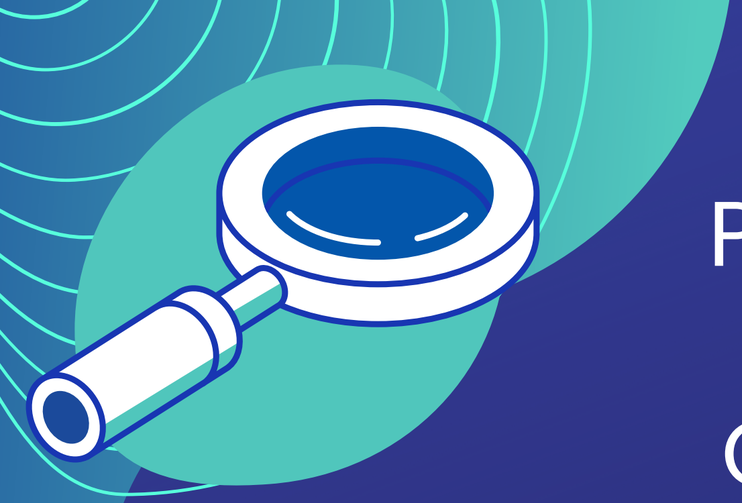


PRESENTACIÓN EJECUTIVA: ANÁLISIS Y MODELOS DE CLASIFICACIÓN DE AGUAS SUBTERRÁNEAS EN MÉXICO

CIENCIA Y ANALÍTICA DE DATOS

ROBERTO ROMERO VIELMA - A00822314

FRANCISCO MEDELLIN ZERTUCHE - A01794044

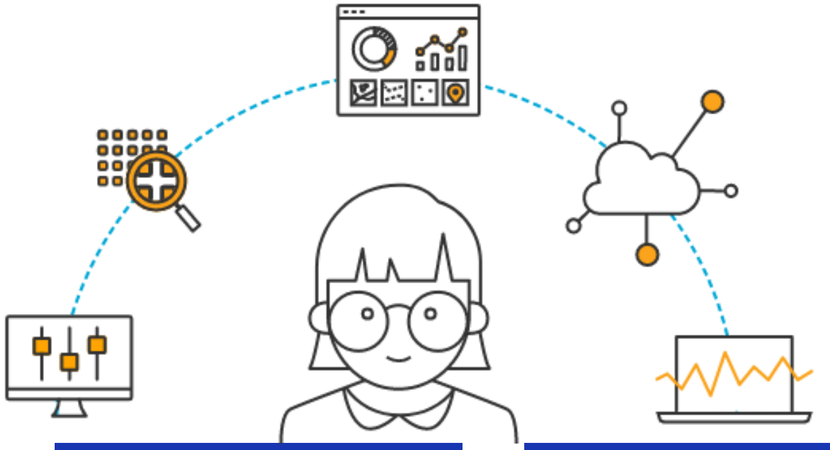
PROFESORA TITULAR: MARÍA DE LA PAZ RICO FERNÁNDEZ

PROFESOR TUTOR: ROBERTO ANTONIO GUEVARA GONZÁLEZ

FECHA DE ENTREGA: 18/11/2022

Pipeline

Este fue el pipeline seguido durante el reto



Raw Data

Importamos el archivo Datos_de_calidad_del_agua_de_sitios_de_monitoreo_de_aguas_subterraneas_2020.csv.

Hicimos uso de Pandas para poder importar los datos a nuestro notebook.

Data Exploration

Visualizamos que columnas estaban dentro del dataset, valores vacíos en cada columna y la distribución de los datos.

Para ello nos apoyamos de métodos como head(), isnull(), info(), hist(), entre otros.

Data Cleaning

Llenamos los valores faltantes, cambiamos el tipo de algunas columnas e hicimos dummy variables.

Los valores faltantes se llenaron con mediana y moda, algunas columnas tenían caracteres especiales que fueron eliminados, además de que la columna CONTAMINANTES fue convetida en dummy variables.

Modeling Data

Con la base de datos ya limpia, procedimos a usar un Decision Tree para obtener las features más importantes, y con esas features creamos nuevos modelos.

Las features más importantes fueron las relacionadas, a los indicadores de si el agua cumple o no con algún componente químico. En base a ello se desarrollaron 2 modelos, uno con Decision Tree y otro con Random Forest.

Model Evaluation

Medimos el desempeño de nuestros modelos con diferentes métricas, además de obtener la matriz de confusión y las curvas de aprendizaje.

