INGENIAS - EQUIPO N°18

DATASET | ANÁLISIS EXPLORATORIO | MACHINE LEARNING

AGUA POTABLE

CLASIFICACIÓN SEGÚN VARIABLES

GUADALUPE LESCANO - MICAELA KOROL

¿POR QUÉ ES IMPORTANTE ANALIZAR LA POTABILIDAD DEL AGUA?

El acceso a agua potable es un derecho humano esencial, pero no siempre está garantizado en todas las regiones.
En muchas zonas rurales o con infraestructura limitada, los análisis físico-químicos no se realizan con frecuencia, ya sea por falta de recursos, equipamiento o personal técnico.

Evaluar la calidad del agua permite prevenir enfermedades, mejorar la salud pública y garantizar el bienestar de comunidades enteras.
Por eso, desarrollar herramientas que permitan monitorear y anticipar riesgos en la potabilidad del agua se vuelve fundamental.

OBJETIVO SOBRE EL DATASET

DISEÑAR UN MODELO PREDICTIVO QUE PERMITA CLASIFICAR MUESTRAS DE AGUA COMO POTABLES O NO POTABLES, A PARTIR DE PARÁMETROS FÍSICO-QUÍMICOS MEDIDOS.

- Fuente: [Keagle – Water Quality Dataset]
- +600.000 registros únicos
- 18 variables numéricas físico-químicas
- Variables adicionales: color, olor, temperatura, fuente de obtención, mes
- Etiqueta objetivo: Target (1 = potable, 0 = no potable)

ANÁLISIS EXPLORATORIO

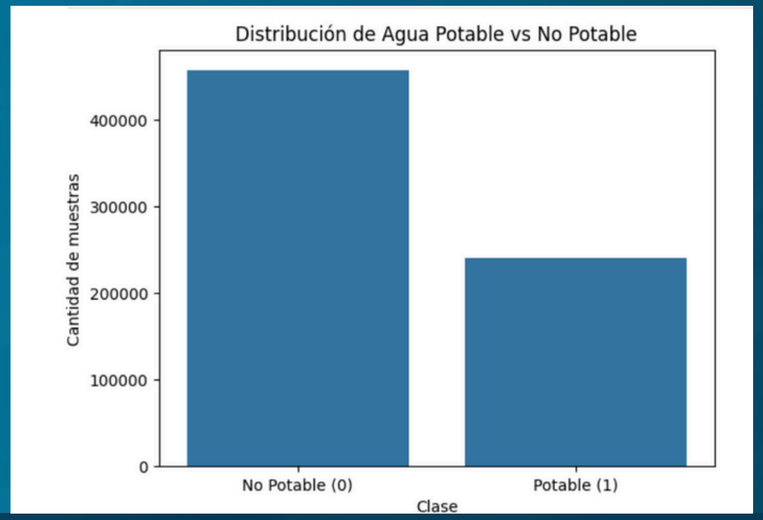
- Análisis de distribución por parámetro
- Proporción de muestras potables y no potables
- Correlaciones
- Outliers y valores faltantes tratados con imputación media y transformación logarítmica

EXPLORACIÓN Y LIMPIEZA DEL DATASET

- Renombramos las columnas para mejorar la legibilidad del dataset.
- Analizamos la distribución de cada columna para entender su comportamiento y detectar valores atípicos o inconsistencias.
- Detectamos un desbalance en la variable objetivo:

Agua no potable: **457.841 muestras**

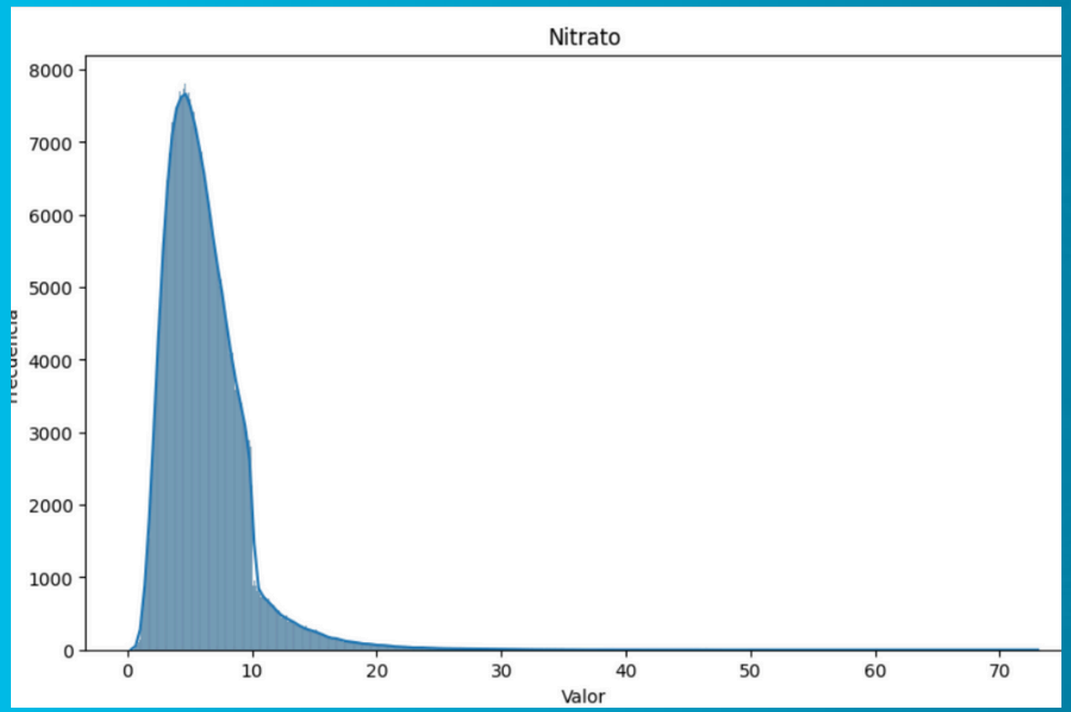
Agua potable: **240.734 muestras**

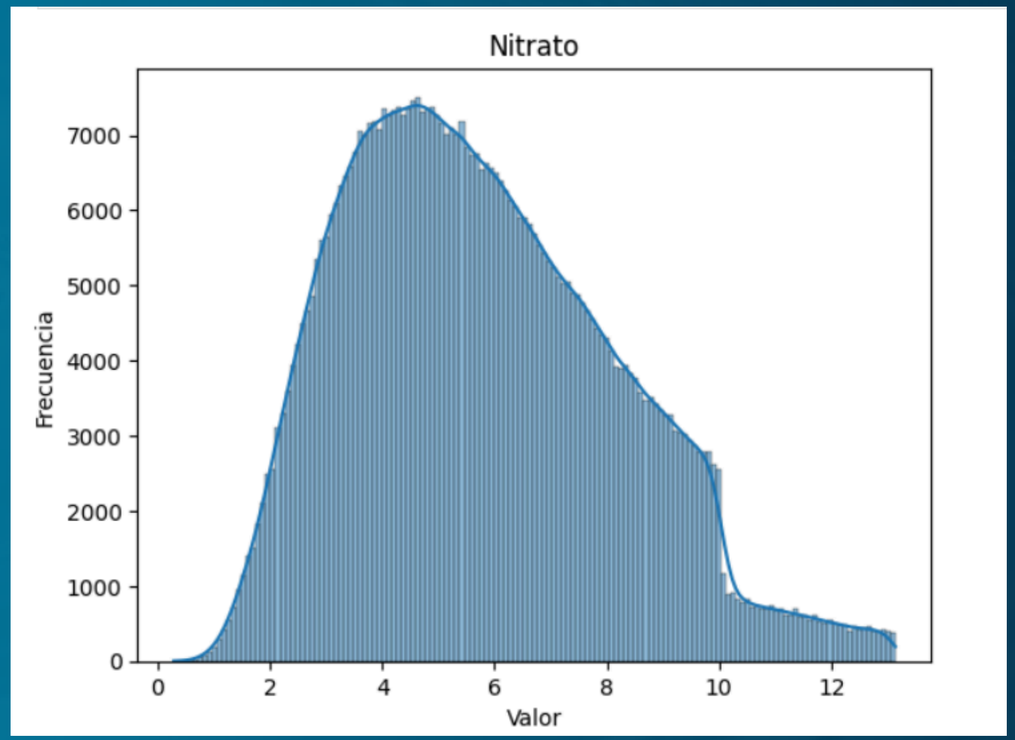


EXPLORACIÓN Y LIMPIEZA DEL DATASET

- Eliminamos outliers y valores nulos para mejorar la calidad del análisis y evitar que valores extremos distorsionen el resultado del modelo.
- Por ejemplo, en la variable Nitrato, se eliminaron ~17.370 filas con valores faltantes o considerados atípicos.
- Esto disminuyó la media (6.13 → 5.76) y también se redujo la dispersión de los datos, lo que significa que ahora están más concentrados cerca del promedio.

EXPLORACIÓN Y LIMPIEZA DEL DATASET





EXPLORACIÓN Y LIMPIEZA DEL DATASET - VARIABLES CATEGÓRICAS

Para que los modelos puedan trabajar con las variables no numéricas, realizamos estas transformaciones:

- Color y Mes:

Codificados como números ordenados según su lógica (Color_encoded, Mes_encoded).

