Calidad del agua.

MODELOS DE PREDICCIÓN SOBRE LA POTABILIDAD DEL AGUA

Introducción

Para este proyecto se uso un dataset proveniente de Kaggle, https://www.kaggle.com/datasets/adityakadiwal/water-potability

Los datos se centran en investigaciones sobre el abastecimiento del agua, podemos ver entonces características como el PH, la dureza, turbidez, conductividad, presencia de solidos, de cloro, de sulfatos, etc, que finalmente determinan si el agua es potable o no.

La fuente consultada presenta rangos establecidos por WHO (World Health Organization) standars para cada una de las características, esta información nos ayudará a realizar un mejor análisis de los datos.

Diccionario de datos

El data set comprende 3276 filas y 10 columnas.

1. ph: PH del agua. (0 a 14)

2. Hardness: Capacidad del agua para precipitar Soap en mg/L.

3. Solids: Sólidos disueltos totales en ppm.

4. Chloramines: Cantidad de Cloraminas (cloro) en ppm.

5. Sulfate: Cantidad de Sulfatos disueltos en mg/L.

6. Conductivity: Conductividad eléctrica del agua en µS/cm.

7. Organic_carbon: Cantidad de carbono orgánico en ppm.

8. Trihalomethanes: Cantidad de Trihalometanos en µg/L.

9. Turbidity: Medida de la propiedad de emisión de luz del agua en NTU.

10. Potability: Indica si el agua es segura para el consumo humano. Potable 1 y No potable 0

Limpieza y tratamiento de los datos

Pasos:

1. validación de columnas duplicadas

2. Identificación y tratamiento de valores faltantes, se utilizo la estrategia Imputer con el promedio de los valores ya que no representaban un cantidad mayor.

3. Validación de estadísticos e identificación de posibles inconsistencias con los datos.

4. visuales univariantes para el objetivo y todas las características. Boxplots e histogramas.

5. Mapa de calor para identificar posibles correlaciones entre las características.

Análisis exploratorio, gráficos.

HISTOGRAMA DE LA COLUMNA OBJETIVO:

Casi el 61% porciento de los datos pertenece al agua no potable, y el 39% restante al agua que si es potable.

