

RETO FINAL

Ciencia y analítica de datos (Gpo 10)

César Iván Pedrero Martínez

Análisis de datos de la calidad
de agua en pozos
subterráneos

Problema



DATOS DE POZOS SUBTERRÁNEOS

Para este proyecto, se usó la base de datos de agua subterránea en la cual se documentó el nivel de calidad de agua en los pozos, así como un análisis de contaminantes, su ubicación, etc.

PASOS DEL ANÁLISIS

Los pasos del análisis de los datos fue el siguiente:

- Limpieza del set de datos
 - Reemplazo de nulos
 - Estandarización de datos mixtos (número y palabra en la misma columna)
- Análisis de los datos
 - Descripción de los datos y outliers
 - K-means para visualización de los pozos y patrones de semáforo
- Reemplazo de variables categóricas (one-hot encoding, label encoding)
- Análisis de la importancia de las variables
- Conclusiones

Limpieza de los datos



VARIABLES CUANTITATIVAS

Reemplazo de nulos

Se reemplazaron los valores nulos de las variables cuantitativas con la media aritmética (ALC_mg/L, CONDUCT_ms/cm, SDT_M_mg/L, etc).

VARIABLES CUALITATIVAS

Reemplazo de nulos

Se reemplazaron los valores nulos de las variables cuantitativas con la moda (CALIDAD_ALC, CALIDAD_CONDUC, CALIDAD_SDT_ra, etc).

VALORES ESPECÍFICOS

Reemplazo de nulos

La columna CONTAMINANTES contenía valores nulos cuando el pozo no tenía ningún contaminante. Los valores nulos se reemplazaron por "NINGUNO"