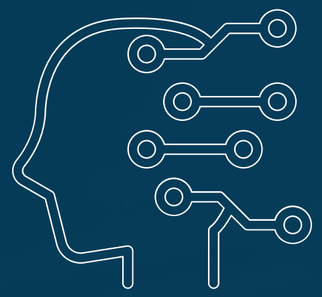
- Fuente de agua:

Aplicamos One Hot Encoding, generando una columna por tipo de fuente (Fuente_River, Fuente_Spring, etc.).

Así preparamos los datos sin perder información clave 🌊

TRANSFORMACIÓN Y ESTANDARIZACIÓN

- Varias variables presentaban valores en notación científica, lo que indicaba un alto sesgo y presencia de valores extremos.
- Por eso, aplicamos una *transformación logarítmica* para normalizar sus distribuciones.
- Por último aplicamos *StandardScaler* para estandarizar las variables numéricas, ayudando a que todas estén en la misma escala y mejorando el desempeño del modelo.



MODELO DE APRENDIZAJE SUPERVISADO

Se utilizó validación cruzada y ajuste de hiperparámetros para optimizar desempeño.
La métrica principal fue F1-score, por tratarse de clases desbalanceadas.

MODELOS EVALUADOS

- Árbol de Decisión
- Random Forest
- SVM (máquinas de soporte vectorial)

DEFINICIÓN Y AJUSTE DEL MODELO

- Definimos el modelo y usamos GridSearchCV para optimizar sus parámetros.
- Dado el desbalance de clases en el dataset, aplicamos **class_weight='balanced'** para asignar mayor peso a la clase minoritaria (agua potable).

EVALUACIÓN Y ANÁLISIS DE VARIABLES IMPORTANTES

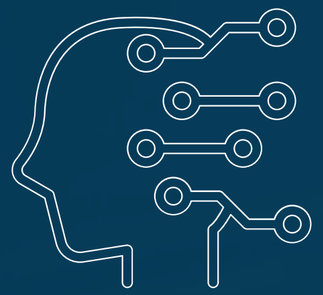
- Calculamos la importancia de las variables para identificar cuáles tienen mayor influencia en las decisiones del modelo.
- Usamos la matriz de confusión para evaluar el desempeño del modelo, **confirmando** que mejora la identificación del agua potable al aplicar class_weight.

🔎 **Impacto en el desempeño del modelo:**

- Se redujeron los Falsos Negativos (FN: 749 → 520)
- Aumentaron los Verdaderos Positivos (TP: 9536 → 9765)
- El modelo mejoró su capacidad para detectar correctamente el agua potable.

RESUMEN DE MODELOS EVALUADOS

- Random Forest obtuvo el mejor desempeño general, con un F1-score alto y recall cercano al 98% para detectar agua potable.
- Árbol de Decisión mostró buen equilibrio entre precisión y recall, con un F1-score de 0.84 para la clase potable.
- SVM logró un recall aceptable, pero con baja precisión, generando más falsos positivos.



MODELO DE APRENDIZAJE NO SUPERVISADO

Aplicamos técnicas de aprendizaje no supervisado para descubrir agrupamientos naturales en las muestras de agua, sin usar la variable objetivo (potable/no potable).
Esto permite identificar patrones ocultos, entender mejor la estructura del dataset y explorar si es posible separar clases sin información previa.

K-MEANS + PCA: EL AGRUPAMIENTO MÁS CLARO

Reducimos la dimensionalidad del dataset con PCA (2 componentes principales) y luego aplicamos K-Means.
Con k = 3, obtuvimos el mayor Silhouette Score (0.456), lo que indica una separación moderadamente clara entre grupos.

Esto sugiere que existen al menos 3 perfiles diferenciados de calidad de agua según sus parámetros físico-químicos.
✅ Este método fue el que mejor agrupó sin usar la etiqueta.

COMPARACIÓN DE MÉTODOS NO SUPERVISADOS

Probamos distintos algoritmos de clustering:

Modelo	Silhouette Score
KMeans	0.193 – 0.456 (con PCA)
DBSCAN	Muy bajo / no separa bien
GMM	0.073 – clusters poco definidos

CONCLUSIONES

- Es posible predecir con buen desempeño la potabilidad del agua usando parámetros físico-químicos.
- Los **Modelos Supervisados** permiten clasificar con precisión muestras sin necesidad de análisis de laboratorio costoso.
- El agrupamiento **No Supervisado** también permite identificar patrones y posibles riesgos.
- Esta herramienta puede ser útil para gobiernos, ONGs o comunidades con acceso limitado a laboratorios.

GRACIAS!

POR SU ATENCIÓN