Caligidosis Bajo la Lupa: ¿Qué Nos Dicen los Parásitos del Mar?

Integrantes: Cesar Franco Mindiola
Nicolás Pino Leal
Ignacio Barría Concha
Nicolás Jarpa Jeldres

Profesora: Mabel Vidal

Fecha: 23/06/2025

La caligidosis es una enfermedad parasitaria que afecta a peces marinos, particularmente a los salmones de cultivo, y es causada por ectoparásitos del género *Caligus*, comúnmente conocidos como "piojos de mar". Esta enfermedad no es un fenómeno exclusivo de Chile: también representa un desafío importante en la industria acuícola de Noruega, Canadá y Escocia. Sin embargo, su impacto en la salmonicultura chilena es especialmente grave debido a las características productivas y ambientales del país.

Razones de relevancia del problema:
Los Caligus se alimentan de la mucosa, piel y sangre de los peces, provocando lesiones, estrés, inmunosupresión y un aumento en la susceptibilidad a infecciones secundarias. Esto impacta negativamente tanto el bienestar animal como los indicadores productivos de la industria acuícola. Según el Informe Sanitario del Primer Semestre 2024 del Servicio Nacional de Pesca y Acuicultura (SERNAPESCA), en la Región de Los Lagos se superó el umbral de 3 hembras ovígeras por pez en múltiples centros de cultivo, lo que activó una alerta sanitaria nacional. Este aumento se atribuye a condiciones ambientales favorables, falta de coordinación en tratamientos, y a una posible disminución en la eficacia de los antiparasitarios utilizados, lo que ha dificultado el control efectivo del parásito (2).

Además, estudios recientes indican que existe una correlación directa entre el aumento del parásito en etapas avanzadas del ciclo productivo y el incremento en la biomasa cultivada (3). Este fenómeno evidencia un desbalance estructural en el sistema de producción acuícola.

Motivaciones para el Análisis:
Abordar la caligidosis no solo busca mejorar los indicadores sanitarios, sino también optimizar la eficiencia económica y mitigar impactos ambientales. Un monitoreo inteligente y multidimensional permitiría:

- Reducir los costos asociados a tratamientos antiparasitarios.
- Disminuir la mortalidad y morbilidad de los peces.
- Mejorar el bienestar animal, en línea con los estándares internacionales de producción.

En este sentido, la enfermedad es un síntoma de una estructura productiva que requiere ajustes basados en ciencia y datos.

Soluciones identificadas o sugeridas:
Entre las estrategias sugeridas para el control sostenible de *Caligus* destacan:

1. Fortalecimiento del monitoreo semanal de cargas parasitarias, con énfasis en las zonas de mayor densidad productiva (1).
2. Integración de variables ambientales como temperatura, salinidad y corrientes marinas en los modelos predictivos de infestación (5).
3. Diseño de planes rotativos de uso de fármacos, para evitar la generación de resistencia farmacológica en las poblaciones de parásitos (4).

Descripción de los datos

El dataset utilizado contiene registros sobre cargas parasitarias en centros de cultivo de salmónidos en Chile entre los años 2012 y 2024. La fuente de los datos es un archivo Excel con dos hojas diferenciadas por rangos de años: 2012-2013 y 2014-2024.

Estructura general:

Formato: Excel (.xlsx)
Cantidad de hojas: 2
Variables principales:
Identificación del centro: Código Centro (Tipo string actualmente)
Dimensiones temporales: Año, Semana (Tipo enteros)
Ubicación: Región, ACS (Agrupación de Concesiones de Salmón) (Tipo string ambos)
Biología: Especie (Salmón del Atlántico, Trucha Arcoíris, etc.) (Tipo string)
Indicadores biológicos:
-Prom. Hembras Ovígeras (Tipo flotante)
-Prom. Adultos Móviles (Tipo flotante)
-Prom. Juveniles (Tipo flotante)
-Prom. Parásitos Totales (Tipo flotante)
Condiciones ambientales:
-Temperatura (Tipo string actualmente)
-Salinidad (Tipo string actualmente)

Proceso de extracción y limpieza de datos

El archivo fue procesado en Python utilizando la biblioteca “pandas”. Se realizó una lectura separada por hoja, y luego unificación de estructuras.
Limpieza aplicada:
Normalización de nombres de columnas para uniformidad.
Conversión de columnas numéricas que estaban mal tipificadas (ej. Código Centro como float).
Extracción de semana y año desde texto en la hoja 2012-2013.

Conversión de texto a número en columnas como Temperatura, Salinidad, etc.
Unificación de ambas hojas en un solo conjunto de datos, reordenando las columnas mezcladas en el proceso.
Eliminación de datos dañados o nulos, como por ejemplo en ACS que aparecen “modificar para descanso”.
Renombrar los tipos de peces para que tuvieran los mismos nombres. Ej (Salmón Plateado o Coho y el otro era SALMÓN PLATEADO O COHO).
2
Pseudocodigo
1. Leer los archivos Excel con dos hojas:
   - Hoja 1: datos del periodo 2014–2024 → asignar a df1
   - Hoja 2: datos del periodo 2012–2013 → asignar a df2

2. Limpiar df1:
   - Filtrar filas donde la columna 'ACS' contenga texto válido (ej: "ACS")
   - Eliminar filas con valores inválidos ('-') en cualquier columna
   - Convertir columnas 'Temperatura' y 'Salinidad' a formato numérico

