tipos de clorofila, una de ellas es la clorofila A; son todos los organismos capaces de realizar el proceso de fotosíntesis y entre ellos están las plantas y las algas. Gracias a la fotosíntesis que la clorofila tiene, las plantas mediante el trabajo del agua y dióxido de carbono pueden transformarlo en oxígeno y carbohidratos para los peces.

La clorofila es presencia de algas en mar, cuando denotamos cantidades importantes de clorofila en el mar abierto, podemos concluir que hay gran cantidad de peces que están pastando sobre esa sección de fitoplancton, esto es muy importante ya que son los productores primarios del ecosistema marítimo. Podemos clasificar la concentración de clorofila A, como alta de color rojo, media de color Amarillo o sapote, Baja de color verde y muy baja de color azul. [17]

2.2 Descripción de las Imágenes satelitales

Las imágenes satelitales son fotografías de la Tierra tomadas desde satélites en órbita alrededor del planeta. Estas imágenes son utilizadas en diferentes campos, como la meteorología, la agricultura, la cartografía, la geología, la gestión de recursos naturales, la seguridad y la defensa, entre otros. La tecnología de imágenes satelitales se ha desarrollado rápidamente en los últimos años, lo que ha permitido una mayor precisión y resolución de las imágenes. Las imágenes satelitales se utilizan para realizar análisis y estudios de diferentes características de la Tierra, como la vegetación, el clima, la topografía, la humedad del suelo, la calidad del agua y el uso del suelo. Existen diferentes tipos de sensores utilizados en la toma de imágenes satelitales, como los sensores ópticos, los infrarrojos y los de radar. Cada tipo de sensor tiene sus propias características y se aplican para diferentes aplicaciones.[18]

Las imágenes satelitales también son usadas en la monitorización de desastres naturales, como terremotos, tsunamis, huracanes y erupciones volcánicas, lo que permite una mejor planificación de la respuesta de emergencia y la evaluación de daños. Las imágenes satelitales

son una herramienta valiosa para la investigación científica y el monitoreo ambiental a largo plazo. Con la creciente disponibilidad de datos satelitales, se han desarrollado diversas aplicaciones y herramientas para el análisis de imágenes satelitales, como la teledetección y el aprendizaje automático. [19]

2.2.1 Definición del sensor MODIS

El sensor MODIS (Moderate Resolution Imaging Spectroradiometer) fue lanzado en órbita por la nasa en 1999 incursionado el satélite Terra y en el 2002 incursionaron el satélite Aqua; es un satélite que observa y captura datos de uno a dos días en las zonas de agua y tierra de todo el planeta, debido a su alta cobertura y a su resolución espacial da una mejor visibilidad del campo, además su usabilidad más importante es su grande franja que cubre un ancho de 2.330 km a una sensibilidad radiométrica de 12 bits , posee sensores para detectar un canal de nubes de tipo CIRRUS, que están relacionadas con el calentamiento global, esto nos ayuda a monitorear los cambios climáticos, actividad volcánica e incendios forestales y juega un papel muy importante en el desarrollo de modelos para la predicción de cambios climáticos. También nos ayuda a detectar los fenómenos que suceden en el mar abierto como la alta concentración de clorofila para la detección de bancos de peces.[20]

2.3 Proceso de transformación Corregistro de Imágenes

El corregistro de las imágenes es un proceso que realiza transformaciones a un grupo de imágenes seleccionadas, para que todas queden en la misma posición y con el mismo tamaño. Este proceso es muy complejo, porque requiere tiempo para la descarga y estudio de las imágenes, para así poder aplicar este proceso, esto ayuda a tener un orden con todas las imágenes y centralizarse en el punto de interés que se desea. Para realizar el proceso corregistro utilizamos un software gratuito llamado SeaDas, el cual sirve para realizar el análisis de datos satelitales como imágenes y tiene una gran función analítica para los ambientes marinos y oceánicos. Trabaja con niveles de clorofila, temperaturas superficiales que son mapeados por los satélites MODIS, Sentinel 3 o Suomi NPP. Existen plataformas como el Ocean Color, Copernicus y Bio-Oracle que nos permiten descargar las imágenes tomadas por los satélites.[21]

2.4 Técnicas de aprendizaje automático Algoritmos de predicción

Los algoritmos de predicción son un conjunto de técnicas utilizadas en el aprendizaje automático para hacer estimaciones sobre valores desconocidos de una variable de interés, gran cantidad de datos o patrones identificados en los datos. Estos algoritmos se utilizan ampliamente en áreas como finanzas, marketing, salud, fabricación y en general, en cualquier campo donde la predicción de valores futuros sea importante. Existen diferentes tipos de algoritmos de predicción, tales como regresión lineal, regresión logística, árboles de decisión, k-vecinos más cercanos, redes neuronales y máquinas de vectores de soporte (SVM, por sus

siglas en inglés). Cada algoritmo tiene sus propias ventajas y desventajas, y su elección de la naturaleza de los datos y del problema de predicción en sí.

Los algoritmos de predicción se utilizan para una variedad de aplicaciones, como el pronóstico del clima, la detección de fraudes, la predicción de ventas, la clasificación de pacientes en riesgo y la predicción de tendencias en los mercados financieros. La eficacia de un algoritmo de predicción depende de la calidad de los datos de entrada, el preprocesamiento adecuado de los datos y la elección del algoritmo adecuado. Es importante evaluar la precisión del modelo de predicción utilizando conjuntos de datos de prueba y de validación para garantizar que las predicciones sean precisas y útiles.[22]

2.4.1 Algoritmos de predicción Redes neuronales artificiales

Las redes neuronales artificiales (ANN, por sus siglas en inglés) son un conjunto de algoritmos de aprendizaje automático que utilizan una arquitectura de nodos interconectados para procesar y transformar datos de entrada en una salida deseada. Las redes neuronales son altamente adaptables y pueden ser entrenadas para predecir valores numéricos o para clasificar datos. Los algoritmos de predicción basados en redes neuronales suelen utilizar una técnica de aprendizaje supervisado, en la cual se proporciona a la red una serie de datos de entrenamiento con valores de entrada y salida conocidos. A través de múltiples iteraciones de ajuste de pesos, la red aprende a relacionar las entradas con las salidas y puede generalizar esta relación para hacer predicciones precisas sobre datos nuevos.

Existen varios tipos de redes neuronales que se pueden utilizar para la predicción, como las redes neuronales de retro propagación, las redes neuronales convolucionales (CNNs) y las redes neuronales recurrentes (RNNs). Las redes neuronales artificiales son muy utilizadas en aplicaciones de predicción en campos como la visión por computadora, el procesamiento del lenguaje natural, el reconocimiento de voz y el análisis de series de tiempo. Además, las redes neuronales han demostrado ser efectivas para resolver problemas de predicción en áreas como finanzas, marketing y salud. [23]

2.4.2 Modelo de regresión lineal BAYESIANRIDGE

Bayesian Ridge es un modelo de regresión lineal que utiliza la teoría de Bayes para realizar una estimación de los parámetros del modelo y su incertidumbre. Es una técnica de aprendizaje automático que se utiliza para predecir valores numéricos a partir de un conjunto de características. La regresión lineal es un método que se utiliza para predecir valores numéricos basados en una o más variables independientes. El modelo de regresión lineal utiliza una línea recta para ajustar los datos, lo que permite realizar predicciones sobre valores desconocidos.

Bayesian Ridge es una mejora del modelo de regresión lineal tradicional, ya que permite incorporar información previa sobre los parámetros del modelo y su incertidumbre. Esto permite una mayor precisión en la estimación de los parámetros y una mayor confianza en las predicciones del modelo. La implementación de Bayesian Ridge en Python se puede realizar

a través de la biblioteca scikit-learn. Esta biblioteca proporciona una implementación eficiente de este algoritmo de regresión, así como otras técnicas de aprendizaje automático. [24].

En conclusión, se utilizó el modelo de regresión lineal por que la variable objetivo de clorofila-a es continua y se utilizó Bayesian Ridge porque fue considerado el modelo más optimo y más preciso haciendo una comparación con métricas de regresión con los demás modelos, el cual se explicará más a fondo en el capítulo 4.

2.5 Conjunto de datos (Dataset)

Un dataset o conjunto de datos es una colección de datos organizados y estructurados de manera que se pueden utilizar para analizar, evaluar o construir modelos de aprendizaje automático. Los datos en un conjunto de datos suelen estar relacionados con un tema específico, como el rendimiento de una empresa, las opiniones de los clientes, o los resultados de una investigación. Los conjuntos de datos pueden estar disponibles en diferentes formatos, como archivos CSV, JSON, Excel, entre otros. También pueden estar disponibles en bases de datos o ser generados a través de la recopilación de datos a través de encuestas, experimentos o registros históricos.

Los conjuntos de datos pueden ser públicos o privados y pueden ser utilizados por empresas, investigadores y científicos de datos para llevar a cabo diferentes tareas, como la minería de datos, el análisis de tendencias, la predicción de resultados futuros, entre otras. La calidad y la cantidad de los datos en un dataset son importantes para garantizar que los modelos de aprendizaje automático perturbados sean precisos y relevantes. Por lo tanto, es importante que

los conjuntos de datos sean cuidadosamente seleccionados y preprocesados antes de ser utilizados para el entrenamiento de modelos. [25]

2.6 Algoritmo de aprendizaje automático (Redes Neuronales)

Las redes neuronales son un conjunto de algoritmos de aprendizaje automático inspirados en la estructura y funcionamiento del cerebro humano. Consisten en una colección de nodos interconectados, conocidos como neuronas, que se organizan en capas para procesar y transformar la información de entrada en una salida. Las redes neuronales se entrenan mediante el ajuste de los pesos de las conexiones entre las neuronas para minimizar una función de costo o error. El proceso de entrenamiento implica presentar a la red una serie de datos de entrenamiento y comparar las salidas predichas con las salidas reales, lo que permite ajustar los pesos de las conexiones para mejorar la precisión de las predicciones.

Entre las ventajas de las redes neuronales se encuentran su capacidad para aprender patrones complejos y no lineales en los datos, su capacidad para generalizar a partir de datos de entrenamiento a datos nuevos, y su capacidad para manejar datos de alta dimensionalidad. Sin embargo, también tienen algunas limitaciones, como la necesidad de grandes cantidades de datos de entrenamiento y la dificultad para interpretar y explicar los resultados de la red.

Las redes neuronales se utilizan en una variedad de aplicaciones, como el procesamiento de imágenes, el procesamiento del lenguaje natural, la visión artificial, la robótica, el reconocimiento de voz y la clasificación de datos. [26]

2.7 Métrica de valor cuadrático medio (coeficiente de correlación)

La métrica de valor cuadrático medio (MSE, por sus siglas en inglés) es una medida utilizada en estadística para evaluar la calidad de un modelo de predicción o regresión. Se calcula como la media de los errores cuadráticos de las predicciones del modelo, es decir, la diferencia al cuadrado entre los valores predichos y los valores observados.

La fórmula para el cálculo del MSE es:

MSE = (1/n) * suma ((y_pred - y_obs) ^2) donde y_pred son los valores predichos por el modelo, y_obs son los valores observados, yn es el número de observaciones. El MSE tiene la ventaja de penalizar duro los errores grandes, ya que los errores se elevan al cuadrado en el cálculo, lo que lo hace especialmente útil para modelos en los que se desea minimizar los errores grandes. Sin embargo, su uso puede ser problemático cuando los errores no están distribuidos normalmente o cuando se tienen valores extremos. La métrica MSE se utiliza con frecuencia en problemas de regresión en el campo de la inteligencia artificial y el aprendizaje automático, y es una de las medidas que se utilizan para evaluar la precisión de los modelos de predicción. [27]

2.8 Librería para manejo imágenes (NetCdf4)

NetCDF (Network Common Data Form) es un formato de archivo de datos científicos utilizados para almacenar y distribuir conjuntos de datos multidimensionales, como datos atmosféricos, oceanográficos y de la tierra. NetCDF4 es la última versión de NetCDF y ofrece

mejoras en el rendimiento, la escalabilidad y la portabilidad. NetCDF4 es una biblioteca de software que permite la creación, lectura y escritura de archivos. La biblioteca está disponible en varios lenguajes de programación, como Python, C++, Fortran, entre otros, lo que la hace una herramienta flexible para la comunidad científica.