Las métricas utilizadas fueron: Accuracy, precision y recall. De los scores obtenidos y de la visualización de las curvas de aprendizaje y la matriz de confusión se obtuvo que el modelo de Random Forest ofrece los mejores resultados.

Limpieza de datos

Conocer los tipos de datos en las columnas

La mayoría de las columnas son datos de tipo objeto, siendo variables de tipo ordinal o categórica. Solo 5 variables aportaban un valor numérico. Las cuales son: "CONDUCT_mS/cm","ALC_mg/L","SDT_M_mg/L","N_NO3_mg/L", "DUR_mg/L"

Eliminación de valores nulos

La base de datos presenta una gran cantidad de datos no nulos. Sin embargo las variables "SDT_mg/L" y "CONTAMINANTES" son las que mayor cantidad de datos nulos tienen. Siendo "SDT_mg/L" la que presenta una totalidad de valores nulos, por lo que fue eliminada.

Llenado de valores nulos

El número de columnas que presentan valores nulos es 13, en donde 11 de ellas contienen de 1 a 6 valores nulos. Aquí se hizo un análisis para comprender estas variables y conocer la forma óptima del tratamiento de nulos. Ya sea por media o moda.

Análisis

Variable objetivo

La distribución de los valores de la variable SEMAFORO se compone de la siguiente manera:

- Verde: 434 (40%)
- Rojo: 387 (36%)
- Amarillo: 247 (24%)

Desviación Estandar

En la mayoría de las columnas se presenta con valores cercanos a 0, indicando poca dispersión de los datos.

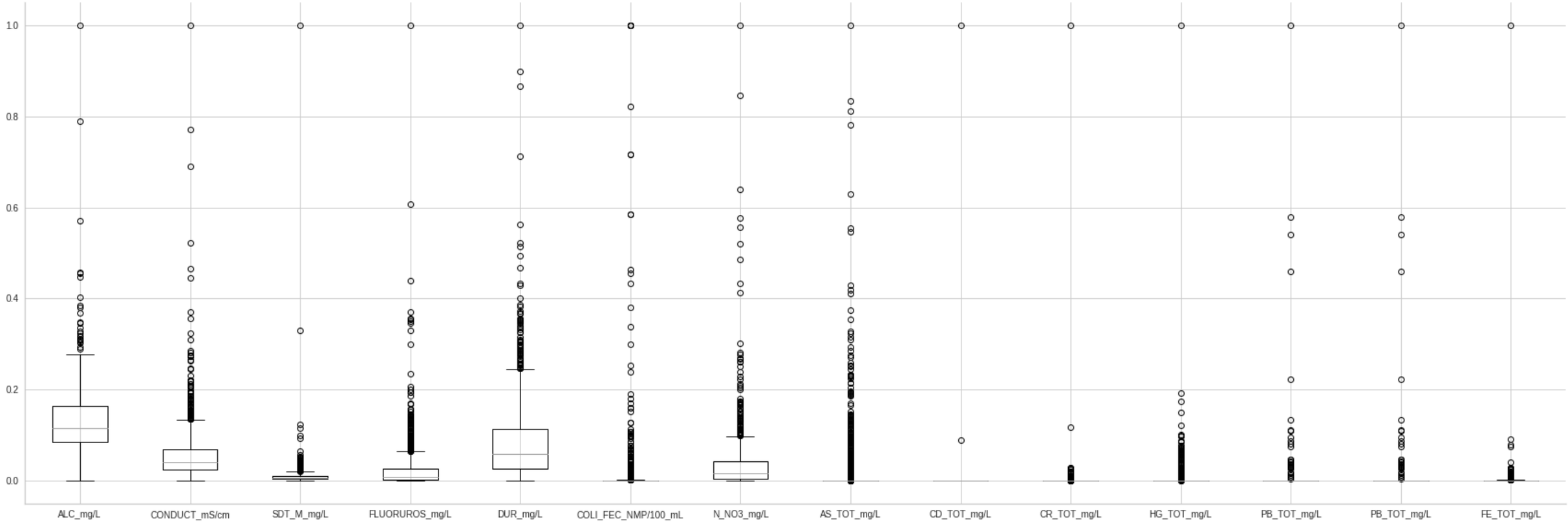
Las columnas que son una excepción a esto son: ALC_mg/L, CONDUCT_mS/cm, SDT_M_mg/L, DUR_mg/L, COLI_FEC_NMP/100_mL.