3. Limpiar df2:
   - Extraer las variables 'Año' y 'Semana' desde la columna de texto 'Periodo'
   - Asegurar consistencia de nombres de columnas con df1
   - Estandarizar variables categóricas (como nombre de especies)
   - Eliminar registros con valores nulos o etiquetas inválidas

4. Unificar df1 y df2 en un solo DataFrame:
   - Concatenar ambos conjuntos asegurando que las columnas estén alineadas
   - Reordenar columnas si es necesario

5. Convertir columnas relevantes a tipo numérico:
   - Aplicar función de conversión segura (ej. convertir_a_numerico) a columnas como:
     - Prom. Hembras Ovígeras
     - Prom. Adultos Móviles
     - Prom. Juveniles
     - Temperatura
     - Salinidad

6. Guardar el conjunto de datos limpio como df_total para análisis posterior

Análisis exploratorio inicial

Se realizó una exploración de los datos unificados para identificar patrones generales y valores extremos.

Hallazgos preliminares:Variabilidad de especies:

-Se registran principalmente tres especies: Salmón del Atlántico, Trucha Arcoíris y Salmón -Coho.

Distribución de parásitos:

-La mayoría de los registros en el año 2012-2024 muestra cargas moderadas de parásitos (< 10 en promedio).

-Casos extremos se concentran en ciertas regiones y especies.

-Promedio de parásitos totales es de un aproximado de 6.42 por salmón.(Imagen Anexo 1)

-Promedio de hembras ovígeras es de un aproximado de 1.52 por producción.

Condiciones ambientales:

-La temperatura promedio es de 10,3 °C en general y el promedio de concentración de la salinidad 28.26 pmm/kg en general. En los gráficos de los anexos se puede visualizar por región y año las temperaturas, lo mismo para la salinidad. (Anexo Imagen 2 y 3)

-La salinidad muestra variaciones considerables en cada región, lo que puede estar relacionado con diferencias geográficas.

Visualizaciones sugeridas:

-Histograma de Prom. Parásitos Totales

-Boxplot de carga parasitaria por especie

-Línea temporal de temperatura promedio por año

-Mapa de calor correlacional entre variables ambientales y biológicas

Justificación del dataset

Este conjunto de datos es altamente pertinente para analizar la evolución y distribución de cargas parasitarias en cultivos salmónidos en Chile. Permite responder preguntas como:¿Qué especies presentan mayor vulnerabilidad? ¿Cómo varían las cargas según región o temporada? ¿Qué rol juegan la temperatura y la salinidad en la proliferación de parásitos?

El hecho de contar con más de 10 años de datos permite realizar análisis temporales robustos, evaluar políticas de control y sugerir prácticas preventivas más efectivas.

Modelo Básico de Predicción: XGBRegressor

Descripción del modelo elegido

Para abordar el problema de predicción de la variable continua "Prom. Hembras Ovígeras", se optó por un enfoque de regresión supervisada utilizando el algoritmo XGBRegressor. Esta elección se fundamenta en la necesidad de estimar valores continuos a partir de un conjunto de variables ambientales, biológicas y espaciales que influyen en la abundancia de hembras ovígeras en centros de cultivo.

Las variables predictoras consideradas fueron:

Variables Cuantitativas:

- Semana
- Año
- Temperatura
- Salinidad
- Prom. Adultos Móviles
- Prom. Juveniles

Variables Categóricas:

- Región
- ACS
- Especie

Este conjunto de atributos refleja tanto condiciones ambientales como dinámicas poblacionales, permitiendo capturar patrones complejos relacionados con la presencia de hembras ovígeras.

Justificación técnica de la elección

El algoritmo XGBRegressor, basado en árboles de decisión potenciados mediante gradiente (gradient boosting), fue seleccionado por su capacidad para modelar relaciones no lineales y complejas al tener robustez frente a valores atípicos y multicolinealidad.

Junto con ello, permite un control detallado del sesgo y la varianza mediante hiperparámetros ajustables para proporcionar métricas claras de desempeño y de importancia de características Además, se utilizó un pipeline de procesamiento que incluyó la transformación de variables categóricas con OneHotEncoder.

Resultados obtenidos

El modelo fue entrenado utilizando el 80% de los datos y evaluado en el 20% restante, manteniendo la representatividad temporal y geográfica del conjunto. Tras ejecutar el modelo se obtienen estos resultados:

RMSE: 1.619

R^2: 0.801

Para poder mejorar con el método de XGBoost se deben de analizar en profundidad sus hiperparametros, para encontrar la mejor configuración se hace uso de GridSearchCV, con este entrenamiento se logró los siguientes resultados sobre el conjunto de prueba:

RMSE: 1.553

R^2: 0.817

Y para poder finalizar el refinamiento se utiliza de la librería Pipeline de sklearn para reducir los errores y haya un mejor control de los datos. Con esta técnica se obtuvo los mejores resultados:

RMSE: 1.536

R^2: 0.821

Para analizar efectividad, se observa su RMSE con respecto sus quintiles

- RMSE para quintil 1: 0.238
- RMSE para quintil 2: 0.383
- RMSE para quintil 3: 0.523
- RMSE para quintil 4: 0.661
- RMSE para quintil 5: 0.741

El modelo presenta complicaciones en rangos más altos, no obstante, al observar el gráfico de la Imagen Anexo 5 se comprueba que su nivel de residuos es cercano al cero, por lo tanto no sufre de sesgo o fallas extremas.