Entre las mejoras de NetCDF4, destacan la capacidad de comprimir archivos para reducir su tamaño y acelerar su transferencia, la posibilidad de usar formatos de compresión sin pérdida de datos, la compatibilidad con el formato HDF5, que permite almacenar conjuntos de datos grandes y complejos en un solo archivo, y la adición de metadatos para la descripción detallada de los datos almacenados. NetCDF4 es ampliamente utilizado en la comunidad científica para almacenar y compartir datos de investigación, especialmente en las áreas de ciencias atmosféricas y oceánicas, pero también en áreas como la climatología, la hidrología, la biología y la geología. [28]

2.9 Corregistro de imágenes satelitales

El corregistro de imágenes satelitales es un proceso de alineación de dos o más imágenes de diferentes fuentes, obtenidas en diferentes momentos, posiblemente por diferentes sensores, para que los objetos en la imagen tengan las mismas coordenadas espaciales. Es un proceso importante en la teledetección y en la obtención de imágenes de alta calidad. El corregistro se realiza para mejorar la calidad de las imágenes y permitir comparaciones entre ellas. Se utiliza en aplicaciones como la detección de cambios en el uso del suelo, la evaluación de la calidad

del aire y el agua, la detección de características geológicas y la monitorización de la actividad volcánica.

Existen diferentes técnicas de corregistro, como la conexión cruzada, la conexión de fase, la transformación de Fourier y la comparación de características. En general, estas técnicas se basan en la búsqueda de puntos de correspondencia entre las imágenes y la estimación de una transformación geométrica que puede llevar a cabo una imagen a la otra. El corregistro de imágenes satelitales es un proceso importante en la teledetección y en la obtención de imágenes de alta calidad. La calidad del corregistro puede afectar significativamente la precisión de las mediciones y las evaluaciones realizadas a partir de las imágenes.[29]

2.10 Técnica de aprendizaje automático Deep Learning

El Deep Learning es una rama del aprendizaje automático que utiliza redes neuronales profundas para resolver problemas de aprendizaje automático. Es un subcampo de la inteligencia artificial que utiliza modelos de redes neuronales artificiales para aprender de grandes cantidades de datos y realizar tareas complejas, como el reconocimiento de imágenes, el procesamiento del lenguaje natural y la toma de decisiones. Las redes neuronales profundas están compuestas de múltiples capas de neuronas artificiales que procesan la información de entrada y generan una salida. Estas redes son capaces de aprender de los datos de entrada y mejorar su capacidad para hacer predicciones precisas a medida que se les proporciona más información.

El Deep Learning ha tenido un gran impacto en áreas como el procesamiento de imágenes, el procesamiento del lenguaje natural y la robótica. También se utiliza en aplicaciones de inteligencia artificial, como reconocimiento de voz, reconocimiento de patrones y análisis de datos. El aprendizaje profundo requiere grandes conjuntos de datos de entrenamiento y un poder de procesamiento significativo para entrenar modelos de redes neuronales profundas. Sin embargo, las bibliotecas de código abierto como TensorFlow, Keras y PyTorch han hecho que sea más fácil para los desarrolladores implementar y entrenar modelos de aprendizaje profundo. [30]

2.11 Entrenamiento de imágenes satelitales

El entrenamiento de imágenes satelitales es un proceso de aprendizaje automático que utiliza datos de imágenes satelitales para entrenar modelos de aprendizaje automático y realizar predicciones sobre nuevas imágenes. Los modelos de aprendizaje automático se entrenan para identificar patrones y características en los datos de las imágenes y utilizarlos para hacer predicciones. El entrenamiento de imágenes satelitales es una aplicación común en campos como la teledetección, la agricultura de precisión y la monitorización ambiental. Se utiliza para clasificar el uso del suelo, identificar características geográficas, detectar cambios en el medio ambiente y mucho más.

El proceso de entrenamiento de imágenes satelitales generalmente implica la recolección de datos de imágenes satelitales, la anotación de los datos para indicar las características o patrones a identificar y el entrenamiento del modelo de aprendizaje automático. El proceso

de anotación de los datos puede ser manual o automático, y se pueden utilizar diferentes algoritmos de aprendizaje automático, como SVM, redes neuronales, árboles de decisión, entre otros. La precisión del modelo de trastorno depende en gran medida de la calidad y cantidad de los datos de imágenes utilizados para el entrenamiento. También es importante seleccionar el algoritmo de aprendizaje automático adecuado y ajustar los parámetros del modelo para obtener los mejores resultados. [31]

2.12 Librería para el procesamiento de imágenes georreferenciadas (GDAL)

GDAL (Geospatial Data Abstraction Library) es una biblioteca de software libre y de código abierto para el procesamiento de datos geoespaciales. Fue desarrollado por Frank Warmerdam en 1998 y actualmente se mantiene por una comunidad de desarrolladores. GDAL proporciona una amplia gama de funciones para leer, escribir y procesar datos geoespaciales en diferentes formatos, incluidos raster y vector. También puede ser utilizado en múltiples plataformas, incluyendo Windows, Linux y Mac OS, y puede ser utilizado en varios lenguajes de programación, como C++, Python y Java.

La biblioteca GDAL se utiliza ampliamente en aplicaciones de GIS (Sistemas de Información Geográfica) para el procesamiento de datos geoespaciales y la creación de mapas. Es compatible con una gran cantidad de formatos de datos geoespaciales, lo que permite a los usuarios trabajar con diferentes tipos de datos y formatos sin tener que preocuparse por la compatibilidad. GDAL también es compatible con otras bibliotecas de software de

geoprocesamiento, como OGR y PROJ, lo que permite a los usuarios crear flujos de trabajo completos para el análisis de datos geoespaciales.[32]

2.13 Librería par visión por computadora (OpenCV)

(Open Source Computer Vision) Es una biblioteca de software libre y de código abierto que se utiliza para desarrollar aplicaciones de visión por computadora y procesamiento de imágenes en tiempo real. Fue desarrollado por Intel en 1999 y actualmente es mantenido por Willow Garage y Itseez.

La biblioteca OpenCV proporciona más de 2500 algoritmos optimizados para tareas de procesamiento de imágenes y visión por computadora, incluyendo detección de rostros, seguimiento de objetos, detección de movimiento, calibración de cámara y muchos más. También admite múltiples plataformas, incluyendo Windows, Linux, Mac OS, iOS y Android, y puede ser utilizada en varios lenguajes de programación, como C++, Python y Java.

OpenCV se ha utilizado en una amplia gama de aplicaciones, incluyendo sistemas de vigilancia, robótica, visión artificial, reconocimiento de caracteres, detección de objetos en tiempo real y más. [33]

2.14 Librería de aprendizaje automático Scikit Learn

Scikit-learn es una biblioteca de aprendizaje automático de código abierto para el lenguaje de programación Python. Fue lanzado en 2007 y es uno de los paquetes más populares para el aprendizaje automático en Python.

Scikit-learn proporciona una amplia gama de algoritmos de aprendizaje supervisado y no supervisado, incluyendo regresión lineal y logística, árboles de decisión, SVM, clustering, análisis de componentes principales (PCA) y más. También proporciona herramientas para la evaluación de modelos, selección de características, preprocesamiento de datos y más. La biblioteca es fácil de usar y está diseñada para ser eficiente en términos de uso de memoria y cálculo. Scikit-learn es muy popular en la comunidad de aprendizaje automático debido a su facilidad de uso y su capacidad para procesar grandes cantidades de datos de manera eficiente. Scikit-learn es compatible con otras bibliotecas de Python para el análisis de datos, como NumPy y Pandas, lo que permite a los usuarios crear flujos de trabajo completos para el análisis de datos y el aprendizaje automático. [34]

2.15 ESTADO DEL ARTE

Diferentes investigaciones relacionadas con pronósticos de valores de clorofila han sido trabajadas en los años anteriores, utilizando métodos de aprendizaje automático como Bagging, Boosting y Random Forest. A continuación, se presenta el estado del arte relacionado con la creación de modelos para el pronóstico de valores de clorofila.

En 2002 C. S. Rousseaux and W. W. Gregg [35] desarrollaron un algoritmo de recuperación de clorofila (Chl) para su uso con imágenes satelitales adquiridas por el instrumento MERIS (Espectrómetro de imágenes de resolución media). El algoritmo fue adaptado a partir de uno originalmente desarrollado para la reflectancia de la irradiación espectral del subsuelo determinada a partir de mediciones a bordo de barcos sobre el agua. Se calibró para concentraciones de Chl a en el rango de 3-185 mg m −3 utilizando la reflectancia espectral calculada a partir de mediciones a bordo en la laguna IJssel (Países Bajos). Luego, se validó el algoritmo para varias aguas interiores y costeras que cubren este rango de concentración. A pesar de la resolución espectral más baja de MERIS en comparación con el espectro radiómetro de a bordo, se espera que el error estándar de estimación sea similar, es decir, ~9 mg m −3de Chl a en lagos, ríos, estuarios y aguas costeras mesotróficas y eutróficas.

En el 2013 M. Jouini, M. Crépon and S. Thiria [36] presenta un método para llenar vacíos de datos en imágenes satelitales de clorofila (CHL) mediante el uso de observaciones oceánicas complementarias de teledetección: temperatura de la superficie del mar (SST) y altura de la superficie del mar (SSH). El método se basa en el principio de métodos de clasificación de mapas auto organizados (SOM) y se basa en la suposición de que un estado del océano puede definirse localmente por sus valores de SST, SSH y CHL. La viabilidad y solidez del método se evaluaron en un conjunto de datos sintéticos y luego se adaptó a datos satelitales reales, introduciendo aprendizaje iterativo y abordando la determinación multimodal de los vectores de referencia. El método demostró ser efectivo en la recuperación de valores perdidos de CHL

bajo una intensa cobertura de nubes y en la mejora de la calidad de los productos semanales de CHL.

En el 2017 Rousseaux & Gregg [16] usando un modelo bioquímico oceánico evaluaron la habilidad de un pronóstico de concentraciones de clorofila en el pacífico ecuatorial en respuesta al evento del Niño. para el periodo 2012-2015 usando un enfoque de pronósticos retrospectivos. Se evaluaron las incertidumbres de clorofila pronosticada al comparar con la concentración mensual de clorofila tomadas con Suomi-National (S-NPP VIIRS) satélite. Este pronóstico logró reducir el escalonamiento de la variabilidad en la concentración de clorofila del pacifico ecuatorial. De los 38 pronósticos realizados la diferencia porcentual promedio entre clorofila pronosticada y clorofila (S-NPP VIIRS) varió entre el 23% (plazo 3 meses) y el 30% (plazo de 9 meses). Mostrando que la concentración de clorofila pronosticada estuvo correlacionada con la clorofila de (S-NPP VIIRS).

Para este estudio hizo falta probar el pronóstico para varios conjuntos, este uso puede mejorar los resultados a diferencia de si se usa un solo conjunto como se trabajó en este modelo.

En 2018 Yajima & Derot [37] aplicaron un modelo de random forest para pronósticos de clorofila -a en cuerpos de agua dulce, el lugar del estudio fue el lago de Urayama el cual tiene una cuenca hidrográfica de 51,6 metros cuadrados. Para este análisis se utilizaron 33 parámetros de entrada del modelo para el embalse, resaltando que en 2011 se instaló un estratificado artificial de pluma de burbujas en el fondo del embalse, por este motivo se tomaron datos solo

de 1999 hasta el 2010 para así evitar sesgos en el aprendizaje del modelo. Se implementó el modelo de ventana deslizante usando como parámetros (lead-time y lag-time), se probó el parámetro de uno a tres meses. Como resultado obtuvieron que usar un número reducido de árboles de decisión permitió reducir el tiempo de cálculo sin afectar los resultados finales, manteniendo el error cuadrático medio estable para más de 200 árboles de decisión, se hicieron pronósticos por estación usando un número óptimo de árboles y un número mínimo del predictor(min-leaf) logrando resultados esperados.

Para este estudio hizo falta usar diferentes modelos para la predicción, como puede ser deep learning con redes neuronales las cuales pueden ser más precisas la hora de hacer pronósticos.