VALORES BOOLEANOS

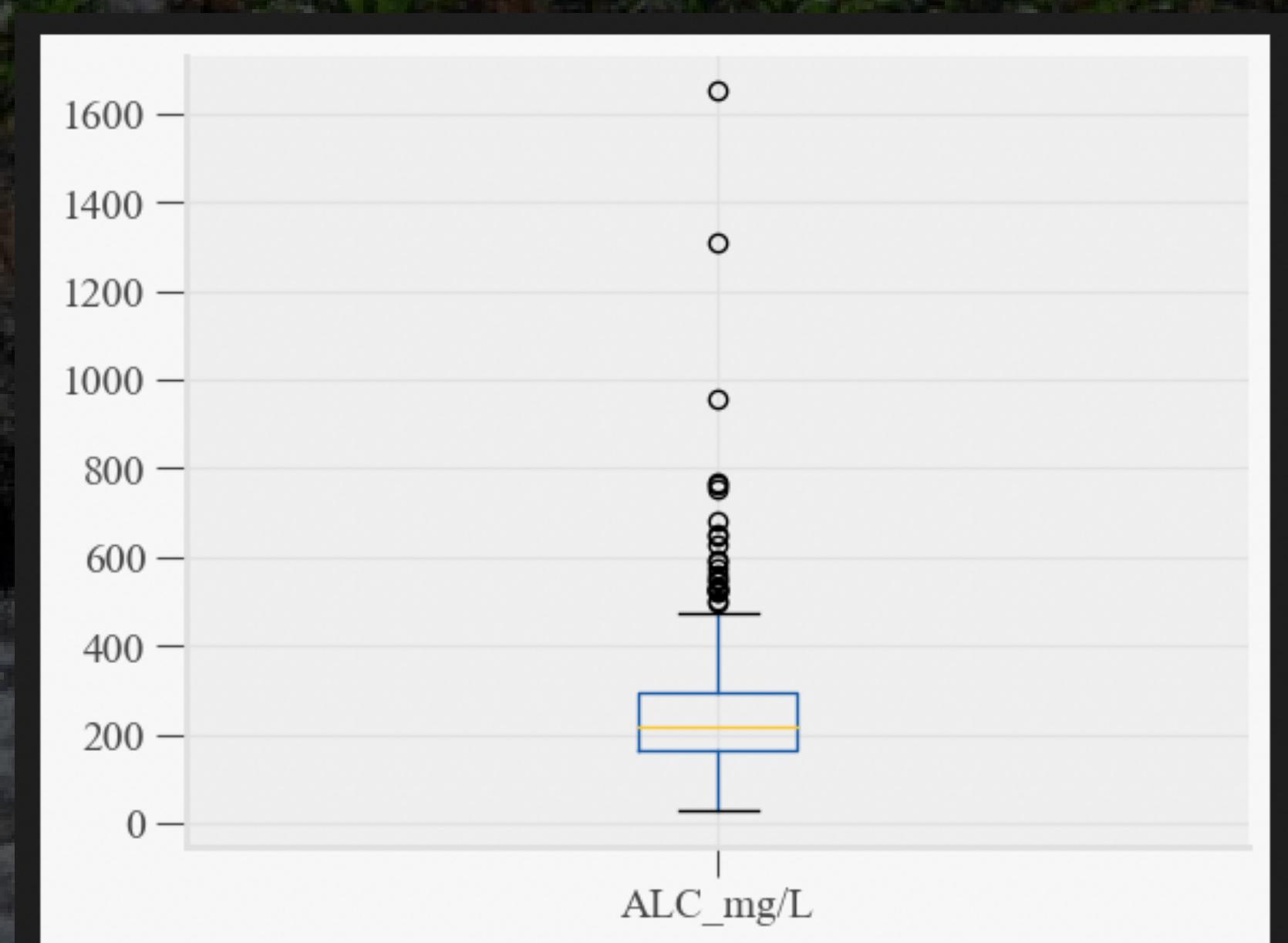
Reemplazo de nulos y cambio de tipo

Había columnas con el prefijo CUMPLE_CON que tenían valores booleanos escritos como las palabras SI/NO/ND. Para estas columnas, se reemplazó la palabra ND por la moda de la columna y se convirtieron los valores a números enteros (SI: 1, NO: 0).

Análisis de outliers

Los outliers más prominentes se encontraban en las columnas con las cantidades de mg/L, mS/cm , NMP/100mL.