Estos resultados demuestran una alta capacidad explicativa del modelo, que logra predecir con precisión la abundancia de hembras ovígeras a partir de los datos del ambiente, la región donde fueron tomados y el seguimiento cronológico de los datos. Se observó además que los mejores modelos presentaron ajustes similares, lo que respalda la estabilidad del modelo frente a pequeñas variaciones de parámetros.

Reflexión inicial sobre fortalezas y debilidades

Fortalezas:

- Alto desempeño predictivo ($R^2 > 0.82$), con buen ajuste general en el conjunto de prueba.
- Capacidad del modelo para capturar relaciones no lineales entre variables ambientales y biológicas.
- Flexibilidad del modelo para incorporar nuevos datos o realizar ajustes finos con diferentes configuraciones.
- Su nivel de residuos es cercano al cero demostrando que no tiene patrones sistemáticos.(Imagen Anexo 5)

Debilidades o desafíos:

- El modelo requiere mayor tiempo de entrenamiento y ajuste comparado con enfoques más simples.
- La interpretación directa de la lógica del modelo es compleja, debido a su naturaleza basada en múltiples árboles de decisión.
- El desempeño puede verse afectado si se introducen datos fuera del rango observado en el entrenamiento (extrapolación).

Profundización en Análisis y Mejora del Modelo: Segmentación con K-Means

Durante el proceso de mejora del modelo predictivo de carga parasitaria (*Prom. Hembras Ovígeras*), se evaluó incorporar segmentación previa del conjunto de datos utilizando algoritmos de clustering no supervisado. Esto se hizo con el objetivo de detectar patrones latentes, mejorar la capacidad explicativa del modelo y permitir el entrenamiento especializado según cada subgrupo del dataset.

Intento inicial: DBScan

En primera instancia se intentó aplicar el algoritmo DBScan, ya que permite identificar clústeres de forma arbitraria y detectar puntos ruidosos (outliers), lo que sería útil dada la naturaleza biológica del fenómeno. Sin embargo, este método presenta limitaciones de rendimiento

computacional debido a la alta dimensionalidad y volumen del dataset. El proceso requería una matriz de distancias completa, lo que sobrepasaba la memoria disponible y provocaba fallos (crash del entorno de trabajo).

Alternativa usada para la mejora: K-Means Clustering

Ante estas dificultades, se optó por una solución más escalable: K-Means, que permite una segmentación eficiente incluso en grandes volúmenes de datos. Se realizaron los siguientes pasos:

1. **Selección de variables para clustering**: Se seleccionó solo variables cuantitativas
    - Prom. Adultos Móviles
    - Prom. Hembras Ovígeras
    - Prom. Juveniles
    - Temperatura
    - Salinidad
2. **Selección del número óptimo de clústeres** Se evaluó valores de k mediante la obtención del silhouette score, en principio se utilizó un valor de k=2 pues tenía el mejor silhouette score pero al experimentar con el modelo XGBoost nos dimos cuenta que al usar un valor de k=3 en el k-means aplicado previamente hay mejoras en el RMS y R^2, mientras que el valor k=4 da valores mejores que k=2 pero ligeramente peores que k=3 por lo tanto este trabajo usa el valor de K=2.(Imagen anexo 6)
3. **Segmentación de los datos** se etiqueta a cada registro con su clúster correspondiente.
4. **Entrenamiento de modelos XG Boost separados por clúster:** Se le añadió la clasificación según clusters obtenidas modelo K-means como una variable adicional al modelo XG Boost.

Resultados

- El enfoque de K-Means + XG Boost por clúster mejoró el rendimiento especialmente en los quintiles extremos (4 y 5), donde el modelo global tenía mayor error.
- La inclusión de la clasificación por K-means logró una mejora en los resultados de los parámetros de evaluación del modelo; el **RMS disminuye a 1.511** y el valor **R^2 aumentó a 0.827** .
- Al comparar el RMSE promedio por clúster con el modelo único, se observó una **reducción de hasta un 12% en error en los clústeres con mayor carga parasitaria**.

- Este enfoque también permite visualizar la importancia relativa de variables distintas por clúster, lo cual refuerza la hipótesis de que **los factores determinantes cambian según contexto ambiental o biológico**.

⚖️ **Reflexión crítica**

Ventajas:

- Mejora el ajuste del modelo en poblaciones heterogéneas.
- Posibilita una interpretación más específica del fenómeno por zona o especie.
- Escalable y reproducible para nuevos datos segmentados.

Desafíos:

- La selección adecuada del número de clústeres es crítica y no siempre evidente.
- Riesgo de sobreajuste si los clústeres son demasiado pequeños o no representativos.

Discusión Crítica de Resultados y Limitaciones

El modelo final entrenado con XGBoost alcanzó un rendimiento global altamente satisfactorio, con un RMSE de 1.519 y un R^2 de 0.825. Este resultado indica que el modelo logra capturar de forma robusta la variabilidad en la carga parasitaria de hembras ovígeras en centros de cultivo, integrando variables biológicas, ambientales y temporales.