En 2018 F. Alfonita [38] utilizó imágenes de satélite Landsat-8 (SR) como herramienta para monitorear procesos funcionales en ecosistemas acuáticos, complementando las técnicas in-situ. El objetivo fue determinar la variación espacial, temporal de la distribución de concentración de clorofila y de la condición trófica de dos ecosistemas acuáticos con condiciones ambientales contrastantes, la Ciénaga de Ayapel (Córdoba) y el Lago de Tota (Boyacá). Se generaron modelos de regresión a partir de la reflectancia de las diferentes bandas del espectro electromagnético, obtenida de las imágenes y de datos tomados en campo. Los resultados indicaron que, para la Ciénaga de Ayapel se presentó buena correlación entre variables, mientras que para el Lago de Tota no se presentó una buena correlación debido a la profundidad de los ecosistemas acuáticos y la presencia de sólidos disueltos y suspendidos. En

general, se considera que esta herramienta es útil a pesar de las limitaciones para cuerpos de agua profundos debido al bajo costo, nivel de resolución y alcance en plan de monitoreo.

En el 2019 H. Cho and H. Park[39] proponen se usó un modelo avanzado basado en redes LSTM llamado Merged-LSTM. Este modelo contiene tres capas LSTM paralelas y capas combinadas que permiten utilizar datos adicionales de diversas fuentes sin problemas durante el proceso de entrenamiento. El modelo se utiliza para predecir la clorofila-a del Río Geumessituado en el medio occidental parte de Corea y se comparó con otros modelos existentes. Los resultados mostraron que el modelo propuesto superó a los modelos anteriores y también mostró un mejor proceso de entrenamiento con mayores dimensiones de datos. El texto sugiere que esto abre la posibilidad de predecir floraciones de algas con modelos más avanzados y fuentes de datos correspondientes.

En el 2020 Shin et al., [40] realizaron un modelo de aprendizaje usando el método de ventana móvil, en este estudio se usaron datos diarios de cantidad/calidad de agua y datos meteorológicos desde 2015 a 2017. Para este estudio se construyó un modelo de predicción para la gestión de la calidad del agua contra la proliferación de algas. Usando varios modelos de aprendizaje automático como Support vector regression (SVR). SVR se usa en una amplia gama de campos como aprendizaje automático, estadísticas y análisis funcional. También se utilizaron redes neuronales recurrentes (RNN), memoria a largo y corto plazo (LSTM), la red neuronal se utilizó debido que la calidad del agua se juzga con variables de naturaleza no lineal,

RNN y LSTM se pueden utilizar para predecir futuras concentraciones de clorofila en función de valores actuales e información pasada.El flujo de trabajo se dividió en 4 pasos: recopilación de datos, procesamiento y estandarización de datos, ajustar el modelo utilizando las variables seleccionadas a través de SVR, RNN y LSTM, Predecir las concentraciones de clorofila. Como resultado se obtuvo la conclusión de que las RNN superaron a los demás modelos de aprendizaje mostrando predicciones aceptables en la mayoría de los casos.

Para esta investigación solo se tuvieron en cuenta ríos y concentraciones en agua dulce, no se tomaron en cuenta variables como nubes que puedan afectar la variable clorofila.

En 2021 Ham et al., [41] hacen un estudio sobre el mecanismo físico involucrado en un sistema de predicción biogeoquímica marina basado en el modelo del sistema terrestre (ESM) que proporciona pronósticos exitosos de concentración de clorofila en el pacífico sur. Estas predicciones retrospectivas se inician el primer día de cada mes calendario desde 1991 hasta 2017, con conjunto de datos de 12 miembros de dos años de duración utilizando predicciones medias de conjunto, la habilidad de predicción se midió mediante el coeficiente de correlación de anomalías temporales entre las anomalías de concentraciones de clorofila. Como resultado se obtuvo que los casos de concentración de valores de clorofila superficial más altos ocurren durante el fenómeno de la niña.

En este estudio hizo falta probar modelos de aprendizaje automático como redes neuronales recurrentes, random forest, entre otros, para así automatizar ciertos procesos y tener mayor precisión en los modelos.

El 2021 Hänninen et al., [42] realizan el "índice de seily" para la predicción de clorofila en el archipiélago del norte, mar báltico al suroeste de Finlandia, para este estudio se generó un modelo mixto aditivo a los datos que fueron recolectados durante los años 2011-2019.

En este estudio se enfocaron dos variables, la clorofila y la temperatura del agua, se dividió la profundidad del agua en dos clases (profundidad < 20 metros) y (profundidad > 20), se trabajó con los valores diarios de clorofila. El estudio se llevó a cabo en forma de ventana móvil, teniendo como datos de entrada para el entrenamiento los datos recopilados desde 2011-2018 además de los primeros 25 días del 2019.Esto se usó para así predecir los siguientes 10 días del año 2019.Como resultado de este estudio se obtuvo que los valores predichos en una ventana están cerca de los verdaderos valores observados al comienzo, que se utilizan como entrenamiento, esto nos indica que el modelo funciona bastante bien.

Para este estudio falto probar el modelo en una ventana móvil más grande, así se sabría con más exactitud la calidad del modelo debido que naturalmente si se extiende la ventana los resultados serían más propensos a tener errores.

En 2021 Eze et al., [43]realizan un estudio sobre el análisis de concentración de clorofila-a de serie temporal, este es un modelo de pronóstico de alta precisión del contenido de clorofila para la industria de la agricultura, lo cual permita tomar medidas de remediación contra apariciones de algas toxicas en los cuerpos de agua. Este modelo combinó descomposición de

modo empírico (EEDM), redes neurales (NN) d memoria de corto plazo (LSTM) y aprendizaje profundo (DL).

Los datos de las series temporales se descomponen con (EEDM) en funciones del modo intrínseco (IMF), se empleó una selección de atributos con el fin de seleccionar el IMF que tengan correlación con la concentración de clorofila para integrarlos como entradas el Deep learning. Se fusionaron las técnicas (EEDM) y aprendizaje profundo para crear un modelo de predicción hibrido de clorofila. Como resultado se tuvo que el modelo propuesto con la aplicación de (EDDM) tiene buenos resultados en la predicción del contendido de clorofila, esto con relación a que no eran muchos datos en las series temporales.

Para este estudio falto utilizar una cantidad de datos más grande, si bien los resultados fueron buenos, los resultados tienden a ser mejores si se utilizan datos masivos (big data).

En 2021 Du et al., [44]realizaron un estudio utilizando redes neuronales convolucionales que son modelos de aprendizaje profundo conocidos para el procesamiento de datos de imágenes para estimar la distribución temporal y espacial de clorofila en una bahía. Para los datos de entrenamiento se construyó un modelo de aprendizaje profundo adquirido del modelo hidronímico y de Oceancolor satelital. Se desarrolló un modelo que estima la concentración de clorofila en imágenes de 48 * 27 y otro modelo que utiliza imágenes segmentadas de 7*7, se utilizaron datos desde enero de 2015 hasta diciembre de 2019.Como resultado se obtuvo que existe un ligero error predictivo en la imagen, pero la tendencia general es que está bien estimada contemplando variables como corrientes de agua, cambios estacionarios entre otros.

Para este modelo falto tomar en cuenta ciertos detalles como las nubes, en una parte del estudio se menciona que algunas imágenes estaban cubiertes por nubes y en ese caso no se pudieron tomar esos datos para estudio.

Como se puede observar se han realizado varios estudios y análisis a la clorofila en años anteriores, con resultados esperados, pero no se observó ninguno que tenga como objetivo la afectación de visualización de la clorofila en imágenes MODIS a causa de las nubes. En este estudio lo que se pretende es pronosticar el valor de clorofila en imágenes parcialmente cubiertas por nubosidad, esto con el fin de acercarnos lo más posible al valor real de la imagen MODIS. Se utilizará un data set robusto conformado por imágenes MODIS desde enero del 2002 hasta marzo del 2022, garantizando un buen conjunto de datos para el entrenamiento de los modelos de aprendizaje profundo.

En 2022 Z. L. Q. R. H. C. M. S. Y. W. Han Ding [45] afirma que la predicción de clorofila a (Chla) es útil para la gestión ambiental del agua. Sin embargo, la mayoría de los estudios previos requerían múltiples variables y un cálculo considerable. Por lo tanto, se desarrolló un algoritmo de predicción basado en similitud (SPA) para esta predicción. Los algoritmos de error relativo medio (MRE) y de distorsión de tiempo dinámico (DTW) se utilizaron para crear SPA (SPA-MRE, SPA-DTW y SPA-MRE&DTW). El modelo se validó en el embalse de China y se comparó con métodos tradicionales de análisis de series temporales. Los resultados mostraron que los métodos SPA funcionaron mejor que los métodos tradicionales y también mostraron

alta robustez y confiabilidad incluso en ausencia del 9% de los datos. El estudio sugiere que SPA es un método de predicción sólido y fiable para la gestión de la proliferación de algas.

En conclusión, se adquirió un estudio de 20 años con lo cual realizo la descarga de imágenes satelitales de la plataforma del OCEAN COLOR, después de esto hizo el corregistro a cada imagen para que todas queden de la misma posición. Se diseñó una aplicación de escritorio para que abra todas las imágenes de forma continua donde nos da el tamaño que contiene la imagen y toda la información correspondiente con la finalidad de saber si las imágenes están descargadas correctamente.

3 METODOLOGIA PARA LA PUESTA PUNTO DE IMÁGENES SATELITALES MODIS

3.1 Metodología desarrollada del proyecto

El proyecto será desarrollado siguiendo la metodología correspondiente al patrón de investigación iterativo [46] que consta de 4 fases por cada ciclo de trabajo: Observación, Identificación, Desarrollo y Pruebas.

El patrón de investigación iterativo es un proceso de investigación en el cual se realizan varias iteraciones de diseño, análisis y recolección de datos antes de llegar a una conclusión. Cada iteración se basa en los hallazgos de la iteración anterior y se enfoca en mejorar o aclarar los resultados. Este proceso es útil para investigaciones en las cuales el problema o el enfoque pueden cambiar a medida que se recolectan más datos o se obtienen nuevos hallazgos. El patrón de investigación iterativo permite a los investigadores adaptar su enfoque y metodología a medida que se desarrolla la investigación, lo que lleva a resultados más precisos y completos. Implementación de un modelo de aprendizaje máquina y/o estadístico para pronosticar valores de clorofila perdidos a causa de las coberturas nubosas sobre la zona de interés del proyecto a partir del conjunto de datos relacionado con las series temporales MODIS.

- Observación: selección de modelos de aprendizaje máquina y/o modelos estadísticos enfocados en tareas de pronósticos de valores teniendo como insumo series temporales.
- Identificación: identificación de zonas libres de cobertura nubosa sobre la región de interés del proyecto la cual es la **cuenca pacifica colombiana**, maximizando la resolución temporal de la serie de imágenes MODIS.
- Desarrollo: conformación de un conjunto de datos que cumplan con los requerimientos de los algoritmos seleccionados para construir modelos de pronóstico.

- Pruebas: ejecución del modelo entrenado sobre imágenes satelitales MODIS relacionadas con productos de clorofila, la herramienta creará nuevos productos minimizando las áreas con cobertura nubosa.

3.2 Metodología de descarga imágenes satelitales

En el marco del proyecto investigativo se desarrolló una guía metodológica para descargar paquetes de imágenes MODIS. Con el fin de compartir esta información con la dirección nacional marítima (DIMAR) para que puedan conformar sus propios conjuntas de datos de forma recurrente y de fácil entendimiento. Esta guía incluye una serie de pasos que deben seguirse para descargar las imágenes MODIS. En primer lugar, se debe registrar en la página de **oceancolor[47],** donde se solicita el producto deseado, que en este caso es la clorofila-a. A continuación, se recomiendan las coordenadas geográficas **Norte 9, Oeste -84, este-7 y Sur 1 (unidades expresadas en grados)** para seleccionar la zona en la que se desea descargar el producto. Por último, se debe seleccionar el satélite VIIRS **se seleccionar el este satélite, debido a que este satélite puede detectar diferentes capas, en este caso específico la capa de nubes**, en la **Figura 1** Selección de característica del sitio web ocean colorse puede observar la selección de características.