Coeficiente de variación

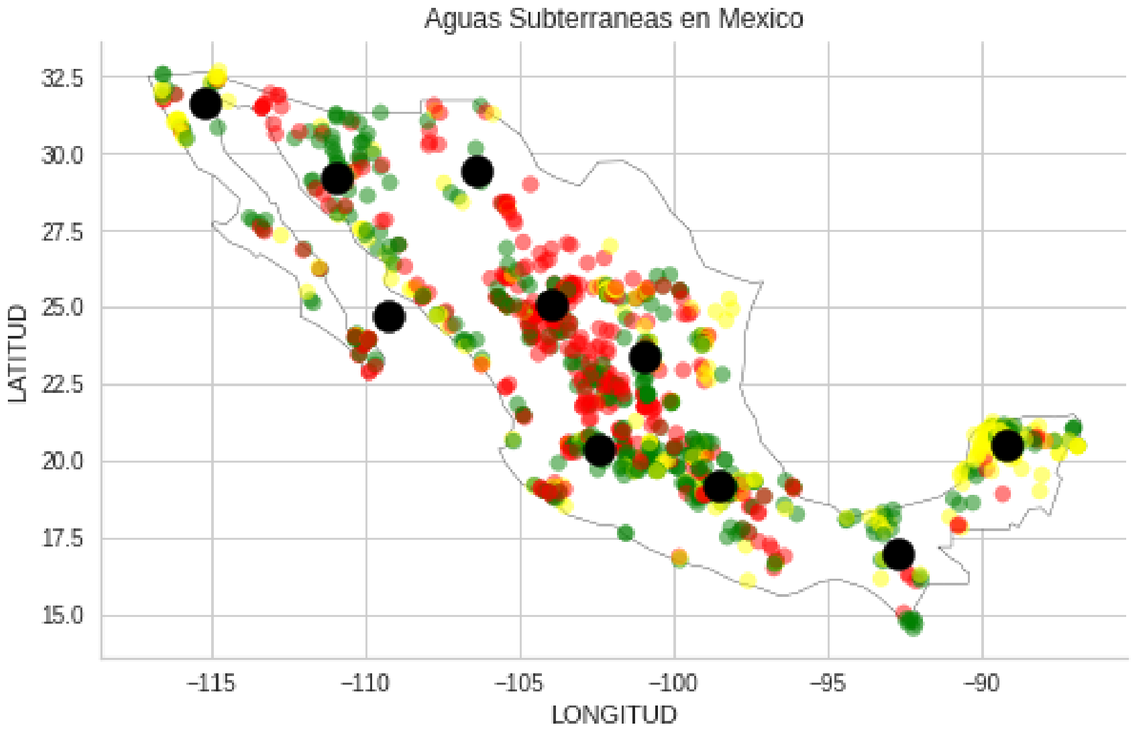
En la mayoría de las columnas se presenta con valores cercanos a 0, un índice claro de poca variabilidad en los datos.

Análisis

Del Boxplot se obtuvo que se presentan muchos outliers, en las variables numéricas del dataset.