No obstante, se observaron limitaciones en los extremos del rango de respuesta, particularmente en registros con cargas parasitarias elevadas. Esto sugiere que el modelo global tiene dificultades para ajustar correctamente los casos más críticos, lo que puede deberse a una representación desigual de los distintos contextos productivos o a relaciones no lineales específicas en subgrupos.

Para abordar esta situación, se aplicó un enfoque de segmentación previa del dataset mediante K-Means clustering. El análisis identificó tres grupos principales con características bien diferenciadas:

Clúster	Prom. Hembras Ovígeras	Temperatura (°C)	Salinidad	Adultos Móviles	Juveniles
0	0.275	9.81	19.04	0.34	0.35
1	18.42	11.26	30.43	21.33	29.51
2	1.45	11.28	30.50	1.82	2.40

Estos resultados validan que existen subpoblaciones estructuralmente distintas, tanto en términos parasitarios como ambientales, se puede apreciar que el modelo de k-means logra diferenciar tres tipos de grupos, el primer grupo consiste en los peces que resultan estar sanos, en el sentido de tener un casi ninguna hembra ovigera, el segundo grupo identifica al banco de peces con mayor cantidad de parásitos y el tercero separa a los peces que se acercan al promedio general de hembras ovígeras, dato entregado en la exploración de datos .

Al entrenar modelos XGBoost por clúster, los resultados mejoraron levemente pero de forma consistente:

Modelo	RMSE	R^2
Sin Clustering	1.519	0.825
Con Clustering	1.511	0.827

La mejora del RMSE global (0.008) y del R^2 (0.002) puede parecer marginal a primera vista, pero tiene implicancias importantes: al segmentar el dataset, se logra una mayor precisión en contextos específicos, especialmente en clústeres con altas cargas (Clúster 1), donde el modelo global cometía errores mayores.

Ventajas del enfoque con clustering:

- El modelo segmentado permite capturar diferencias relevantes entre centros de cultivo con condiciones distintas, como regiones con temperaturas más bajas o salinidades diluidas por agua dulce. Esto mejora la sensibilidad del modelo frente a entornos de producción heterogéneos, que son frecuentes en la salmonicultura chilena.

- Ayuda a mitigar errores en los extremos, que son epidemiológicamente relevantes, además de ayudar a balancear la atención del modelo hacia subgrupos menos frecuentes pero epidemiológicamente importantes.

- Al identificar clústeres con condiciones similares, se puede interpretar mejor qué variables afectan a cada grupo, lo cual es especialmente útil para tomar decisiones sanitarias diferenciadas por zona.

Limitaciones persistentes:

- Aunque el rendimiento mejora en segmentos críticos, las métricas globales (como R^2 o RMSE) solo muestran mejoras marginales. Esto sugiere que existen factores relevantes no modelados que podrían ser relevantes para mejorar el modelo.

- El uso de K-Means obliga a fijar un número “k” de clústeres, ya que el dataset consiste en más de 150.000 datos y se necesita buena precisión.

- Aunque el modelo segmentado mejora la precisión local, incrementa la complejidad operacional, ya que ahora hay múltiples modelos que mantener y evaluar.

Conclusión crítica de Resultados y Limitaciones

La integración de clustering previo al modelado representa un avance metodológico que, aunque no transforma drásticamente el desempeño global del modelo, mejora su capacidad de adaptación a subcontextos biológicos y ambientales diferenciados. Esto lo vuelve más útil para una eventual implementación práctica, como la generación de alertas o diagnósticos específicos por región o tipo de cultivo. A futuro, la incorporación de nuevas variables (como densidad de cultivo, tipo de tratamiento aplicado, u otros indicadores inmunológicos) podría mejorar aún más la precisión y utilidad del modelo.

Posibles líneas de trabajo

Incorporación de variables productivas y de manejo

Actualmente, el modelo utiliza variables ambientales y biológicas, pero no incluye información clave como:

- Densidad de cultivo por centro
- Tipo y frecuencia de tratamientos antiparasitarios aplicados
- Edad o peso promedio del pez en el ciclo productivo

Estas variables podrían mejorar sustancialmente la capacidad predictiva, especialmente en centros con manejo intensivo o tratamientos fallidos.

2. Implementación de un sistema de alerta temprana basado en el modelo

Utilizar las predicciones de hembras ovígeras para establecer umbrales críticos de riesgo sanitario (ej. >3 hembras/pez) y generar alertas automáticas. Esto requeriría acoplar el modelo a un sistema de monitoreo en tiempo real con sensores de temperatura, salinidad y carga parasitaria semanal.

3. Evaluación y ajuste del modelo por estacionalidad

Aunque se incluye la semana del año como variable, una línea futura sería entrenar modelos separados por estación (verano, otoño, invierno, primavera) o al menos analizar el desempeño temporalmente, ya que los ciclos de *Caligus* están fuertemente influenciados por las condiciones estacionales.

4. Validación cruzada por región o año

Para robustecer la generalización del modelo, se podría hacer una validación cruzada más estricta:

- Entrenar con datos de ciertas regiones (ej. Los Lagos) y testear en otras (ej. Aysén).
- Entrenar con datos de ciertos años (ej. 2012–2020) y testear en años recientes (2021–2024).