Esta guía podrá ser encontrada en el (anexo A)

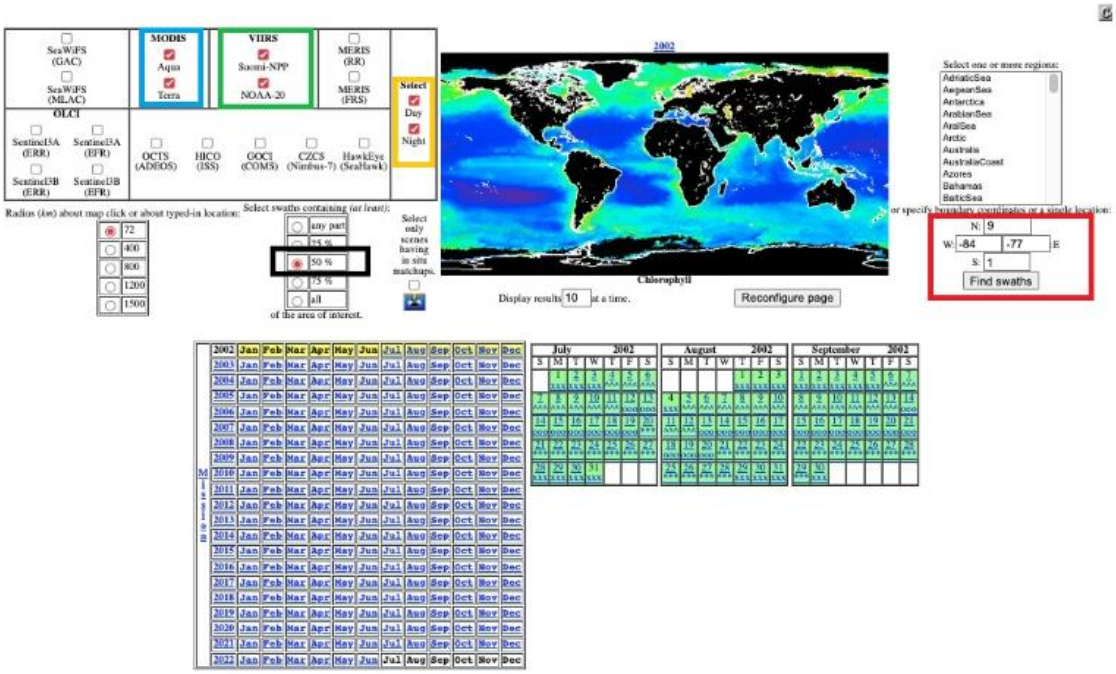
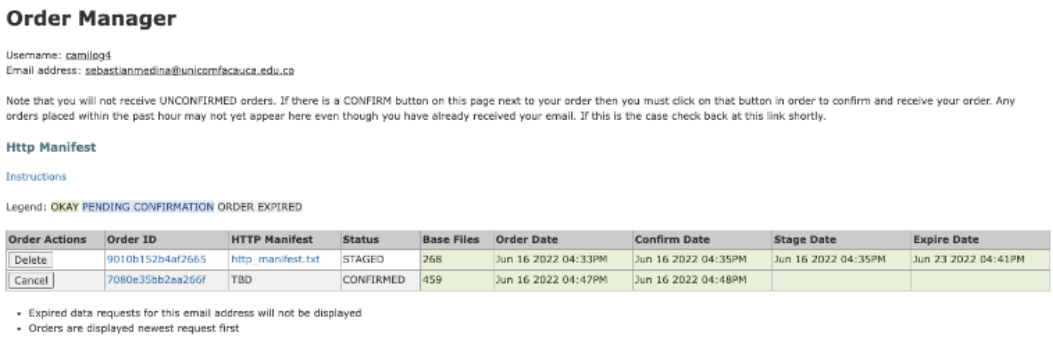


Figura 1 Selección de característica del sitio web ocean color

Para descargar imágenes MODIS. se hace una solicitud al servidor del paquete. Una vez hecha la solicitud, se recibirá un mensaje de confirmación en el correo con el que se hizo la petición. Se deberá aceptar dicha petición para que se genere un archivo manifiesto en formato txt. Este archivo contendrá un enlace que permitirá iniciar la descarga. Al pegar el enlace en cualquier navegador, se descargan paquetes de 1GB en formato rar que contienen las imágenes MODIS, en la **Figura 2** Confirmación de peticiónse observa la confirmación de la petición.



Order Manager

Username: camilog4
Email address: sebastianmedina@unicomfacauca.edu.co

Note that you will not receive UNCONFIRMED orders. If there is a CONFIRM button on this page next to your order then you must click on that button in order to confirm and receive your order. Any orders placed within the past hour may not yet appear here even though you have already received your email. If this is the case check back at this link shortly.

Http Manifest

Instructions

Legend: OKAY PENDING CONFIRMATION ORDER EXPIRED

Order Actions	Order ID	HTTP Manifest	Status	Base Files	Order Date	Confirm Date	Stage Date	Expire Date
Delete	9010b152b4af2665	http_manifest.txt	STAGED	268	Jun 16 2022 04:33PM	Jun 16 2022 04:35PM	Jun 16 2022 04:35PM	Jun 23 2022 04:41PM
Cancel	7080e35bb2aa266f	TBD	CONFIRMED	459	Jun 16 2022 04:47PM	Jun 16 2022 04:48PM		

- Expired data requests for this email address will not be displayed
- Orders are displayed newest request first

Figura 2 Confirmación de petición para la descarga de imágenes satelitales

Al descargar las imágenes y descomprimirlo se pueden abrir en algún software de visualización de imágenes georreferenciadas, en este caso se utilizó **Seadas[48],se utiliza este software debido a que es recomendado por la misma página de ocean color, página web en donde se descargan las imágenes**, en la **Figura 3** Imagen de clorofila descargadase puede observar una imagen en dicho software.

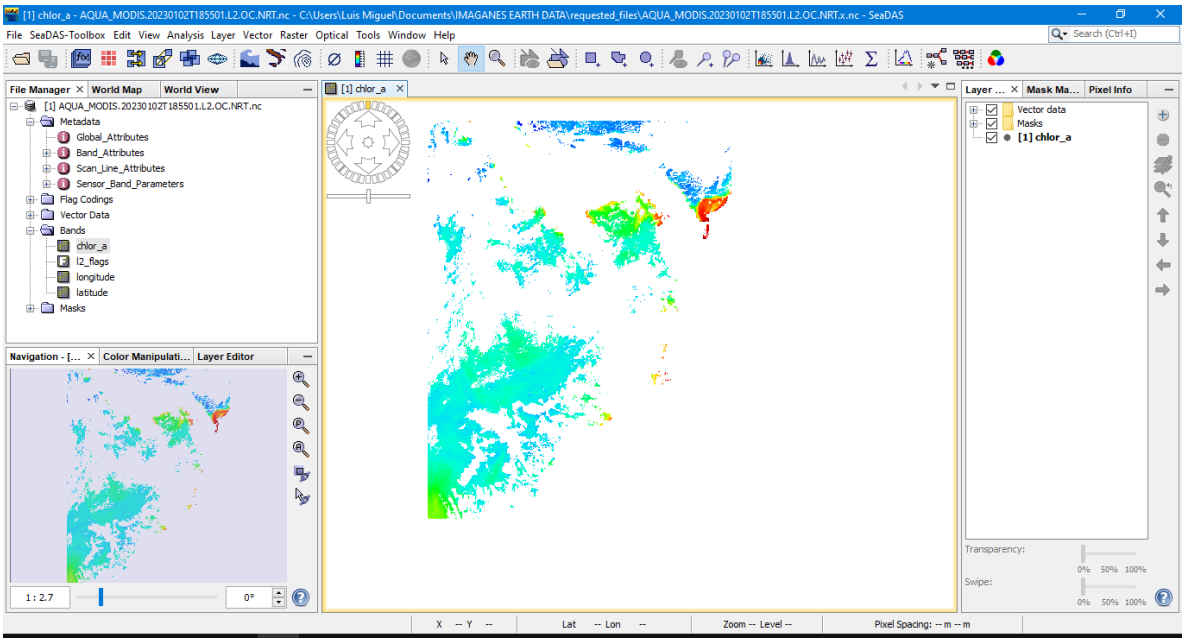


Figura 3 Imagen de clorofila descargada y visualizada en el software SEADAS

En conclusión, debido a que la metodología para descargar las imágenes MODIS de cada año implica seguir muchos pasos, se ha creado un documento el **cual cabe resaltar que está adjunto en el anexo 1** que explica el proceso paso a paso. De esta manera, se busca facilitar la comprensión del proceso de descarga y hacerlo más accesible para las personas.

3.3 Corregistro de las imágenes satelitales

En el marco del proyecto investigativo se desarrolló un Script para el co-registro masivo de imágenes MODIS. El objetivo de esta herramienta es automatizar el proceso, ya que al principio de la investigación solo se pudo hacer el co-registro de imagen por imagen, lo que resultó en una pérdida significativa de tiempo debido al tamaño del conjunto de datos de imágenes. La herramienta está construida en tkinter, una librería para construir interfaces

gráficas. Se componen de cinco campos que permiten seleccionar diferentes carpetas y archivos. El primer campo se utiliza para elegir la carpeta que contiene las imágenes a las que se les quiere hacer el co-registro. El segundo campo se utiliza para seleccionar un mosaico de archivos en formato XML, generado previamente por el software sellado. Este archivo contiene los parámetros necesarios para el co-registro del producto clorofila-a.

El tercer campo se utiliza para seleccionar la carpeta donde se guardarán las imágenes con co-registro ya realizado. El cuarto campo está diseñado para seleccionar una carpeta dentro del programa sellado que genera el archivo XML. Por último, se tiene el campo de archivo de texto, donde se guarda cada línea ejecutada automáticamente en la terminal de comandos. Esto permite llevar a cabo un registro de todas las imágenes co-registradas

En la **Figura 4** Ejecución interfaz Corregistrose puede ver cómo funciona la interfaz de la herramienta

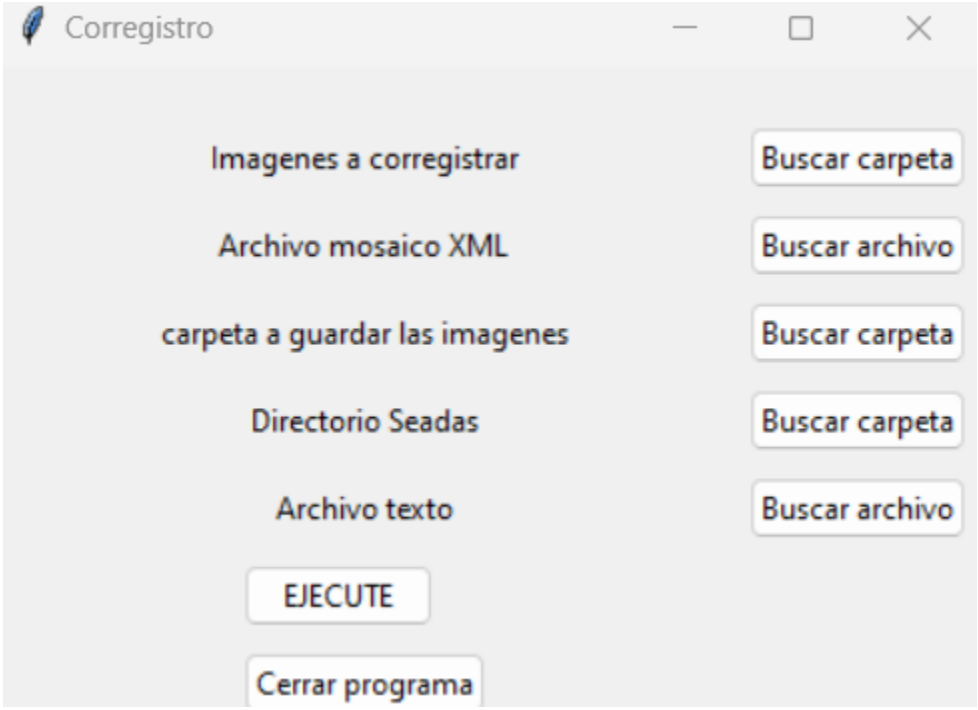


Figura 4 Ejecución interfaz Corregistro de imágenes satelitales MODIS

Al ejecutar el programa, se escribirá un comando en la consola. Este comando se ejecutará tantas veces como imágenes existan dentro de la carpeta seleccionada. En la **Figura 5** se puede observar el proceso en la consola al ejecutar el programa.