Ejemplo
ALC_mg/L

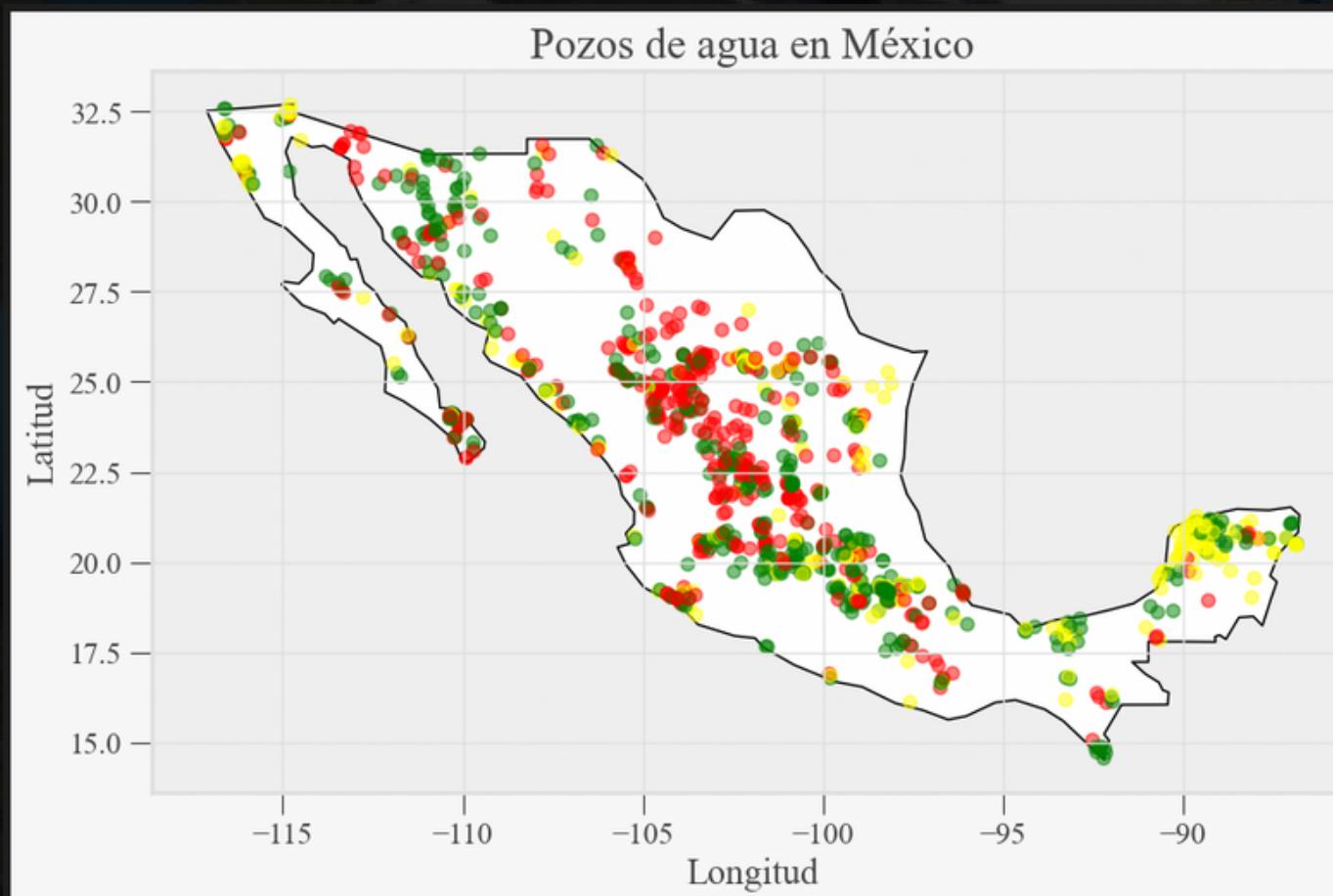


Para más información de los outliers, referirse al notebook de la primer entrega ([click aquí](#)).

