Memoria Final - Proyecto de Machine Learning en la Calidad del Agua

# 1. Introducción

El presente proyecto tiene como objetivo analizar la calidad del agua superficial utilizando técnicas de Machine Learning (ML) y Ciencia de Datos. A través de diferentes métodos (clasificación, regresión, series temporales y clustering), se exploran patrones y relaciones en datos hidrológicos reales provenientes de distintas regiones del mundo.

# 2. Primeras pruebas con datasets alternativos

Inicialmente, se realizaron pruebas exploratorias con el dataset de potabilidad del agua disponible en Kaggle ("Water Potability"). Aunque útil desde el punto de vista didáctico, su naturaleza sintética y limitada impidió un análisis profundo y contextualizado. <https://www.kaggle.com/datasets/adityakadiwal/water-potability>

# 3. Adopción de un dataset real y global

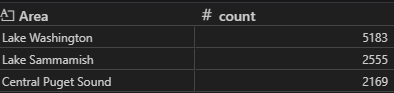
Posteriormente, se descubrió un recurso mucho más completo: “A Comprehensive Surface Water Quality Monitoring Dataset (1940–2023)”, el cual contiene más de 2.8 millones de registros de calidad de agua superficial de cinco países (EE.UU., China, Inglaterra, Irlanda, Canadá). <https://figshare.com/articles/dataset/A_Comprehensive_Surface_Water_Quality_Monitoring_Dataset_1940-2023_2_82Million_Record_Resource_for_Empirical_and_ML-Based_Research/27800394>

# 4. Problemas con el dataset combinado

Se intentó inicialmente trabajar con el archivo Combined\_data.csv, que reúne los datos de todos los países. Sin embargo, al ejecutar los primeros modelos de predicción, se comprobó que el tiempo de procesamiento era excesivo, dificultando la iteración rápida y efectiva. Por ello, se optó por analizar cada país de forma separada y al final elegir Inglaterra por temas prácticos.

# 5. Caso especial: Estados Unidos - Región de Seattle

Durante el análisis de EE.UU., se eligió la región de Seattle como caso de estudio específico. En particular, se investigaron las áreas: Lake Washington, Lake Sammamish y Central Puget Sound. Históricamente, estas zonas han enfrentado problemas de contaminación, pero en la actualidad muestran valores muy altos de calidad (CCME Values).

A screenshot of a black screen

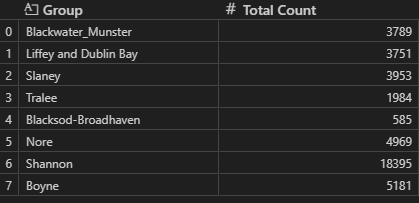
AI-generated content may be incorrect.

A map of the seattle area

AI-generated content may be incorrect.

# 6. Agrupación por palabras clave: Irlanda

Otra línea de investigación consistió en generar agrupaciones manuales por áreas geográficas usando palabras clave (por ejemplo, en el notebook ireland\_sample100). Se aislaron grupos relevantes como Slaney, Blackwater\_Munster, Liffey and Dublin Bay, Tralee, etc., para análisis focalizado.



# 7. Análisis por tipo de cuerpo de agua: Inglaterra

Se ensayaron segmentaciones según el tipo de cuerpo de agua (Waterbody Type): Lake, River, Effluent, Sewage, entre otros. Esta aproximación podría usarse para modelar comportamientos diferenciados y detectar zonas críticas pero, todavía, no se ha completado el análisis.

A black rectangular object with white text

AI-generated content may be incorrect.

# 8. Modelado aplicado en todos los países

En casi todos los datasets nacionales se aplicaron los siguientes métodos:

- EDA  
- Clasificación binaria (agua segura vs. no segura)  
- Clasificación multiclase (Poor, Marginal, Fair, Good, Excellent)  
- Regresión (predicción continua del índice CCME)  
- Series temporales (modelos SARIMA)  
- Clustering (KMeans, visualizaciones por spider plots)  
\*Aunque se realizaron pruebas para varios países, Inglaterra fue seleccionada como el país principal para la presentación final, debido a su riqueza en datos y diversidad de resultados. A screenshot of a graph

AI-generated content may be incorrect.