Ejecutable + ruta a archivo XML	Imagen a corregistrar	Imagen corregistrada
gpt.bat -e path_xml_file	Ssource=Path_modis_image.nc	Path_folder_destination/image_with_coregistration.nc

Figura 5 Línea ejecutable en la consola

3.4 Verificación de las imágenes

Se ha desarrollado una herramienta que permite verificar la calidad de las imágenes MODIS a través de los metadatos que contienen. Para utilizarla, es necesario seleccionar una carpeta que contenga las imágenes MODIS que se desean verificar. Una vez seleccionada y ejecutada la herramienta, en la consola se mostrarán los metadatos de cada imagen, como la fecha, el satélite que tomó la imagen, la latitud y longitud, el propietario, entre otros. La **Figura 6** Ejecución de interfaz gráfica, verificación de contenido imágenesilustra el funcionamiento de la herramienta.

Figura 6 Ejecución de interfaz gráfica, verificación de contenido imágenes

La **Figura 7** muestra la ejecución y los metadatos de cada imagen satelital MODIS.

Nombre	A2014006193500.L2_LAC_OC.nc
Satélite	MODIS
id	2018.0/L2/A2014006193500.L2_LAC_OC.nc
Creador	NASA/GSFC/OBPG
Latitud max	10.450525
Latitud min	-0.8281756
Longitud max	-79.46037
Longitud min	-87.345474

Figura 7 Metadatos de verificación de imagen satelital

4 IMPLEMENTACION SOFTWARE DE LA HERRAMIENTA OPERATIVA

Para la construcción del algoritmo de pronóstico se creó un framework con seis módulos, con el fin de dividir las tareas específicas y tener un código mantenible y escalable en el tiempo, cada módulo tiene funciones concretas que le dan funcionamiento al código y lo hacen más entendible. A continuación, se explicará brevemente **los seis módulos en seudocódigo, con su respectivo proceso describiéndolo de manera cronológica desde el modulo 1 hasta el 6. En el Anexo B se especifica los requerimientos de la herramienta operativa y en el Anexo C se especifica la arquitectura del sistema.**

4.1 Módulo para el procesamiento de NetCDFmodule.

Este módulo utiliza algunas bibliotecas Python, como netCDF4, os, glob y GDAL, para procesar imágenes satelitales en formato NetCDF. El objetivo principal del módulo es cargar las imágenes satelitales en el formato NetCDF desde un directorio dado, procesarlas y guardarlas en un formato más común y fácil de usar, como archivos de tipo TIFF.

El proceso comienza con la carga de las imágenes satelitales en formato NetCDF desde un directorio específico utilizando la biblioteca glob. Luego, se utiliza la biblioteca netCDF4 para extraer los valores de clorofila y nube de las imágenes y se guardan en archivos TIFF procesados. Para guardar los archivos procesados, se utiliza la biblioteca GDAL.

El resultado final de este módulo son imágenes procesadas de clorofila y nube en formato TIFF, las cuales se utilizan posteriormente en otros módulos del algoritmo para la estimación de

valores perdidos de clorofila. **La Figura 8** Módulo NetCDF muestra la estructura del módulo para el procesamiento de NetCDF.

Seudocódigo 1. Módulo NetCDF

```
Entradas:   Directorio de imágenes NetCDF
            Archivo ráster de referencia
Salida:     Lista de imágenes TIIF de Clorofila y Nubes
1:          Constructor de la clase
2:          lista = imagenesNetCDF
3:          Para imagen en lista
4:              imgCloud = imagen['clor_a_count'] // extraer característica nubes
5:              imgClor = imagen['chlor_a'] // extraer característica clorofila
6:              ImgNombres = imagen
7:          Fin para
8:          Crear carpeta
9:          Carpeta1 = 'modis_clo'
10:         Carpeta2 = 'modis_cloud'
11:         Guardar imágenes modis_clo y modis cloud en las carpetas correspondientes
```

Figura 8 Módulo NetCDF

4.2 Módulo para el procesamiento de ImageProcessing

En este módulo se utilizan librerías Python como numpy, os, glob y gdal. Este módulo es importante para realizar el procesamiento de las imágenes y aplicar una máscara de nubes a las mismas, lo que es esencial para obtener resultados precisos en la estimación de valores perdidos de clorofila. Además, el uso de una imagen de referencia permite mantener la georreferenciación y proyección de las imágenes de salida, lo que es importante para la posterior

integración y análisis de los datos. En la **Figura 9** Módulo ImageProcessingse evidencia la estructura del módulo.

Seudocódigo 2. Módulo ImageProcessing

```
Entradas:   Directorio con imágenes modis en formato .tif
            Directorio de salida
Salida:     Máscara de nubes en formato TIFF fullCloudMask
            Imágenes Modis normalizadas con la máscara de nubes
1:          Constructor de la clase
2:          lista = imagenesModisTiff
3:          Para imagen en lista
4:              sateliteImagen = imagen //Leer imagen como array
5:              ImgNombres = imagen
6:          Fin para
7:          Crear máscara de nubes
8:          Combinar todas las matrices de máscara de nubes
9:          Guardar la combinación en un archivo .tif fullCloudMask
10:         Normalizar cada imagen modis
11:         Para imagenNormalizada en sateliteImagen
12:             modisNormalizada = imagenNormalizada * sateliteImagen
13:         Fin Para
14:         Guardar cada imagen normalizada como archivo .tif en el directorio de salida
```

Figura 9 Módulo ImageProcessing

4.3 Módulo para el procesamiento de builderDataset

Este módulo construye un conjunto de datos de imágenes satelitales MODIS.Contiene una clase llamada builderDataset tiene un constructor que toma cinco argumentos: _path_to_modis, _spatialResolution, _timeResolution, _px_features y _name_dataset. El primer argumento es la

ruta al directorio que contiene las imágenes MODIS. El segundo argumento es la resolución espacial, que determina el tamaño de la ventana que se utilizará para extraer los valores de píxeles. El tercer argumento es la resolución temporal, que actualmente no se usa en el código. El cuarto argumento es el número de píxeles no cero que se requieren encada ventana para que se extraigan los valores de píxeles. El quinto argumento es el nombre del archivo donde se almacenará el conjunto de datos construido.

El método loadImages() cambia al directorio que contiene las imágenes MODIS y utiliza la biblioteca glob para obtener una lista de todos los archivos con extensión .tif. Luego, utiliza la biblioteca gdal para leer cada archivo y almacenarlos en una lista de imágenes. El método builder() construye el conjunto de datos utilizando una estrategia de ventana deslizante. Para cada imagen, recorre todos los píxeles y extrae los valores de los píxeles de una ventana de tamaño 2 * _spatialResolution + 1 x 2 *_spatialResolution + 1. Solo se extraen los valores de los píxeles de la ventana si el número de píxeles no cero en la ventana es igual a _px_features. Luego, los valores de los píxeles se almacenan en una lista que representa una fila del conjunto de datos.

Por último, el script tiene dos métodos para guardar el conjunto de datos. saveDataset() guarda el conjunto de datos como un archivo CSV. saveDataset2() guarda el conjunto de datos como un archivo CSV, pero con un encabezado para cada columna. En la **Figura 10** Módulo BuilderDataset se evidencia la estructura del módulo.

Seudocódigo 3. Módulo BuilderDataset

```
Entradas:  Directorio con imágenes modis en formato
           Resolución espacial
           Resolución temporal
           Numero de pixeles diferente de cero
           Ruta del directorio donde se almacenará el conjunto de datos
Salida:    Conjunto de datos a partir de imágenes Modis
1:         Constructor de la clase
2:         lista = imagenesModisTiff
3:         Para imagen en lista
4:             img = imagen //Leer imagen como array
5:         Fin para
6:         Si resolución temporal > 0 entonces
7:             //Pasar
8:         Sino
9:             Para datos en img
10:                Si datos[x][]y] != 0 entonces
11                     Obtener datos según resolución espacial y numero de pixeles
12                     Guardar los datos en una lista
13             Guardar los datos en un archivo csv
14:        Fin para
```

Figura 10 Módulo BuilderDataset

4.4 Módulo para el procesamiento de Trainer

Este código tiene como objetivo entrenar un modelo de regresión para predecir el valor de una variable objetivo ("clor_val") a partir de un conjunto de variables explicativas que se encuentran en un archivo csv ("datasetFile").

Este módulo lee un archivo csv que contiene un conjunto de datos, separa los datos en características y etiquetas (en este caso, 'clor_val' es la etiqueta), divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba, ajusta un modelo de regresión de Bayes

a los datos de entrenamiento y lo guarda en un archivo .sav. Finalmente, se imprimen las métricas de evaluación del modelo, como el error absoluto medio y el coeficiente de correlación R cuadrado.

En resumen, este código carga un conjunto de datos, divide el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba, entrena un modelo de regresión utilizando el conjunto de entrenamiento, hace predicciones sobre el conjunto de prueba, evalúa el rendimiento del modelo utilizando métricas de validación y guarda el modelo entrenado en un archivo. En la **Figura 11** Módulo Trainer se evidencia la estructura del módulo.

Seudocódigo 4. Módulo Trainer	
Entradas:	Conjunto de datos para entrenamiento y prueba del modelo //*datasetFile* Ruta del archivo donde se guardará el modelo
Salida:	Error absoluto medio Puntuación de correlación R^2 del modelo
1:	Cargar conjunto de datos
2:	Dividir conjunto de datos en entrenamiento y prueba
3:	x_train, x_test, y_train, y_test = dividirDatos() //datos entrada y datos objetivo
4:	Crear un objeto de regresión bayesiana
5:	Ajustar el modelo con el conjunto de entranamiento
6:	Modelo = RegresionBayesiana
7:	Modelo = Modelo.ajustar(x_train, y_train)
8:	Realizar una prediccion del conjunto de datos
9:	Predecir = Modelo.predecir(x_test)
10:	Calcular el error absoluto medio
11	Calcular la puntuación de correlación
12:	Guardar el modelo

Figura 11 Módulo Trainer

4.5 Módulo para el procesamiento de Tester

Este código carga varios modelos de regresión entrenados y luego los usa para predecir los valores de píxeles faltantes en una imagen MODIS. La imagen se carga desde un archivo y se procesa por píxeles, identificando los píxeles faltantes y seleccionando una ventana de píxeles adyacentes. Luego, se utiliza una función llamada "roitolog" para extraer los valores de los píxeles no faltantes en la ventana seleccionada y se usan como entrada para un modelo de regresión que estima el valor del píxel faltante. Se utilizan diferentes modelos de regresión según el número de píxeles no faltantes en la ventana. El resultado es una nueva imagen con los píxeles faltantes estimados y un mapa de confianza que indica la certeza de la estimación para cada píxel. En la **Figura 12** Módulo Testerse evidencia la estructura del módulo.

Seudocódigo 5. Módulo Tester

```
Entradas:   Matriz de valores de píxel de la imagen
            Resolución espacial
Salida:     Imagen de prediccion
            Mapa de confiabilidad
1:          Cargar modelos de entrenamiento
2:          Cargar la imagen Modis como un array
3:          Para imgModis in imagen
4:              Recorrer filas y columnas de imgModis
5:              Si fila y columna = 0.0 entonces
6:                  Obtener mapa de confidencia
7:                  Obtener nueva imagen pronosticada Modsos
8:          Final para
9:          Guardar nueva imagen Modis en formato TIFF
10:         Guardar mapa de confiabilidad en formato TIFF
```

Figura 12 Módulo Tester

4.6 Módulo para el procesamiento de Predictor

En este módulo se realizan dos tareas diferentes según los valores booleanos de dos variables: module_ImageProcessing y module_builderDataset. Si la primera es verdadera, se realiza el procesamiento de imágenes de clorofila y máscaras de nubes, y si la segunda es verdadera, se construye un conjunto de datos.

Este es un código de Python que utiliza varios módulos para procesar y construir un conjunto de datos a partir de imágenes satelitales.

El módulo tiene dos bloques condicionales que determinan qué módulos se utilizan. Si "module_ImageProcessing" es "True", se utiliza el módulo "ImageProcessing" para procesar imágenes. Si "module_builderDataset" es "True", se utiliza el módulo "builderDataset" para construir un conjunto de datos.

