

Entrega 3: Presentación del Modelo y Análisis de Resultados

Fuente de datos

Los datos utilizados provienen del monitoreo realizadas entre 2013 y 2024 por el Ministerio de Ambiente y Desarrollo Sostenible de Argentina. Se recolectaron muestras a lo largo de la costanera del Río de la Plata y los datos fueron obtenidos de plataformas públicas como Kaggle y CiAM.

El dataset contiene 1182 registros y 30 variables.

<https://www.kaggle.com/datasets/palomachiacchiara/muestreos-de-calidad-de-agua-de-la-riiglo/data>

<https://ciam.ambiente.gob.ar/repositorio.php?tid=1&stid=105&did=408#>

Análisis Exploratorio de Datos

Se realizaron múltiples análisis y visualizaciones:

- Distribución de la variable objetivo, La clase "extremadamente deteriorada" fue la más frecuente, evidenciando un desbalance de clases.
- Matriz de correlación, Reveló fuertes relaciones entre indicadores microbiológicos y la calidad del agua.
- Gráficos de dispersión y violin plots, Permitieron visualizar la relación entre variables como temperatura, OD, pH y la calidad del agua.
- Tratamiento de outliers, Se aplicó el método del rango intercuartílico (IQR) para suavizar valores extremos.
- Reducción de dimensionalidad, Se utilizó PCA para conservar el 95% de la varianza con 25 componentes principales.

Preguntas de Investigación o Hipótesis

Vamos a responder las preguntas de Investigación o Hipótesis, que nos planteamos al inicio del proyecto, basándonos en el análisis exploratorio, la construcción de modelos y los resultados obtenidos.

1. ¿Cuáles son las variables que más influyen en la clasificación de la calidad del agua?

Realizamos observación al análisis de correlación y los modelos aplicados, las variables que más influyen en la clasificación de la calidad del agua.

- Indicadores microbiológicos, Coliformes fecales, Escherichia coli y Enterococos mostraron alta correlación entre sí y con la variable objetivo, lo que indica su fuerte relación con la contaminación fecal.

Alumno: Diego Estrada

- Parámetros físico-químicos, Oxígeno disuelto (OD) y pH tienen correlaciones significativas con el Índice de Calidad del Agua (ICA) y la variable de calidad. Nitrato, amonio (NH_4) y fósforo total también se destacan por su relación con procesos de eutrofización.
- Otros factores relevantes, Turbidez, microcistinas y DQO (Demanda Química de Oxígeno) también mostraron correlaciones negativas con la calidad del agua, indicando deterioro.

Estas variables fueron fundamentales para la construcción de los modelos predictivos y conservaron su relevancia incluso después de aplicar técnicas de reducción de dimensionalidad como PCA.

2. ¿Existen diferencias significativas en la calidad del agua entre estaciones del año o zonas geográficas?

Si observamos el análisis exploratorio, se puede afirmar que sí existen diferencias estacionales en la calidad del agua.

- Las estaciones del año fueron codificadas como variables binarias (invierno, otoño, primavera, verano).
- La matriz de correlación reveló asociaciones entre ciertas estaciones y la calidad del agua.
- La estación de verano mostró una correlación positiva con la temperatura del agua y del aire, lo que podría favorecer la proliferación de microorganismos.
- La estación de invierno presentó una correlación negativa con la temperatura y una leve asociación con mejores condiciones de calidad del agua.

3. ¿Es posible predecir la categoría de calidad del agua utilizando modelos de aprendizaje supervisado?

En este proyecto se implementaron y compararon tres modelos de clasificación supervisada.

- Random Forest (Accuracy en validación cruzada, 88.20%)
- K-Nearest Neighbors (Accuracy en validación cruzada: 85.72%)
- Red Neuronal (Accuracy en validación cruzada: 92.58%)

Al aplicar técnicas de optimización de hiperparámetros y regularización, el modelo MLP fue el que obtuvo el mejor desempeño. Resultado.

92.58% de accuracy en validación cruzada.

83.12% de accuracy en el conjunto de prueba.

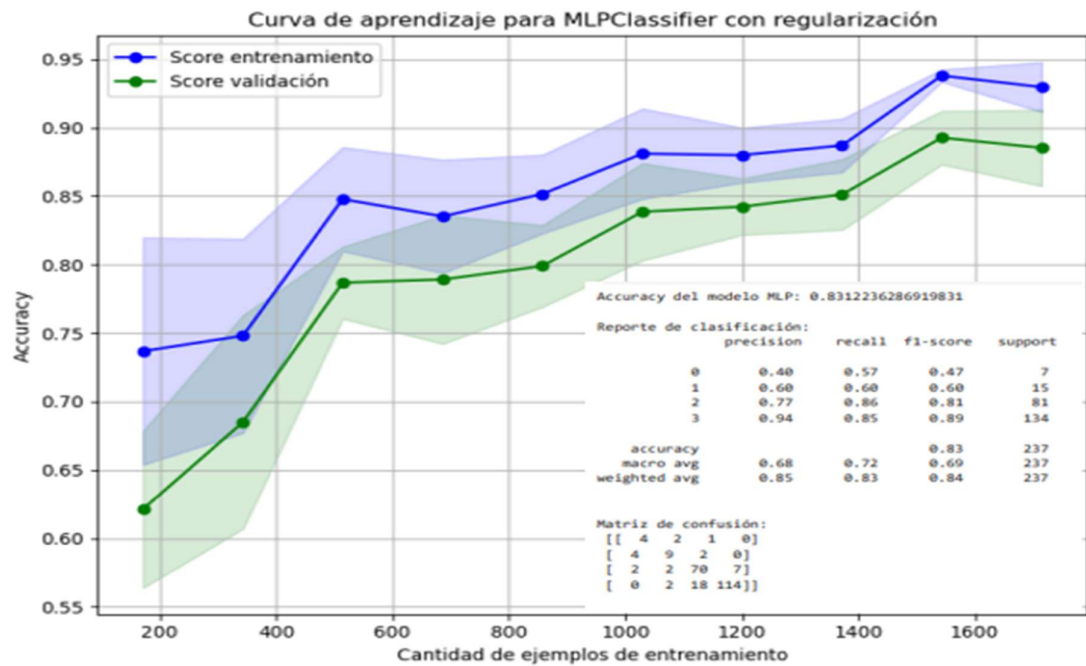
Excelente equilibrio entre precisión y recall, incluso en clases minoritarias.

Curva de aprendizaje estable, sin indicios significativos de sobreajuste.

Alumno: Diego Estrada

Métricas de Evaluación del Modelo MLP

Clase	Precisión	Recall	F1-score
Levemente deteriorada (0)	0.40	0.57	0.47
Deteriorada (1)	0.60	0.60	0.60
Muy deteriorada (2)	0.77	0.86	0.81
Extremadamente deteriorada (3)	0.94	0.85	0.89
Accuracy general			83.12%



Conclusión.

El modelo Red Neuronal MLP optimizado demostró ser la mejor alternativa para predecir la calidad del agua. Su rendimiento fue sólido, equilibrado y con buena capacidad de generalización, lo que lo convierte en una herramienta confiable para su aplicación en otras regiones, como Tierra del Fuego.