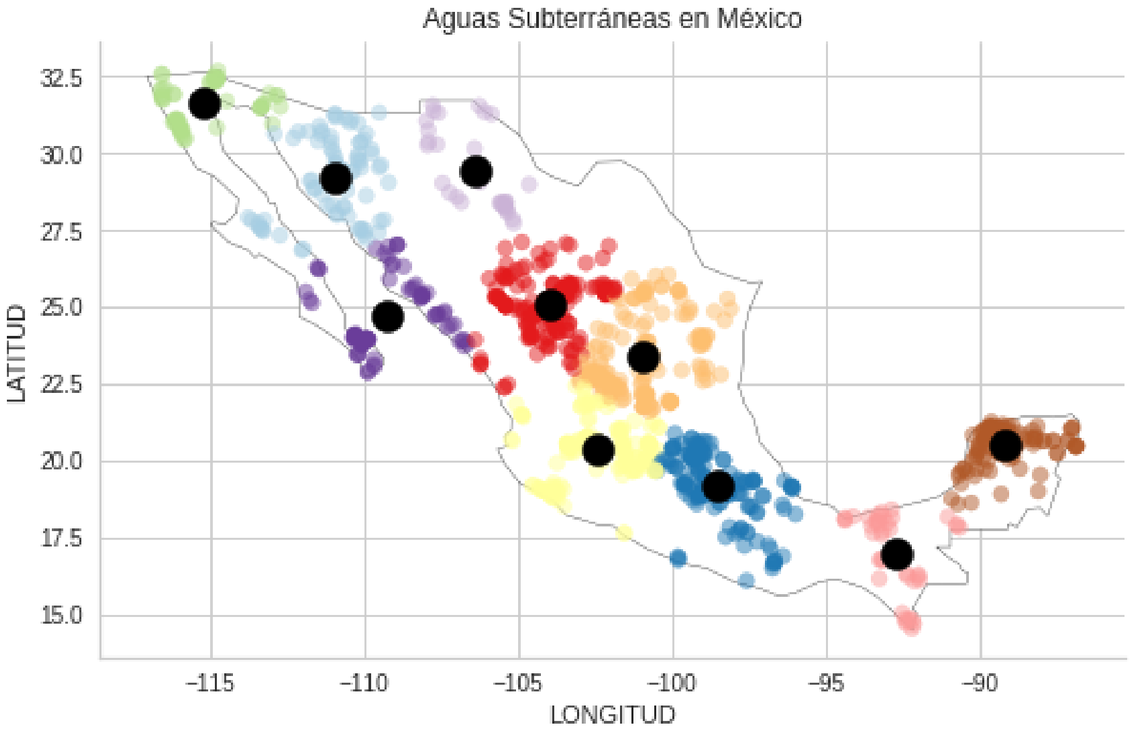


K-Means



Análisis del Mapa con sus clusters y su color de semáforo

Usamos el método del codo para determinar el número de clusters o centros necesarios para explicar la calidad del agua dependiendo de las coordenadas geográficas, sin embargo el valor obtenido (K=3), no era el adecuado para explicar la calidad del agua por medio de las coordenadas, por lo cual probamos con diferentes valores dentro del parámetro n_clusters para ver si alguno, se adaptaba mejor a la información proporcionada, al final usamos K=10, pero no clasificaba de la mejor forma la calidad del agua.



Análisis de los cuerpos de agua y sus clusters (K=10)

Mediante el análisis visual llegamos a la misma conclusión de que no existe relación entre la calidad del agua y su ubicación geográfica agrupada por medio de K-Means. Las clases de la variable objetivo SEMAFORO se encuentran muy juntas. Por lo que una opción para mejorar la clasificación, es utilizar modelos supervisados de tipo Ensamble.

Clasificación

De los resultados de K-Means quedó claro que, se requería de modelos de clasificación más complejos, por esta razón se utilizaron los árboles de decisión y los bosques aleatorios para poder hacer mejores predicciones de las distintas clases de la variable SEMAFORO.

En el modelo del Decision Tree se utilizaron los parámetros, max depth y class weigth, el primero se usó para determinar la profundidad máxima del Decision Tree, en este caso 5, el segundo parámetro se utilizó “balanced” para balancear las clases, dado que las 3 clases de la variable SEMAFORO cuentan con diferentes proporciones.

Con Random Forest se seleccionaron los mismos parámetros, teniendo la misma profundidad máxima que el Decision Tree, en tanto que en el parámetro de balanceo se selecciono “balanced subsample”, para calcular las ponderaciones en base a la muestra Bootstrap de cada árbol generado. A continuación se presentan los scores de ambos modelos:

Reporte de clasificación del Decision Tree

	precision	recall	f1-score	support
class 0	0.79	0.84	0.81	49
class 1	1.00	0.68	0.81	78
class 2	0.80	1.00	0.89	87
accuracy			0.85	214
macro avg	0.86	0.84	0.84	214
weighted avg	0.87	0.85	0.84	214

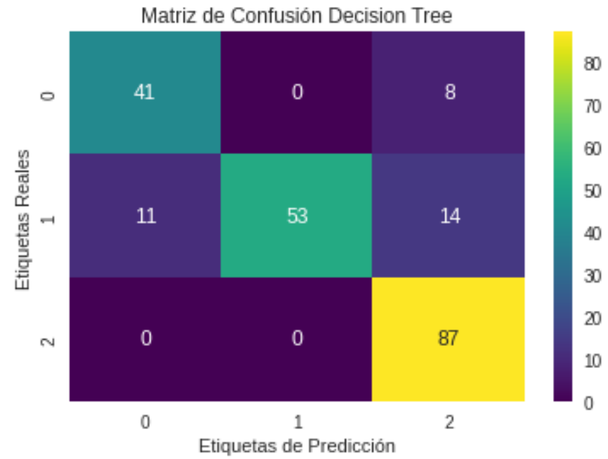
Reporte de clasificación del Random Forest

	precision	recall	f1-score	support
class 0	0.84	0.84	0.84	49
class 1	1.00	0.90	0.95	78
class 2	0.91	0.99	0.95	87
accuracy			0.92	214
macro avg	0.91	0.91	0.91	214
weighted avg	0.92	0.92	0.92	214