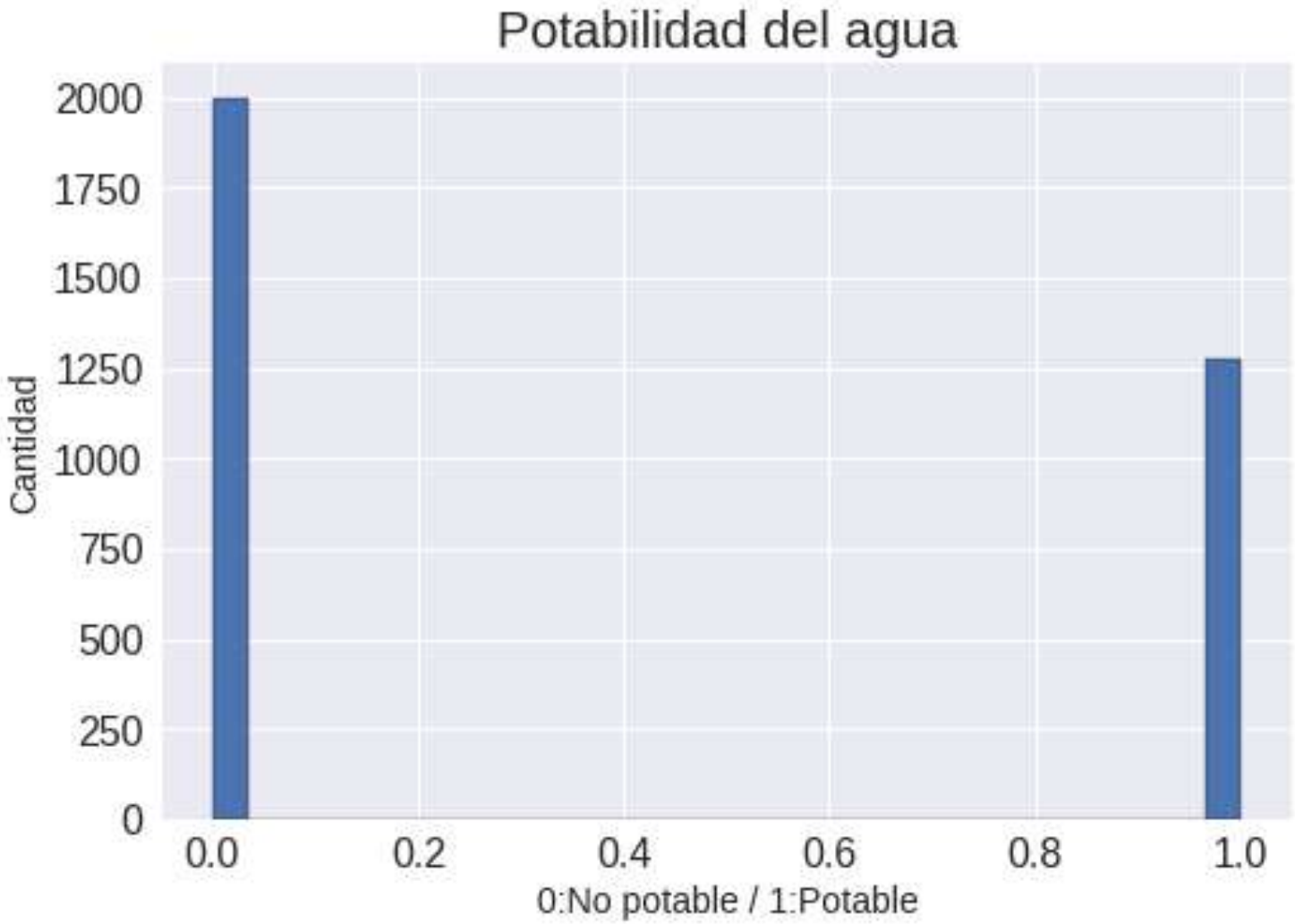
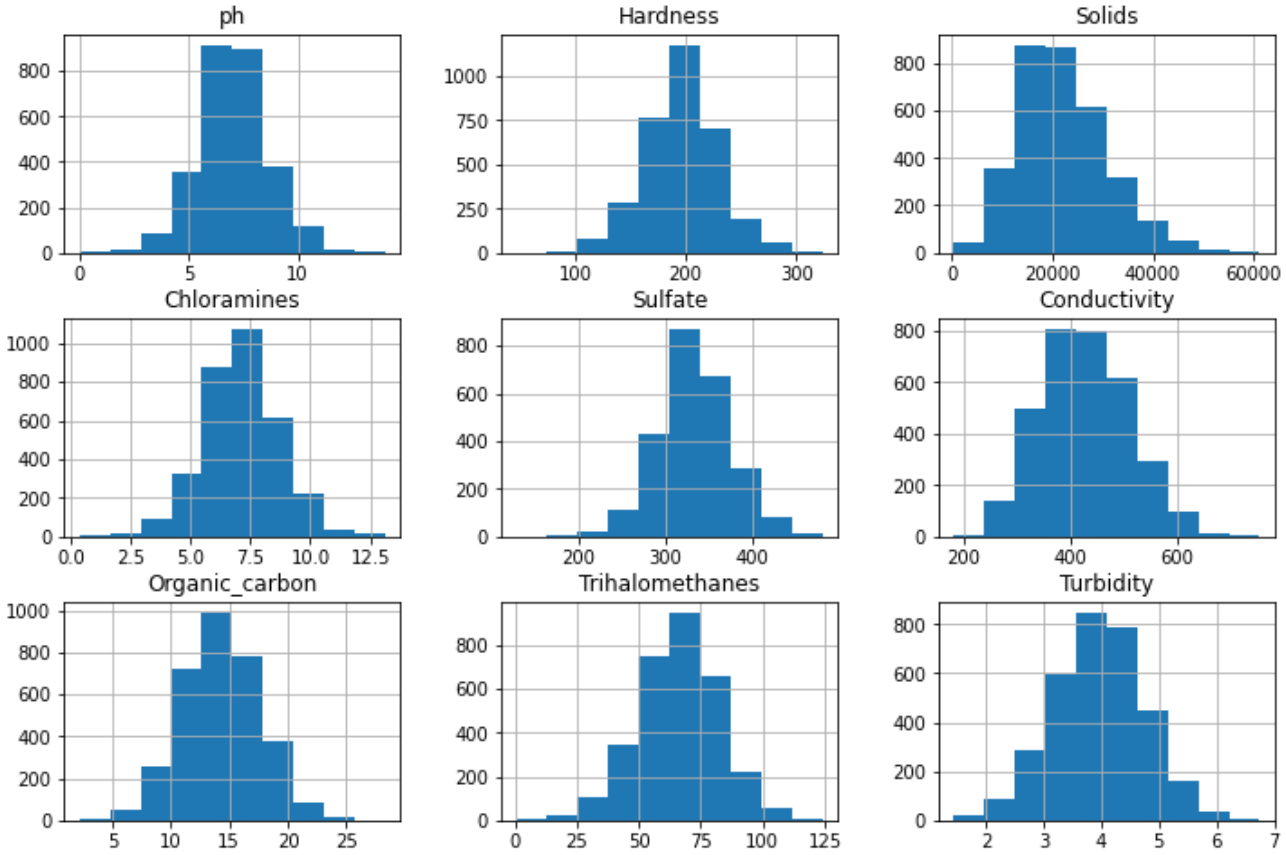