A chart with different colored squares

AI-generated content may be incorrect.

# 9. Regresión para predicción de calidad: Inglaterra

Se obtuvo un coeficiente R² de 0.999 en la predicción del índice CCME, lo que indica un excelente ajuste del modelo a los datos.

A screenshot of a black screen

AI-generated content may be incorrect.

# 10. Comparación: China vs. Inglaterra (Feature Importance)

Se compararon las importancias de variables del modelo XGBoost entre China e Inglaterra. En Inglaterra dominan los ortofosfatos, mientras que en China es clave la temperatura del agua, lo cual me impresionó porque en todos los demás dataset estaba ganando, con diferencia, en importancia, los ortofosfatos.

A graph with blue bars

AI-generated content may be incorrect.

A graph with blue bars

AI-generated content may be incorrect.

# 11. Clasificación multiclase y matriz de confusión: Inglaterra

El modelo multiclase alcanzó un recall y F1-score casi perfectos. La matriz de confusión muestra errores mínimos entre clases contiguas.

A screenshot of a computer

AI-generated content may be incorrect.

A graph of blue squares with white text

AI-generated content may be incorrect.

# 12. Clustering en China sin diferencias

A pesar de generar agrupaciones en China, los resultados no mostraron separación clara entre calidades de agua. Por este motivo se descartó como caso de estudio principal.

A black screen with white text

AI-generated content may be incorrect.

# 13. Clustering en Inglaterra y Spider Plot

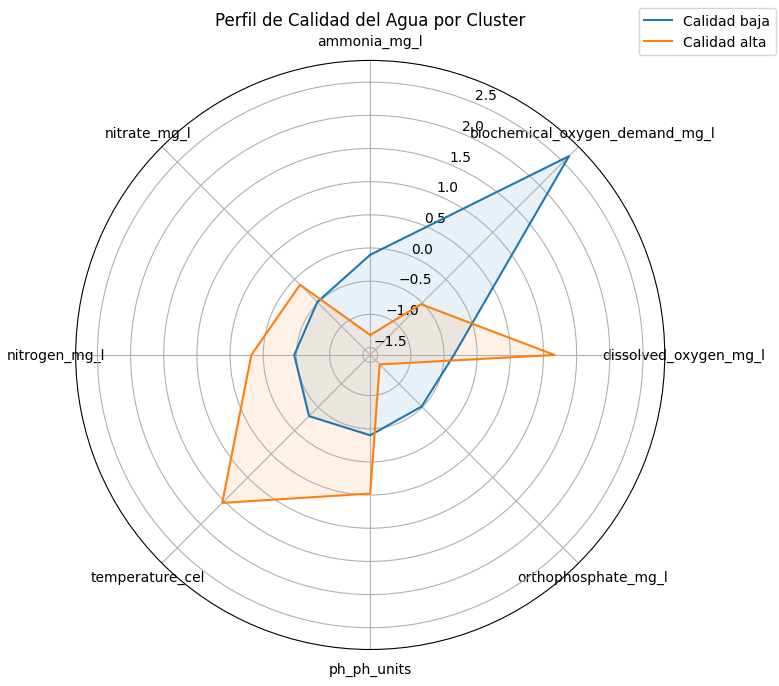
El análisis de clusters para Inglaterra reveló dos grupos claramente diferenciados en cuanto a calidad del agua. Se construyó un Spider Plot para visualizar el perfil químico medio por grupo.

A screen shot of a computer

AI-generated content may be incorrect.