Esto simulará el rendimiento del modelo en contextos nuevos o con condiciones cambiantes.

5. Comparación de segmentación con otros métodos

Si bien K-Means fue eficiente, podrías explorar otras técnicas de clustering o reducción de dimensionalidad:

- **HDBSCAN**: para identificar clústeres con formas arbitrarias (una alternativa a DBSCAN más escalable).

- **t-SNE o UMAP** + clustering para separar datos en base a relaciones no lineales complejas.
  Esto podría descubrir patrones no capturados por K-Means y permitir modelos aún más especializados.

Conclusión

El presente estudio permitió la construcción de un modelo predictivo avanzado para estimar la carga parasitaria de *Caligus* spp. en salmónidos de cultivo, utilizando como variable objetivo el promedio de hembras ovígeras. A partir de un conjunto de datos extensivo (2012–2024), se integraron variables ambientales, biológicas, espaciales y temporales para capturar las complejas interacciones que determinan la dinámica de infestación en sistemas acuícolas intensivos.

El modelo base, construido mediante el algoritmo XGBoost, alcanzó un desempeño notable ($R^2 = 0.825$), validando su capacidad para modelar fenómenos no lineales y con interacciones múltiples. La utilización de este tipo de regresores ensamblados resulta especialmente pertinente en contextos donde los efectos no son aditivos ni independientes, como ocurre con enfermedades multifactoriales como la caligidosis.

La posterior implementación de un enfoque de segmentación previa utilizando K-Means introdujo una capa adicional de refinamiento metodológico. Si bien las mejoras globales en métricas de desempeño fueron moderadas ($R^2 = 0.827$; RMSE reducido en 0.008), la segmentación reveló patrones estructurales relevantes dentro del sistema productivo. En particular, los clústeres con mayores cargas parasitarias se asociaron a condiciones de temperatura y salinidad elevadas, lo que confirma hipótesis previas sobre la influencia ambiental en la proliferación de este ectoparásito.

Más allá de su rendimiento cuantitativo, el modelo segmentado exhibe ventajas cualitativas significativas: permite entrenar predictores especializados según perfiles ecológicos o productivos, facilita interpretaciones localizadas para la toma de decisiones sanitarias, y mejora la sensibilidad del sistema predictivo ante eventos críticos de alta carga, los cuales representan un riesgo epidemiológico considerable.

No obstante, el análisis también evidenció limitaciones importantes. La relativa estabilidad del error residual sugiere que existen determinantes de la infestación no incluidos en el modelo, tales como historial farmacológico, presión parasitaria de centros cercanos, o cambios en la gestión del cultivo. Asimismo, el uso de múltiples modelos por clúster introduce desafíos operativos y computacionales que deben ser considerados al pensar en una futura implementación industrial o institucional.

Finalmente, el trabajo refuerza la idea de que la carga parasitaria no puede entenderse ni modelarse de manera uniforme en la salmonicultura chilena. La heterogeneidad geográfica, climática y biológica del país exige enfoques adaptativos, dinámicos y basados en evidencia, donde la inteligencia artificial puede jugar un rol clave para avanzar hacia una sanidad acuícola más predictiva y preventiva.

Bibliografía:
(1)https://www.sernapesca.cl/app/uploads/2023/10/informe_sanitario_salmonicultura_en_centros_marinos_1_semestre_2021.pdf?utm (Capitulo 4)
(2)ttps://www.sernapesca.cl/app/uploads/2025/03/Informe-Sanitario-PRIMER-SEMESTRE-2024.pdf (Capitulo 5)
(3)https://www.salmonexpert.cl/idi/variables-productivas-serian-claves-en-la-dinamica-de-infeccion-de-caligus/1324743
(4)https://www.salmonexpert.cl/caligus-cohabitacion-idi/definen-niveles-de-transmision-de-caligus-entre-salmon-atlantico-y-robalo/1209730?utm
(5)https://www.salmonexpert.cl/caligus-piojo-piojos/revelan-asociacion-entre-factores-ambientales-y-expresion-genetica-durante-la-caligidosis/1873400?utm