Gráfico univariante de todas las características, histogramas:



Según la fuente de la información, el rango de PH recomendado para la salud es de 6.5 a 8.5. En el gráfico 1 podemos observar que hay algunos valores por fuera de este rango, lo que puede afectar la potabilidad del agua.

La cantidad deseable de Solidos en el agua es de 500 mg/l y máximo de 1000 mg/l, el grafico 3 muestra gran cantidad de datos por fuera de este límite, afectando también la potabilidad del agua.

El nivel de cloro máximo debe ser de 4 mg/L, los datos en el grafico muestran un nivel elevado de cloro, lo mismo sucede con el carbono orgánico cuyo valor máximo debe ser de 4 mg/L.

La turbidez del agua muestra algunos valores por encima del nivel permitido, 5 NTU.

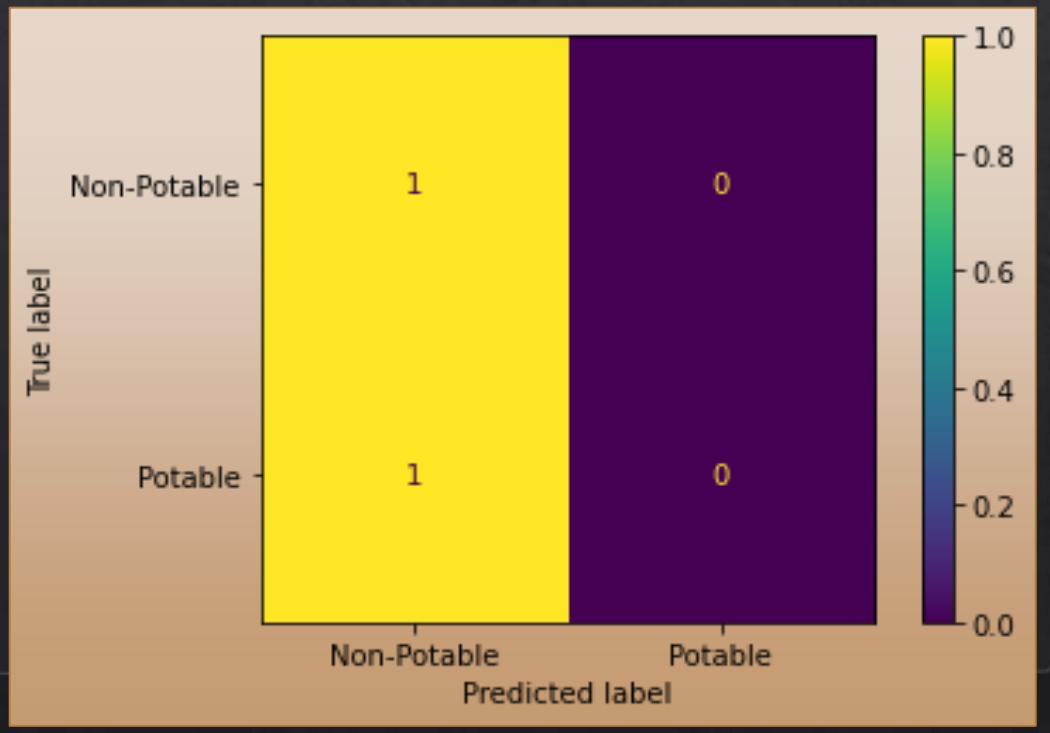
Problema de clasificación binaria

Se probaron 3 modelos de clasificación:

- **1. Regresión Logística**
- Afinando hiperparametros L1 y L2:
- accuracy de entrenamiento: 0.60
- accuracy de prueba: 0.62

	precision	recall	f1-score	support
0	0.62	1.00	0.77	510
1	0.00	0.00	0.00	309
accuracy			0.62	819

Matrix de confusion:

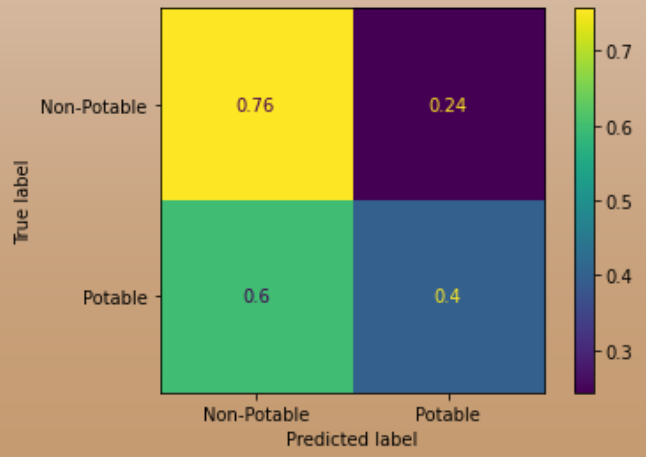


2. KNN (k vecinos más cercanos)

- Usando PCA (reducción de dimensionalidad):
- accuracy de entrenamiento: 0.76
- accuracy de prueba: 0.62

Classification Report for Testing Set

	precision	recall	f1-score	support
0	0.68	0.76	0.71	510
1	0.50	0.40	0.45	309
accuracy			0.62	819

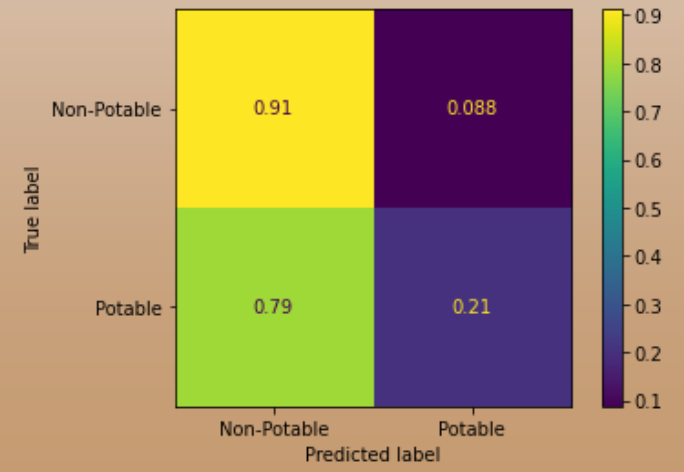


3. Arboles de decisión

- Usando PCA y afinando hiperparametros:
- accuracy de entrenamiento: 0.67
- accuracy de prueba: 0.65

Classification Report for Testing Set

	precision	recall	f1-score	support
0	0.65	0.91	0.76	510
1	0.59	0.21	0.31	309
accuracy			0.65	819



conclusiones

El mejor modelo en este ejercicio es el de arboles de decisión, ya que presenta menor cantidad de falsos positivos y predice mejor a los verdaderos negativos en comparación con el modelo de KNN y de árbol de decisión que parece presentar overfiting. Además es el que arroja mejor resultado de accuracy para los datos de testing.

Aplicar PCA ayudo a mejorar levemente las predicciones.

En el modelo de regresión logística se presento overfiting al tratar de mejorar el resultado afinando hiperparametros.

Gracias!

ANDREA ARBOLEDA