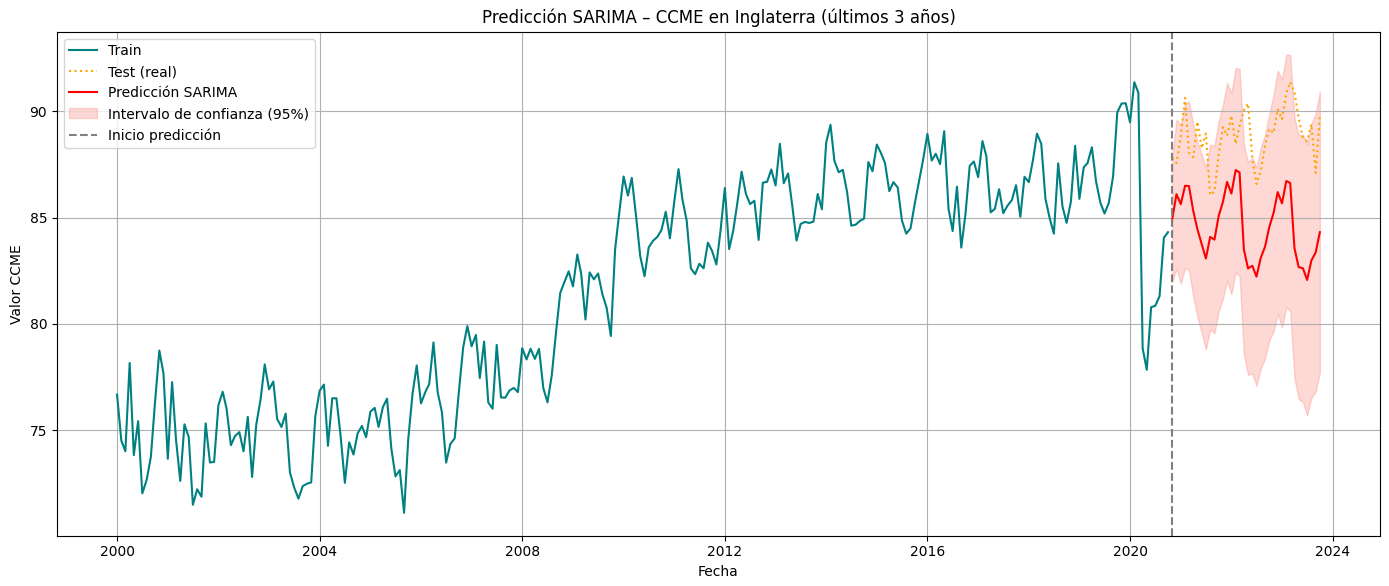
A black background with white text

AI-generated content may be incorrect.



# 14. Series temporales - Inglaterra

Se entrenó un modelo SARIMA para los últimos tres años en Inglaterra. La predicción muestra patrones estacionales definidos y una ligera tendencia positiva.

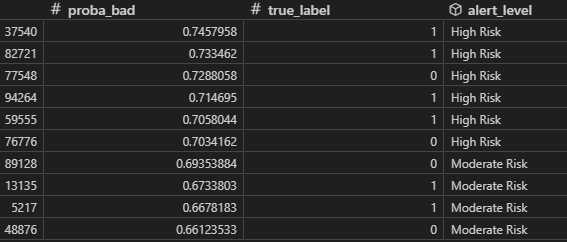


Con Resultados:  
A black background with white text

AI-generated content may be incorrect.

# 15. Sistema de alerta de riesgo

Se ensayó también un sistema de alerta temprana, categorizando zonas en High Risk, Moderate Risk y seguras, usando probabilidades de un modelo binario.



# 16. Ideas para el futuro

- Usar el combined\_data.csv completo con mejor infraestructura de cómputo para detectar agrupaciones globales por país o tipo de contaminación.  
- Aplicar modelos de series temporales diferenciados por tipo de cuerpo de agua o región administrativa (distritos).  
- Desarrollar dashboards interactivos para monitoreo en tiempo real y modelos actualizables.

- Aplicar técnicas geoespaciales y mapas interactivos

# Conclusión

Este proyecto no solo demuestra la aplicabilidad del Machine Learning en la calidad del agua, sino también destaca el valor de los datos abiertos y bien estructurados. El enfoque exploratorio por países y regiones permitió obtener hallazgos relevantes para la gestión y preservación de recursos hídricos.

**Autor: [Angelos Ampatzidis]**  
**Proyecto: Machine Learning aplicado a calidad del agua (Inglaterra)**  
**Fecha: [20/05/2025]**