Si la condición "if module_ImageProcessing is True:", se definen varias rutas de archivo para cargar y guardar imágenes y máscaras. Luego se cargan imágenes TIFF de nubes y se crea una máscara general de nubes correspondiente a la serie temporal. Finalmente, se normalizan las imágenes MODIS según la máscara general de nubes.

Si el bloque "if module_builderDataset is True:", se define la ruta de archivo para cargar imágenes normalizadas MODIS y se llama a la función "builder()" del módulo "builderDataset" para construir un conjunto de datos a partir de esas imágenes. El conjunto de datos se guarda en un archivo CSV llamado "dataset_3_neigh.csv". En la **Figura 13** Módulo Predictorse evidencia la estructura del módulo.

Seudocódigo 6. Módulo Predictor	
Entradas:	Valor booleano que indica si se debe procesar las imágenes//*imageProcessing* Valor booleano que indica si se debe construir el conjunto de datos //builderDataset
Salida:	Conjunto de datos o procesamiento de imágenes
1:	**Si** imageProcessing = Verdadero entonces
2:	Se carga la ruta de las imágenes netCDF de clorofila Almacenar en un objeto del módulo NetCDF
3:	Cargar la ruta de las imágenes de nubes
4:	Crear la mascara general de nubes
5:	Normalizar las imágenes Modis conforme a la máscara de nubes
6:	**Si** builderDataset = Verdadero entonces
7:	Se cargan las imágenes normalizadas de clorofila
8:	Se crea un objeto builderDataset con la ruta de las imágenes, la resolución temporal, número de características por píxel y nombre del archivo de salida
9:	Se construye en conjunto de datos utilizando las imágenes cargadas

Figura 13 Módulo Predictor

5 RESULTADOS Y ANALISIS

5.1 Descripción del modelo

El Mean Absolute Error se encarga de promediar el error absoluto entre los valores reales y predichos, entre más cercano a cero más preciso es el modelo. Para el coeficiente de correlación R2, que varía entre -1 y 1, si es cercano a uno significa que las variables están relacionadas directamente. Los resultados anteriores indican que el modelo de **Linnear Regression y Bayesian Ridge**, se obtienen los valores más cercanos a los esperados en ambas vecindades.

Para brindar más precisión al pronóstico, en donde la efectividad del modelo a elegir juega un papel importante en la predicción que se haga. La configuración de los parámetros de cada modelo debe ser óptima para obtener un mejor entrenamiento. Las herramientas de aprendizaje automático reúnen las técnicas para descubrir modelos automáticamente y que a su vez tengan buen rendimiento en las tareas predictivas con muy poca participación del usuario.

Para entrenar el modelo con éxito, dada la abundancia de datos en el conjunto de dataset, se tomó la decisión de separarlos en una proporción de 70/30, asignando un 70% para el entrenamiento y un 30% para las pruebas, tal como se describe en los artículos del estado del arte. La validación del modelo no se llevó a cabo en el campo debido a las dificultades asociadas con la captura de valores de clorofila-a, que requieren viajes que resultan complicados de realizar por motivos de la DIMAR(Dirección General Marítima). Por esta razón, se decidió validar el proyecto utilizando todas las métricas de precisión disponibles. En una fase posterior, la DIMAR(Dirección General Marítima) tendrá la oportunidad de realizar la validación en el campo.

Las bibliotecas de código abierto están disponibles para usar los métodos de aprendizaje automático en Python, una de las más importantes en este campo es TPOT[49] (Tree-based Pipeline Optimization Tool), el cual es un algoritmo evolutivo que diseña y optimiza automáticamente la construcción de tuberías (pipelines) a través de la programación genética con el fin de encontrar un modelo de alto rendimiento para un determinado conjunto de datos. TPOT[49] es de código abierto y utiliza componentes de Scikit Learn para transformar datos, descomponer características, seleccionar las características y modelos óptimos.

En conclusión se eligió este modelo porque al compararlo con los diferentes modelos de regresión obtuvo el mejor resultado para el pronóstico de valores de clorofila y se utilizó la herramienta de aprendizaje automático TPOT[49] para optimizar el modelo. En la **Figura 14** Comparación de modelosse puede observar la comparación de los modelos.

Tabla 1. Resultados modelos con datasets de 2 y 8 vecindades.

Modelo	Error Vecindad 8 Px	Correlación Vecindad 8 Px	Error Vecindad 2 Px	Correlación Vedcindad 2 Px
Tree Decision	0.006087610174919019	0.8950951988964873	0.020458200557692312	0.7140430679302319
Linear Regression	0.004788768422705357	0.9421596401240928	0.01544390321375302	0.8439548796644282
LGBM Regressor	0.0049213686111166015	0.9292263548659834	0.020754620463105253	0.6954271017772398
SGD Regressor	0.016966233823612675	0.5563317139691233	0.04623685479977479	0.1892350114133790 6
Kernel Ridge	0.00517747393555572	0.9358741416829514	0.020639953374064008	0.8197554569119427
Elastic Net	0.025752672460926548	-1,32	0.05252432908435682	-0.0307729576
Bayesian Ridge	0.004789997764351299	0.9421565725656275	0.01544448924585 0106	0.8441034335671147
Gradient Boosting	0.0049368790731045145	0.9373169824436688	0.017095735491978622	0.8147426244359499
SVR	0.038360229105311616	-0.7858742980	0.05200003781086 3785	0.1202420595713 4415

Figura 14 Comparación de modelos

5.2 Descripción de la herramienta

Debido a la densidad de la cobertura nubosa, se ha optado por utilizar información espacial obtenida de imágenes MODIS. Estas imágenes contienen píxeles con valores de clorofila que pueden ser utilizados para pronosticar valores cercanos. Para ello, se ha utilizado una vecindad de píxeles nxn con n impar. Se ha elegido que n sea impar para poder tener un píxel central, el cual se utilizará como referencia tanto para encontrar el valor central perdido como para aplicar

un kernel 3x3 y determinar las conectividades del pixel central. Las conectividades que se han considerado son 2, 3, 4, 5, 6, 7 y 8, que son las vecindades que puede tener el valor central.

La **Figura 15** Imagen binaria con Vecindad Variable nxnmuestra la vecindad de pixeles.

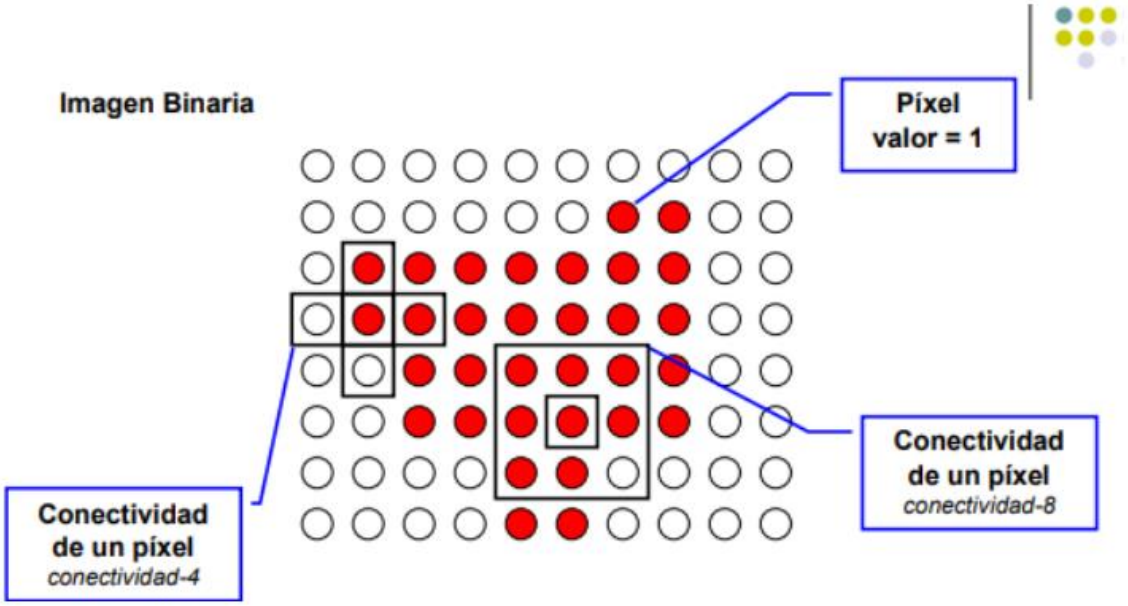


Figura 15 Imagen binaria con Vecindad Variable nxn

En el caso de una vecindad de tamaño 2 (n=2), donde N es un píxel del periodo, PV es un píxel válido y PP es el píxel que se desea pronosticar, el algoritmo puede encontrar doce combinaciones posibles dentro de una vecindad de 2 al utilizar un kernel de 3x3 con un solo píxel. Para entrenar la conectividad 2, el algoritmo recorre cada píxel de la imagen completa y almacena la información obtenida en un archivo CSV.

La **Figura 16** Posibles combinaciones de vecindad 2muestra todas las posibles combinaciones de vecindad dos.

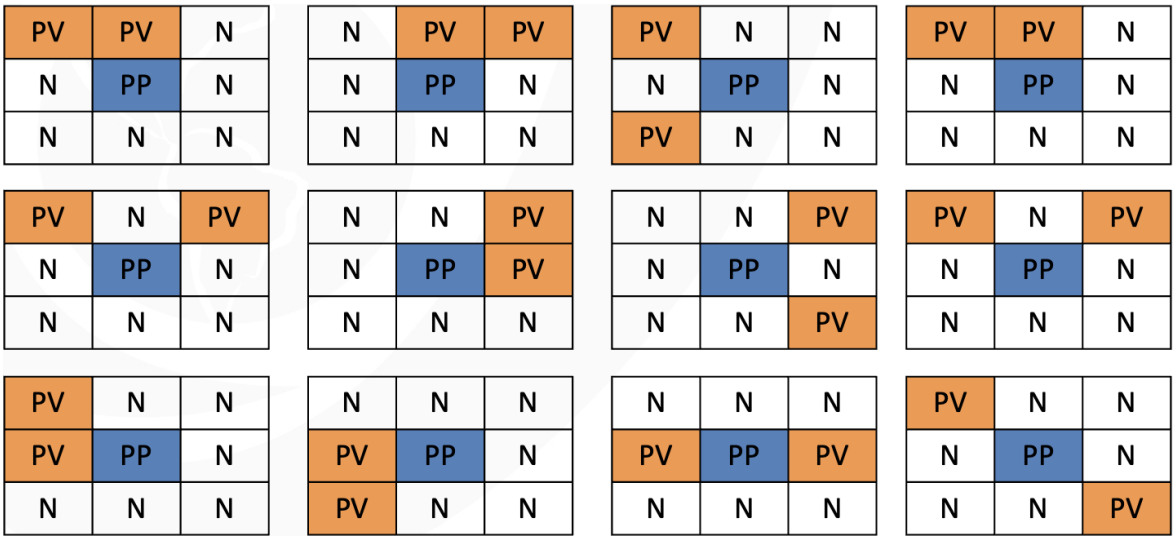


Figura 16 Posibles combinaciones de vecindad 2

Para poder entrenar el modelo, es necesario contar con un conjunto de datos (dataset) que incluya las características y el valor que se desea predecir, ya que es imposible entrenar un modelo sin estas variables. En el caso de una vecindad de 2, se recorre cada píxel de la imagen para identificar la vecindad del píxel central. El objetivo es proporcionarle al modelo estos datos para que aprenda y, en el futuro, pueda predecir el valor de un píxel faltante. En este modelo utilizamos la tarea de regresión, ya que se trata de valores continuos. Como métrica de validación, utilizamos el error cuadrático medio y el coeficiente de correlación (R2), el cual varía en un rango entre -1 y 1. Si R2 es cercano a 1, significa que las variables están directamente relacionadas: cuando las características aumentan, la variable objetivo también lo hace, y viceversa. Si R2 es cercano a -1, significa que las variables están relacionadas, pero de

forma inversa: cuando las características aumentan, la variable objetivo disminuye, y viceversa. El error nos indica qué tan bien se ajustó el modelo a los datos.

La **Figura 17** Conformación conjunto de datosmuestra un ejemplo de cómo se construye el dataset.