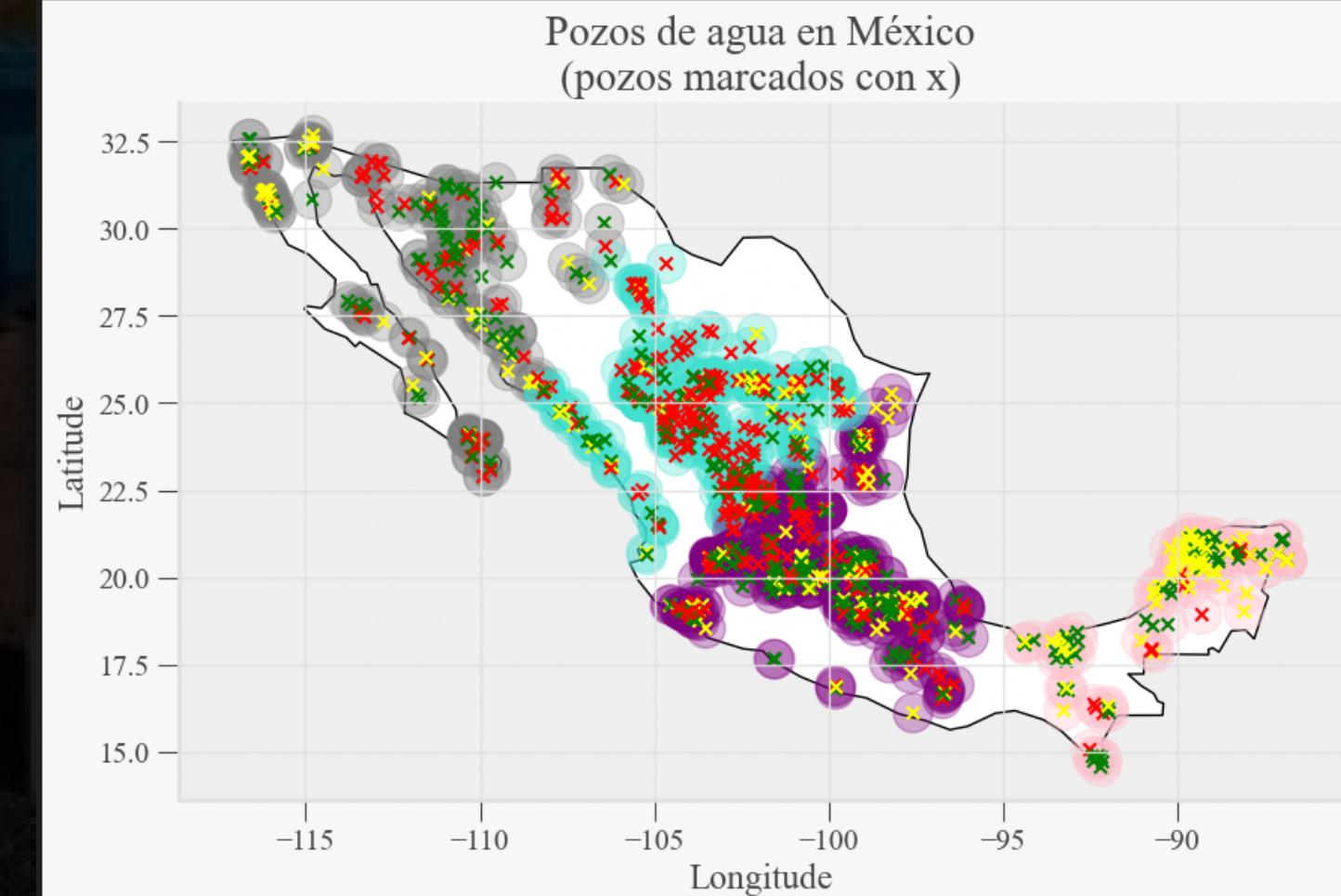
Análisis con K-means

Se graficaron los datos geográficos del set de datos utilizando geopandas

Pozos en la república de acuerdo a su color de semáforo



Pozos en la república después de una agrupación con K-means



Se usó una agrupación con 4 clusters (número de clusters dado por el 'elbow method'). Los pozos se denotan por una X mientras que los clusters se denotan por una sombra circular detrás del pozo.

Los colores de los clusters son: gris, cyan, morado y rosa.

Codificación de las variables categóricas

Contaminantes

La columna de contaminantes tenía datos separados por comas. Se hizo un one-hot encoding por cada valor posible dentro de las columnas y separados por coma (ej: CONTAMINANTES_DT, CONTAMINANTES_FE, CONTAMINANTES_FLUO, etc).

Categorías

Para los demás datos categóricos, se hizo un label encoding para no crear más columnas dentro del set de datos y tener valores numéricos que representaran su categoría. Esto fue útil en especial para la columna semáforo, ya que fue más fácil manejar una sola columna con diferentes categorías numéricas durante la predicción (ej: Estado{NUEVO LEON, AGUASCALIENTES...} -> Estado{0,1,2...18}).