Anexos de graficos

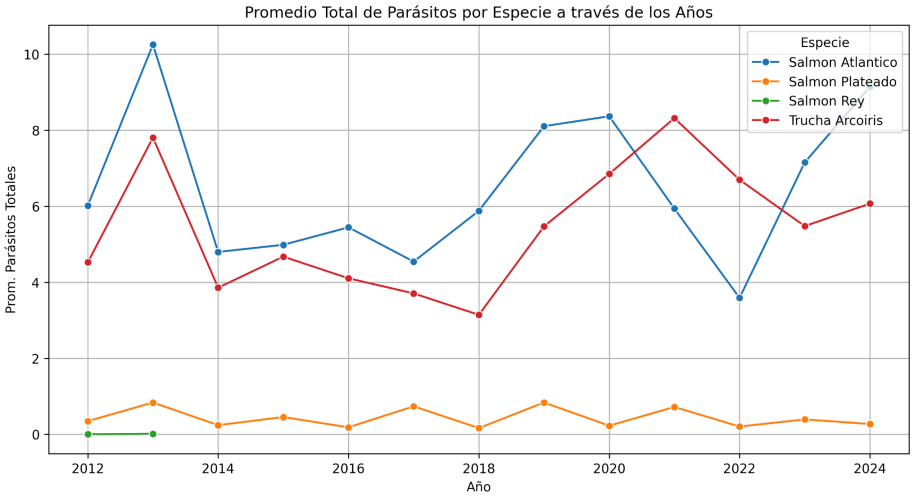


Imagen Anexo 1: Promedio Total de Parásitos

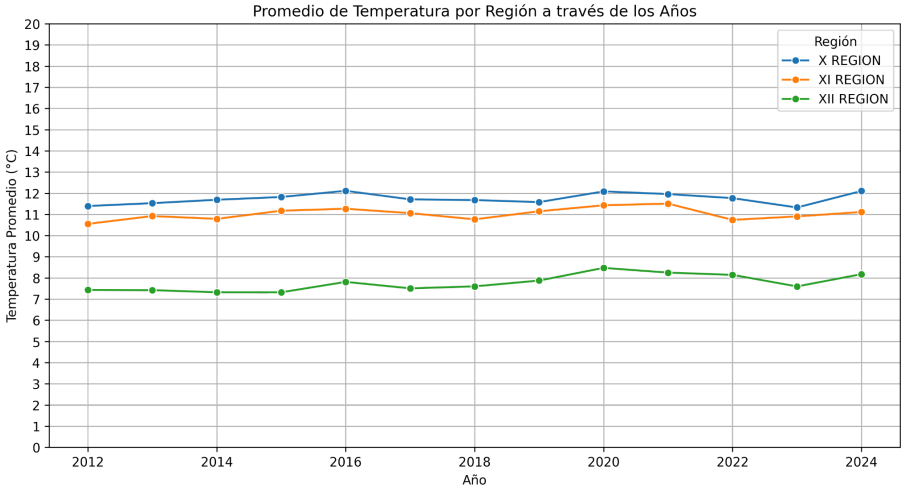


Imagen Anexo 2: Promedio de Temperatura

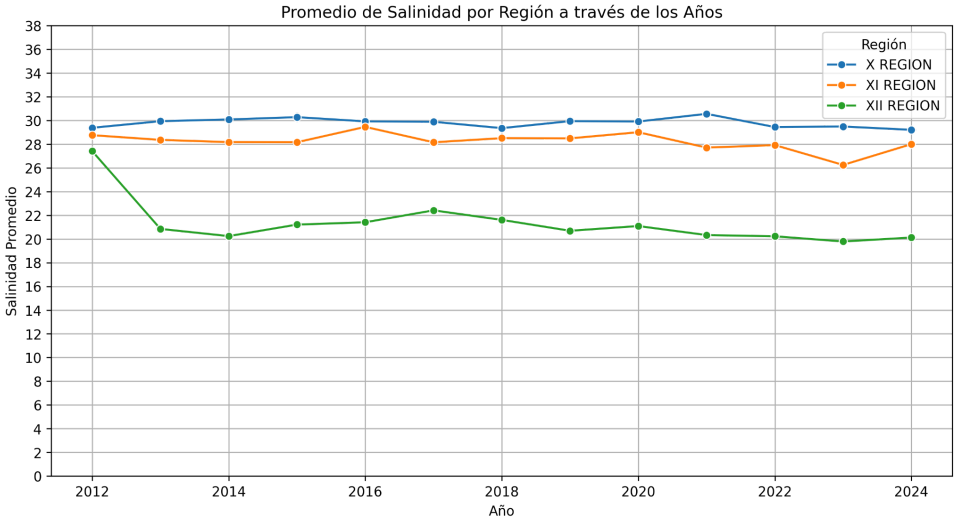
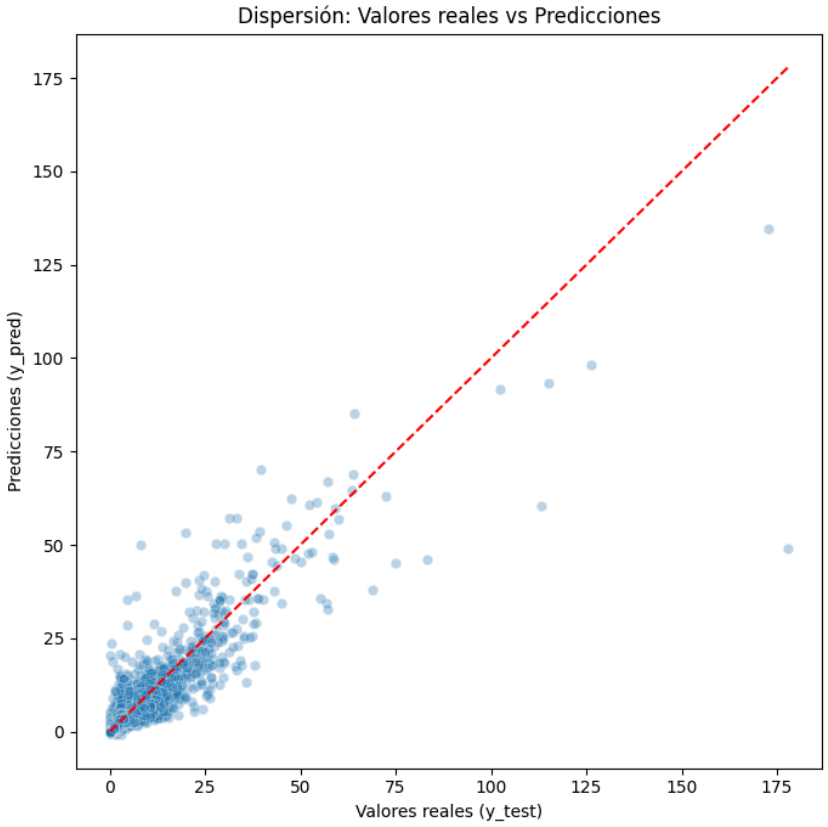