F1	F2	Clorofila
0,5	0,34	0,52
0,49	0,36	0,46
0,34	0,35	0,4

Figura 17 Conformación conjunto de datos clorofila con vecindad

Se entrenaron siete modelos de aprendizaje automático utilizando Scikit Learn, que es una biblioteca de Python que permite procesar datos antes de aplicar un modelo de aprendizaje automático. Scikit Learn ofrece una amplia variedad de modelos de aprendizaje supervisados, como regresión lineal, logística, árboles de decisión y bosques aleatorios. En este proyecto se utilizó el modelo de regresión.

Se utilizó OpenCV para recolectar el contenido de las imágenes satelitales, detectar características, crear regiones de interés y almacenar la información de cada una de las

imágenes. Además, se utilizó GDAL para trabajar con las imágenes, ya que estas contienen información de georreferenciación y metadatos. GDAL permitió cargar las imágenes junto con su georreferenciación y sus metadatos para trabajar con ellas.

Para entrenar el modelo, se crearon siete conectividades, que son el máximo número de entradas que se pueden encontrar aplicando un kernel 3x3 a la imagen. Estas conectividades van desde dos píxeles con vecindad central hasta ocho píxeles, teniendo en cuenta que el caso de ocho píxeles de vecindad es el mejor y el de dos píxeles es el peor. La **Figura 18** Posibles modelos con diferentes conectividadesmuestra los 7 modelos.

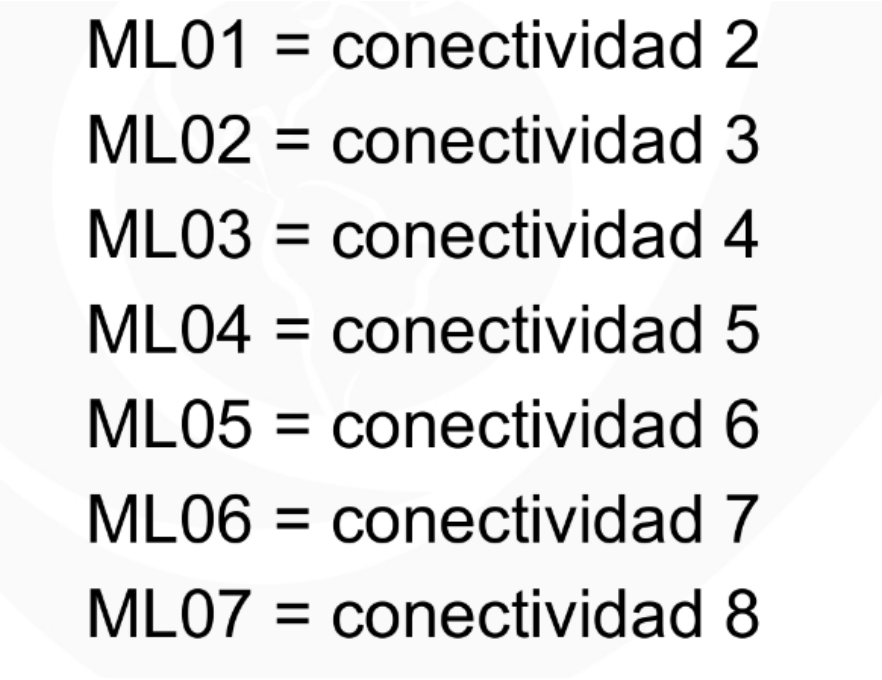
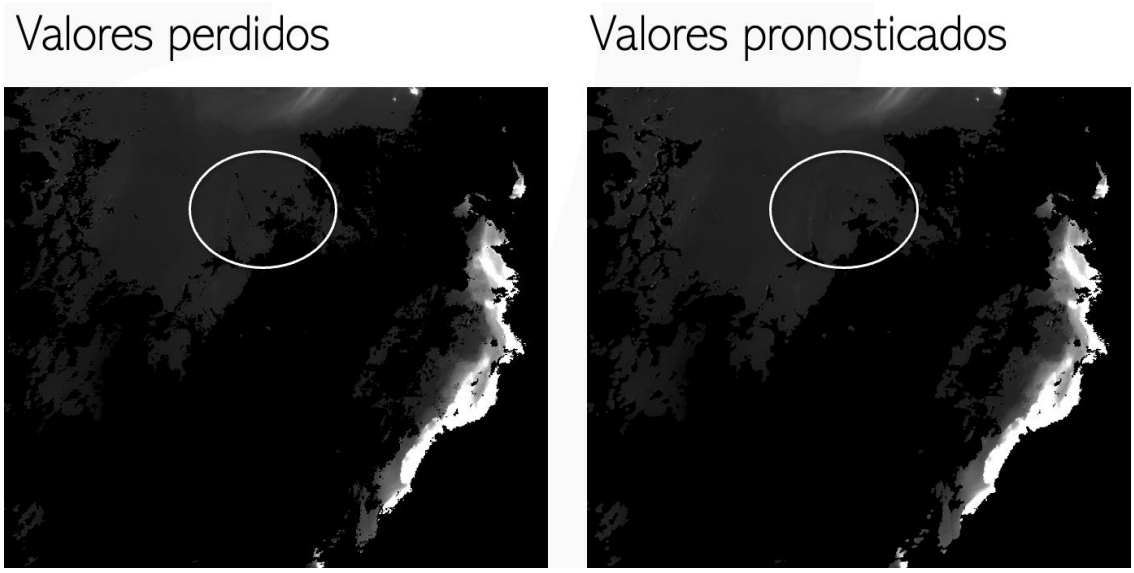


Figura 18 Posibles modelos con diferentes conectividades

5.3 Resultados

En la **Figura 19** Comparación de valores perdidos- valores pronosticadosse puede apreciar que la imagen de la izquierda presenta varios valores perdidos de clorofila debido a la nubosidad. Estos valores perdidos aparecen como píxeles negros, encerrados en un círculo, mientras que los píxeles que contienen clorofila se observan en tonos grises. Sin embargo, al aplicar el modelo de regresión, se pueden completar ciertos píxeles que no eran visibles en la imagen original de la izquierda. Esto se puede apreciar claramente en el círculo encerrado en la imagen de la derecha.



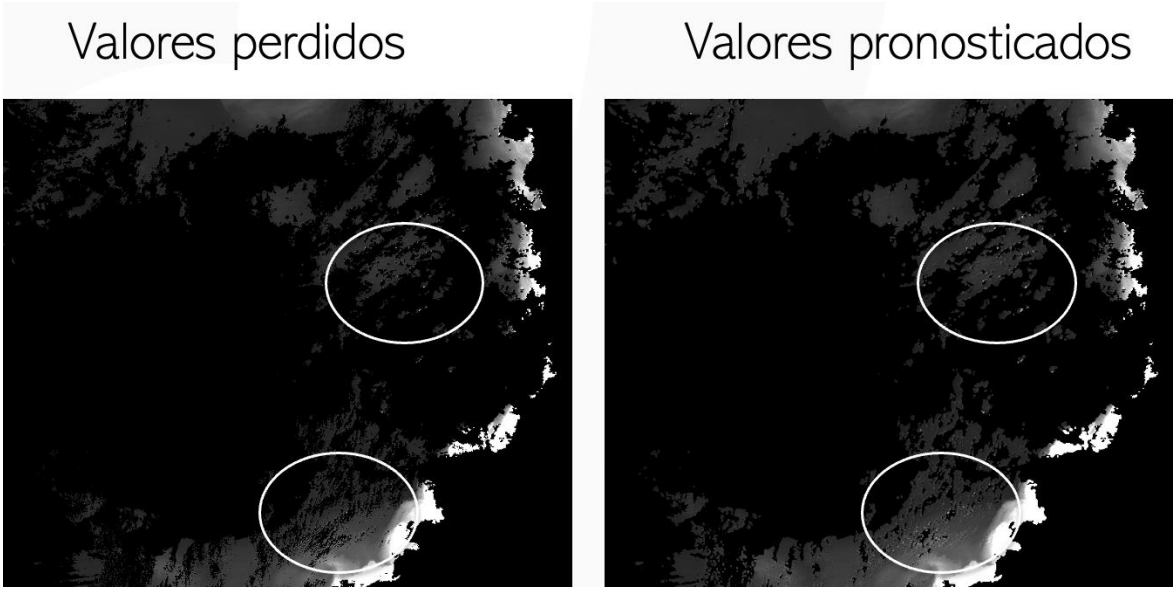


Figura 19 Comparación de valores perdidos- valores pronosticados

En la **Figura 20** Formula que se implementó para la Confianza del modelo, se puede observar la confianza del modelo al utilizar 100 iteraciones. Se puede apreciar que los píxeles válidos de clorofila se van agregando de manera brusca. Para calcular el valor de confianza, se toma el número de iteraciones y se divide por el kernel aplicado en el modelo, y luego se multiplica por el número de píxeles válidos de la imagen. Es importante tener en cuenta que, a medida que aumentan las iteraciones del modelo, la confianza disminuye.

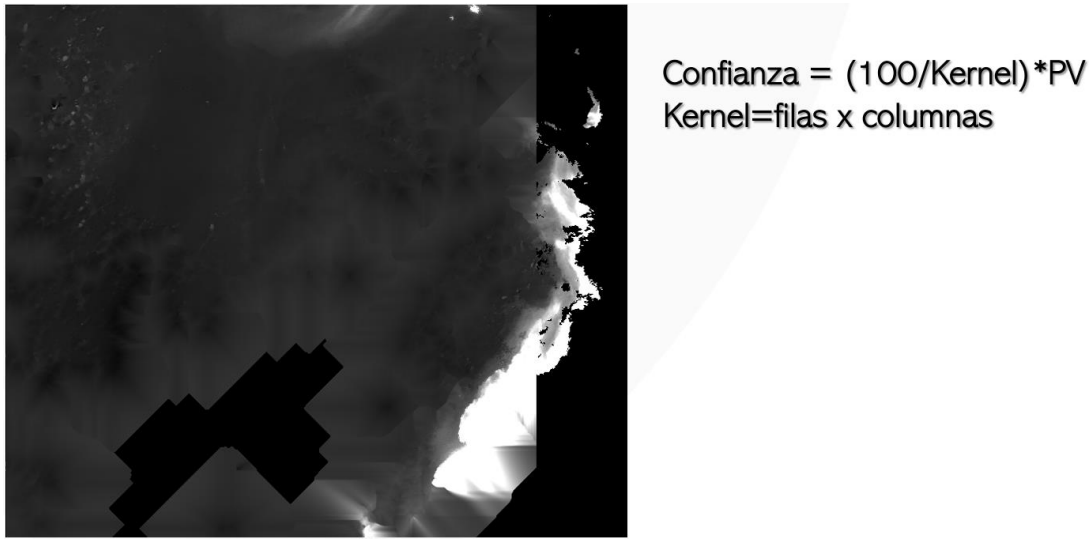


Figura 20 Formula que se implementó para la Confianza del modelo

6 CONCLUSIONES Y TRABAJOS A FUTURO

6.1 Conclusiones

Se desarrolló una guía metodológica para el proceso de descarga de las imágenes MODIS, explicando paso a paso todo el procedimiento que se debe realizar para obtener los lotes de imágenes que posteriormente fueron analizados, con la posibilidad de descargar información entre rangos de fechas y coordenadas específicas. Estas imágenes contienen información de clorofila y cobertura nubosa en las zonas de mar abierto, siendo esenciales para la construcción de un dataset que posteriormente se utilizó para el entrenamiento de un modelo de machine learning.

Se presenta un algoritmo de pronóstico construido en un marco con seis módulos, cada uno con funciones específicas que permiten un código escalable y mantenible. Uno de los módulos es el NetCDF, que carga imágenes NetCDF, las procesa y las almacena en directorios específicos en formato TIFF utilizando la biblioteca gdal. Otro módulo es el ImageProcessing, que es un script de procesamiento de imágenes que utiliza diferentes bibliotecas de Python y permite la creación de una máscara de nubes y la aplicación de la misma a las imágenes de satélite. El tercer módulo es el builderDataset, que construye un conjunto de datos de imágenes satelitales utilizando imágenes MODIS y se basa en la información temporal o espacial. En general, estos módulos contribuyen a la creación de un algoritmo de pronóstico sólido y efectivo.