Resultados

Decision Tree

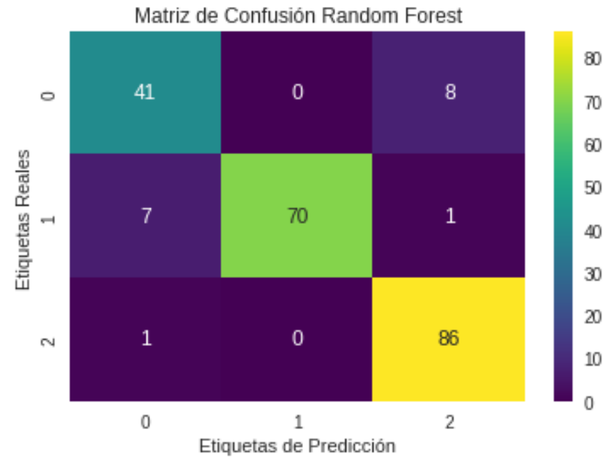
Del modelo con Decision Tree se obtuvieron los scores de: 0.85, 0.86 y 0.84 para las métricas de accuracy, precision y recall, respectivamente. La matriz de confusión arrojó los siguientes resultados:



Tal como se puede observar la mayoría de las clases están siendo clasificadas de manera correcta, sin embargo todavía existen muchos casos, donde el algoritmo no alcanza a predecir de forma correcta el semáforo al que pertencen.

Random Forest

Del modelo con Random Forest se obtuvieron los scores de: 0.92, 0.91 y 0.9 para las métricas de accuracy, precision y recall, respectivamente. La matriz de confusión arrojó los siguientes resultados:



Tal como se puede observar la mayoría de las clases están siendo clasificadas de manera correcta, la clase 0 se encuentra con el mismo número de predicciones correctas, que en el modelo de Decision Tree, sin embargo la clase 1 tuvo un aumento significativo, en tanto que la clase 2 tuvo una ligera disminución.

Principales hallazgos del análisis y clasificación de los datos por medio de los modelos de Decision Tree y Random Forest

- En total se cuenta con un 40% de mantos subterráneos con semáforo verde, mientras que en el color rojo se encuentra el 36% y el resto en color amarillo.

- Las variables que más le aportarán a nuestro modelo, serán las relacionadas a si el agua cumple o no con algún indicador, dado que estas servirán como patrones que a simple vista no se ven, pero indican la calidad del agua.

- El clasificador más óptimo será el de Random Forest, dado que en todos los scores de diferentes métricas como "accuracy", "recall" y "precision" muestra mejores resultados en comparación al modelo del Decision Tree, no sólo eso sino que además la matriz de confusión muestra que el modelo con Random Forest, clasifica mejor los Verdaderos Positivos, Falsos Positivos, Falsos Negativos y Verdaderos Negativos.

- La curva de aprendizaje muestra que el modelo con Random Forest tiene mejores resultados en todas las métricas utilizadas, además de no mostrar un sobreentrenamiento, en tanto que la curva de aprendizaje del Decision Tree, si bien nos muestra que tiene buenos resultados, no se compara a los resultados obtenidos con el modelo de Random Forest.

Conclusiones

1.

El modelo con Random Forest arroja mejores predicciones en comparación al modelo de Decision Tree, dado que las métricas de accuracy, recall y precision muestran un mejor desempeño con Random Forest.

2.

Dado que las clases de la variable SEMAFORO se encuentran muy pegadas entre sí, el modelo de K-Means no será útil para determinar si las coordenadas geográficas determinan la calidad del agua.

3.

Para el modelo de clasificación de la variable SEMAFORO, las features que más valor tendrán para el modelo, serán las que indiquen si el agua tiene o no algún componente químico en particular.

Referencias bibliográficas

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. " O'Reilly Media, Inc.".