Imagen Anexo 3: Promedio de Salinidad



Imagen

Anexo 4: Dispersión del Modelo XGBoost

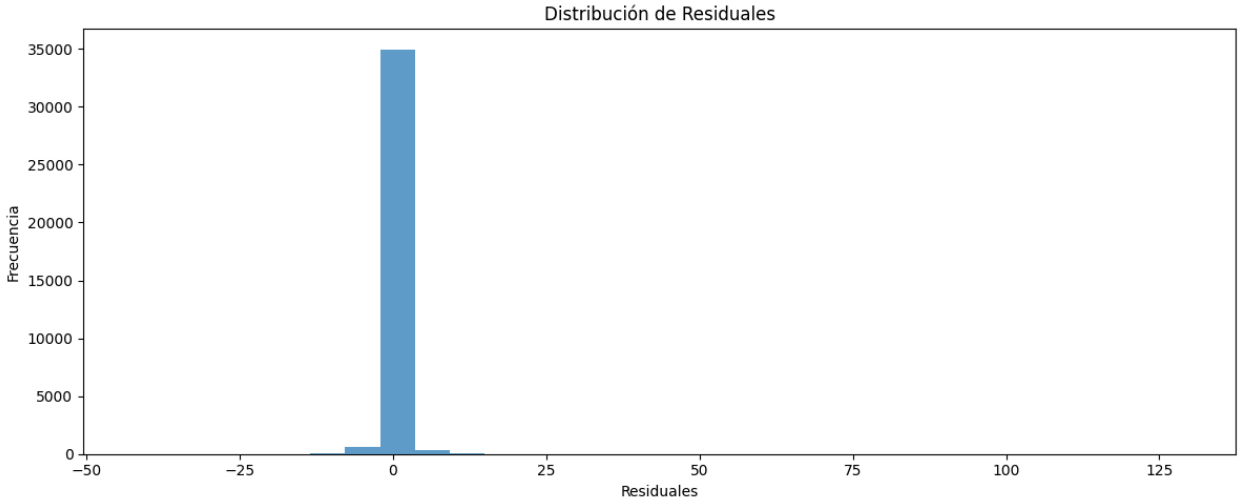


Imagen Anexo 5: Distribución de residuos del modelo XGBoost

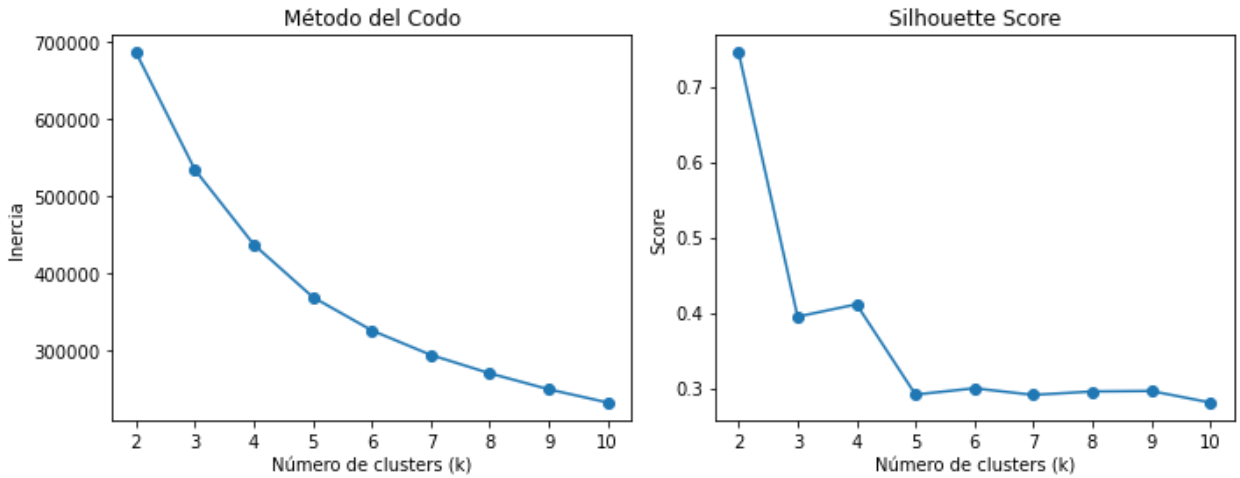
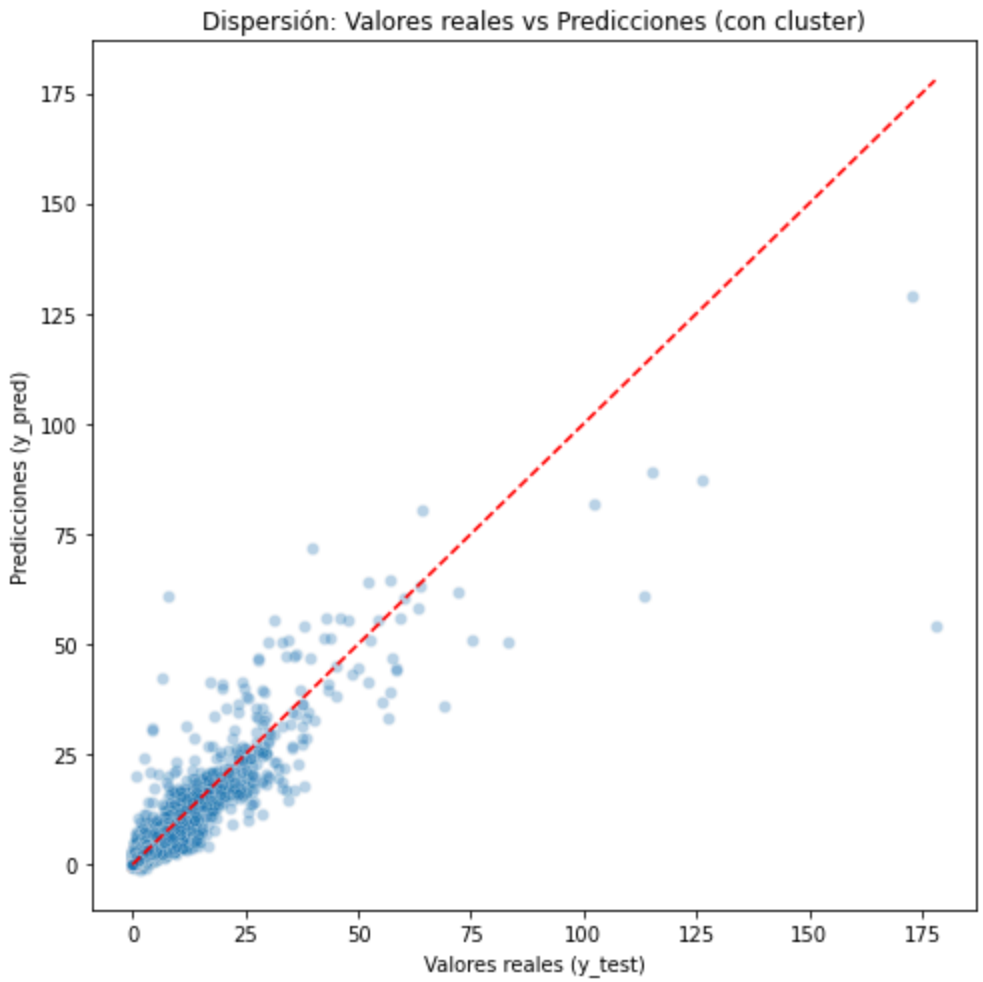