Se construyó una herramienta operativa para el ajustar y alinear cada imagen georreferenciada, este proceso es importante para corregir distorsiones y desplazamientos entre las imágenes, haciendo que cada píxel de cada imagen corresponde geo referencialmente a ese mismo píxel de todas las imágenes dentro del rango de coordenadas previamente establecido. Al poseer una gran cantidad de imágenes, algunas de ellas podrían tener errores o estar corruptas, para ello se desarrolló un algoritmo el cual permite verificar los metadatos de cada imagen MODIS y así poder corroborar que está correcta y puede ser procesada para realizar el respectivo análisis, mostrando en consola datos como: fecha, cantidad de pixeles, satélite…etc.

6.2 Trabajos a futuro

Los autores platean como trabajos a futuro:

- Probar con diferentes modelos de aprendizaje automático para saber cuál es el modelo más optimo
- Enriquecer el conjunto de datos con las nuevas imágenes que puedan ser publicadas a futuro
- Se propone diseñar una versión web para poder realizar las consultas desde cualquier servidor
- Implementar un análisis temporal, analizando después del nivel de confiabilidad durante periodos de tiempo

7 REFERENCIA

[1] MADR, "Acuicultura en Colombia, cadena de la Acuicultura," Minist. Agric. y Desarro. Rural, pp. 1–30, 2021, Accessed: Nov. 28, 2021. [Online]. Available: **https://sioc.minagricultura.gov.co/Acuicultura/Documentos/2020-12-30%20Cifras%20Sectoriales.pdf**

[2] Universidad Externado de Colombia, "Análisis sobre la práctica pesquera no sostenible en el pacífico colombiano - Derecho del Medio Ambiente," 2020. https://medioambiente.uexternado.edu.co/analisis-sobre-la-practica-pesquera-no-sostenible-en-el-pacifico-colombiano/ (accessed Nov. 28, 2021).

[3] OCDE, "Pesca y acuicultura en Colombia," J. Sea, vol. 1, p. 34, 2016, Accessed: Nov. 28, 2021. [Online]. Available: **https://www.oecd.org/tad/fisheries/Fisheries_Colombia_SPA_rev.pdf.**

[4] E. Chuvieco Salinero, Teledetección ambiental la observación de la tierra desde el espacio. Ariel, 2008.

[5] M. M. Krekova, G. M. Krekov, I. V. Samokhvalov, and V. S. Shamanaev, "Numerical evaluation of the possibilities of remote laser sensing of fish schools," Appl. Opt., vol. 33, no. 24, p. 5715, Aug. 1994, doi: 10.1364/ao.33.005715.

[6] V. V. Klemas, "Advances in fisheries applications of remote sensing," in *Measuring and Modeling of Multi-Scale Interactions in the Marine Environment - IEEE/OES Baltic International Symposium 2014, BALTIC 2014*, 2014, doi: 10.1109/BALTIC.2014.6887836.

[7] F. S. Uslu, H. Binol, and A. Bal, "Food inspection using hyperspectral imaging and SVDD," in *Sensing for Agriculture and Food Quality and Safety VIII*, May 2016, vol. 9864, p. 98640N, doi: 10.1117/12.2223938.

[8] V. Klemas, "Remote sensing of environmental indicators of potential fish aggregation: An overview," *Baltica*, vol. 25, no. 2, pp. 99–112, Dec. 2012, doi: 10.5200/baltica.2012.25.10.

[9] A. Baudena *et al.*, "Fine-scale structures as spots of increased fish concentration in the open ocean," *Sci. Rep.*, vol. 11, no. 1, pp. 1–13, Aug. 2021, doi: 10.1038/s41598-021-94368-1.

[10] C. Ferronato, V. A. Guinder, M. P. Chidichimo, C. López-Abbate, and M. Amodeo, "Zonation of protistan plankton in a productive area of the Patagonian shelf: Potential implications for the anchovy distribution," *Food Webs*, vol. 29, p. e00211, Dec. 2021, doi: 10.1016/j.fooweb.2021.e00211.

[11] NASA, "MODIS Web," *MODIS Web*, 2014. https://modis.gsfc.nasa.gov/data/dataprod/chlor_a.php (accessed Nov. 28, 2021).

[12] J. Muñoz, L. Pencue, A. Figueroa, and C. Guzmán, "Crop monitoring in high andean ecosystems of the upper basin of the palacé river using planet images," in *Advances in Intelligent Systems and Computing*, Nov. 2018, vol. 687, pp. 155–169, doi: 10.1007/978-3-319-70187-5_12.

[13] E. L. Pencue-Fierro, Y. T. Solano-Correa, J. C. Corrales-Muñoz, and A. Figueroa-Casas, "A Semi-Supervised Hybrid Approach for Multitemporal Multi-Region Multisensor Landsat Data Classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 9, no. 12, pp. 5424–5435, Dec. 2016, doi: 10.1109/JSTARS.2016.2623567.

[14] C. Muñoz-Ordóñez, J. Muñoz-Ordóñez, L. Pencue-Fierro, and A. Figueroa, "Mapeo de cobertura terrestre utilizando aprendizaje máquina," *Investig. e Innovación en Ing.*, vol. 8, no. 3, pp. 85–101, Nov. 2020, doi: 10.17081/invinno.8.3.4706.

[16] K. S. Pratt, "Design Patterns for Research Methods: Iterative Field Research," AAAI Spring Symp. Exp. Des. Real, no. 1994, pp. 1–7, 2009, Accessed: Nov. 26, 2020. [Online]. Available: **http://kpratt.net/wp-content/uploads/2009/01/research_methods.pdf**

[17] Ecologia Verde, "Qué es la clorofila y sus tipos," 2020. https://www.ecologiaverde.com/que-es-la-clorofila-y-sus-tipos-2579.html#:~:text=Clorofila%20A&text=Está%20presente%20en%20los%20cloroplastos,energía%20lumínica%20en%20energía%20química. (accessed May 23, 2022).

[18] Axess Ntwork, "Las imágenes satelitales y sus aplicaciones en la vida cotidiana," 2019. https://axessnet.com/las-imagenes-satelitales-y-sus-aplicaciones-en-la-vida-cotidiana/#:~:text=Las%20imágenes%20satelitales%20también%20conocidas,emitida%20o%20reflejada%20por%20un (accessed Mar. 02, 2023).

[19] Administración de Parques Nacionales (APN), "Protocolo para el pre procesamiento de imágenes satelitales Landsat para aplicaciones de la Administración de Parques Nacionales," *Sistema de información de Biodiversidad*, 2005.

[20] Terra, "Los beneficios MODIS," Aug. 23, 2011. http://www.terra-i.org/es/news/news/The-benefits-of-MODIS.html (accessed May 23, 2022).

[21] SCIELO, "Nuevo Sistema Satelital de Monitoreo Oceánico," 2011. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-38802011000200010 (accessed May 23, 2022).

[22] Wikipedia, "Predictive modelling," Apr. 2016. https://en.wikipedia.org/wiki/Predictive_modelling (accessed Mar. 02, 2023).

[23] SAS, "Artificial Neural Networks What they are & why they matter." https://www.sas.com/en_us/insights/analytics/neural-networks.html (accessed Mar. 02, 2023).

[24] Scikit Learn, "Linear Models," 2021. https://scikit-learn.org/stable/modules/linear_model.html#bayesian-regression (accessed Mar. 02, 2023).

[25] IBM, "What is a data set?," 2010. https://www.ibm.com/docs/en/zos-basic-skills?topic=more-what-is-data-set (accessed Mar. 02, 2023).

[26] Luke Dormehl, "What is an artificial neural network? Here's everything you need to know | Digital Trends," *Https://Www.Digitaltrends.Com/Cool-Tech/What-Is-an-Artificial-Neural-Network/*, 2018.

[27] Aditya Mishra, "Metrics to Evaluate your Machine Learning Algorithm," Feb. 2018. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (accessed May 31, 2022).

[28] Unidata, "Network Common Data Form (NetCDF)," p. 07, 2022, Accessed: Oct. 01, 2022. [Online]. Available: https://www.unidata.ucar.edu/software/netcdf/

[29] M. L. Kessler, "Image registration and data fusion in radiation therapy," *British Journal of Radiology*, vol. 79, no. SPEC. ISS. 2006. doi: 10.1259/bjr/70617164.

[30] T. Amaratunga, "What Is Deep Learning?," in *Deep Learning on Windows*, 2021. doi: 10.1007/978-1-4842-6431-7_1.

[31] G. Ruppert, M. Hussain, and H. Müller, "Accuracy Assessment of Satellite Image Classification Depending on Training Sample," *Austrian Journal of Statistics*, vol. 28, no. 4, 2016, doi: 10.17713/ajs.v28i4.522.

[32] Kolmite, "GDAL – BIBLIOTECA DE ABSTRACCIÓN DE DATOS GEOESPACIALES," Nov. 2022, Accessed: Jul. 02, 2022. [Online]. Available: https://kolmite.com/gdal-biblioteca-de-abstraccion-de-datos-geoespaciales/

[33] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[34] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn," *GetMobile: Mobile Computing and Communications*, vol. 19, no. 1, 2015, doi: 10.1145/2786984.2786995.

[35] C. S. Rousseaux and W. W. Gregg, "Forecasting ocean chlorophyll in the equatorial Pacific," *Front Mar Sci*, vol. 4, no. JUL, 2017, doi: 10.3389/fmars.2017.00236.

[36] M. Jouini, M. Lévy, M. Crépon, and S. Thiria, "Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method," *Remote Sens Environ*, vol. 131, 2013, doi: 10.1016/j.rse.2012.11.025.

[37] H. Yajima and J. Derot, "Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases," *Journal of Hydroinformatics*, vol. 20, no. 1, 2018, doi: 10.2166/hydro.2017.010.

[38] F. Alfonita, "IDENTIFICACIÓN ESPACIO TEMPORAL DE CLOROFILA-α MEDIANTE IMÁGENES DE SATÉLITE EN SISTEMAS ACUÁTICOS," *Comput Ind Eng*, vol. 2, no. January, 2018.

[39] H. Cho and H. Park, "Merged-LSTM and multistep prediction of daily chlorophyll-A concentration for algal bloom forecast," in *IOP Conference Series: Earth and Environmental Science*, 2019, vol. 351, no. 1. doi: 10.1088/1755-1315/351/1/012020.

[40] M. Penteliuc and M. Frincu, "Prediction of Cloud Movement from Satellite Images Using Neural Networks," in *Proceedings - 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2019*, 2019. doi: 10.1109/SYNASC49474.2019.00038.

[41] Y. G. Ham, Y. S. Joo, and J. Y. Park, "Mechanism of skillful seasonal surface chlorophyll prediction over the southern Pacific using a global earth system model," *Clim Dyn*, vol. 56, no. 1–2, 2021, doi: 10.1007/s00382-020-05403-2.

[42] J. Hänninen, K. Mäkinen, K. Nordhausen, J. Laaksonlaita, O. Loisa, and J. Virta, "The 'Seili-index' for the Prediction of Chlorophyll-α Levels in the Archipelago Sea of the northern Baltic Sea, southwest Finland," *Environmental Modeling and Assessment*, 2022, doi: 10.1007/s10666-022-09822-9.

[43] E. Eze, S. Kirby, J. Attridge, and T. Ajmal, "Time Series Chlorophyll-A Concentration Data Analysis: A Novel Forecasting Model for Aquaculture Industry," 2021. doi: 10.3390/engproc2021005027.

[44] Z. Du, S. Wu, M. P. Kwan, C. Zhang, F. Zhang, and R. Liu, "A spatiotemporal regression-kriging model for space-time interpolation: a case study of chlorophyll-a prediction in the coastal areas of Zhejiang, China," *International Journal of Geographical Information Science*, vol. 32, no. 10, 2018, doi: 10.1080/13658816.2018.1471607.

[45] Z. L. Q. R. H. C. M. S. Y. W. Han Ding, "Single-variable method for predicting trends in chlorophyll a concentration based on the similarity of time series," *ScienceDirect*, vol. 140, 2022.

[46] Asana, "Cómo entender los procesos iterativos," *2022*, 12AD. https://asana.com/es/resources/iterative-process (accessed Jan. 03, 2022).

[47] Nasa, "Ocean color." https://oceancolor.gsfc.nasa.gov/ (accessed May 26, 2023).

[48] NASA, "NASA SeaDas." https://seadas.gsfc.nasa.gov (accessed May 30, 2023).

[49] Randal S. Olson, "TPOT." https://epistasislab.github.io/tpot/ (accessed Jun. 03, 2023).

8 ANEXO

Anexo A

Anexo B

Anexo C