Imagen Anexo 5: Silhouette score y metodo del codo para seleccionar los valores óptimos de k en k-means



Anexo 7: Dispersión del Modelo XGBoost usando k-means con k=3

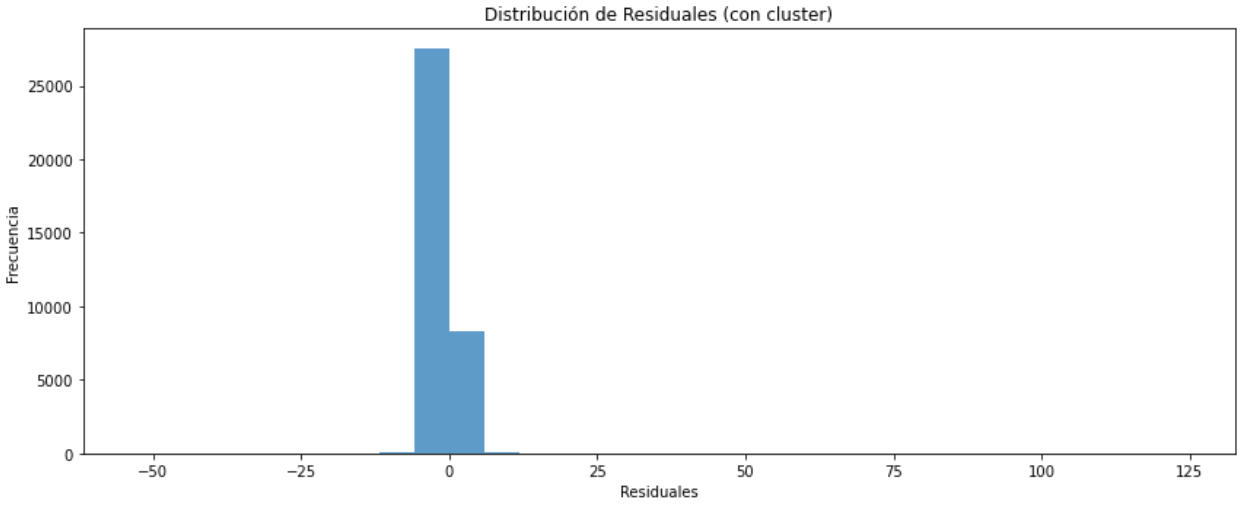


Imagen Anexo 8: Distribución de residuos del modelo XGBoost usando k-means con k=3