Importancias: Modelo de clasificación

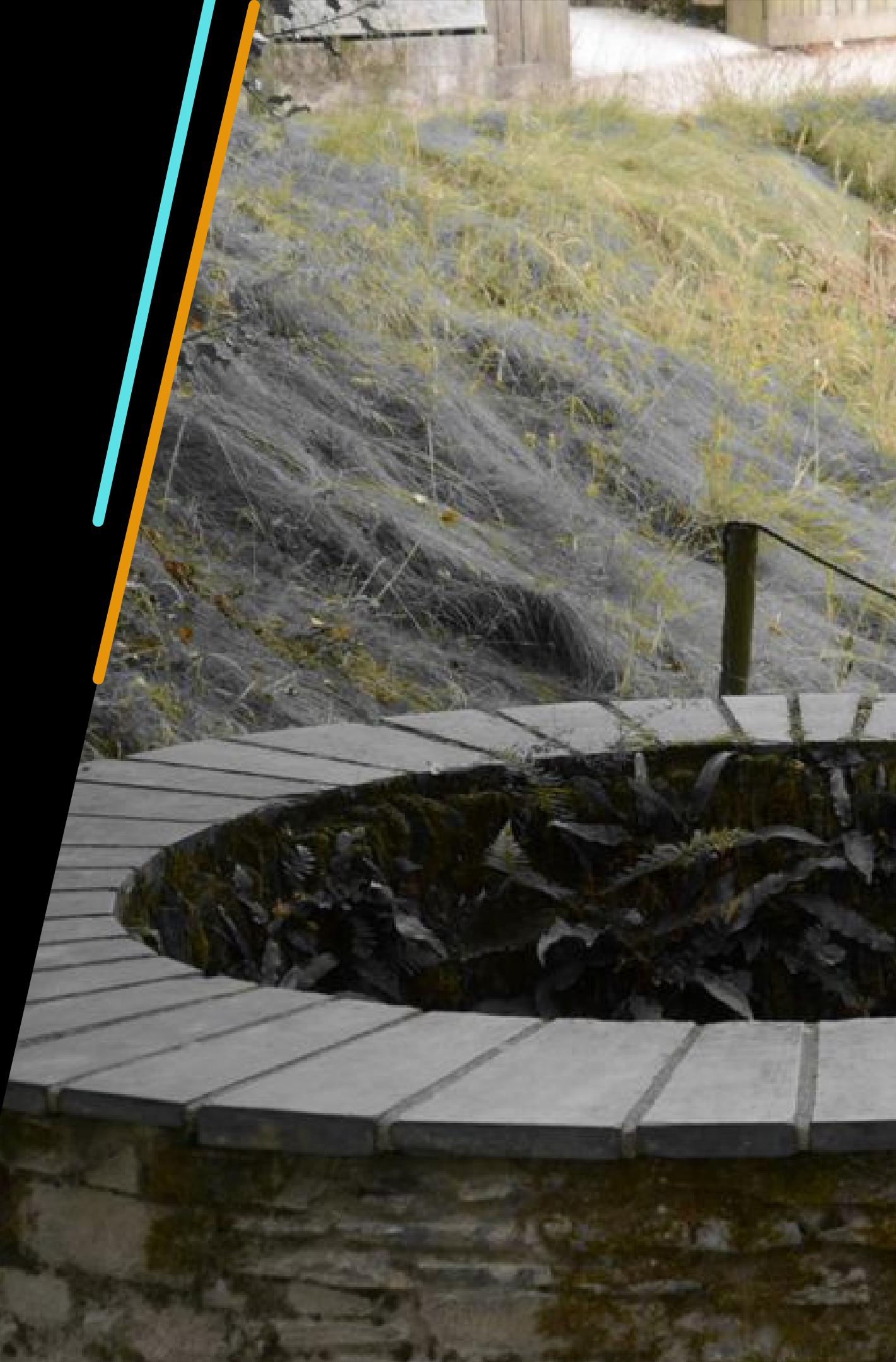
El modelo que proveyó el mejor resultado fue el [RandomForestClassifier](#) (se intentó un árbol de decisión, pero este daba demasiadas variables con 0% al momento de analizar las importancias).

```
RandomForestClassifier  
RandomForestClassifier(n_estimators=20, random_state=0)
```

Se entrenó un modelo con 20 estimadores y el porcentaje de división de los datos para entrenamiento y pruebas fue del 70:30.

```
accuracy_score(y_test, y_pred) ?  
✓ 0.3s  
0.9906542056074766
```

El modelo tuvo 99% de precisión al comparar los datos de testing con los inferidos



Resultados

De la predicción usando RandomForestClassifier

Variable más importante **CONTAMINANTES_NINGUNO**

Esto era de esperarse ya que es una columna booleana que indica si el agua tiene contaminantes o no (sin contaminantes indica un semáforo verde).

14.8 %

Contaminante más importante **Fluoruros**

El contaminante con más importancia en los datos de los pozos fueron los fluoruros.

6 %

Un dato curioso es que la **latitud** tuvo una importancia mayor que la **longitud** en el análisis

Latitud

1 %

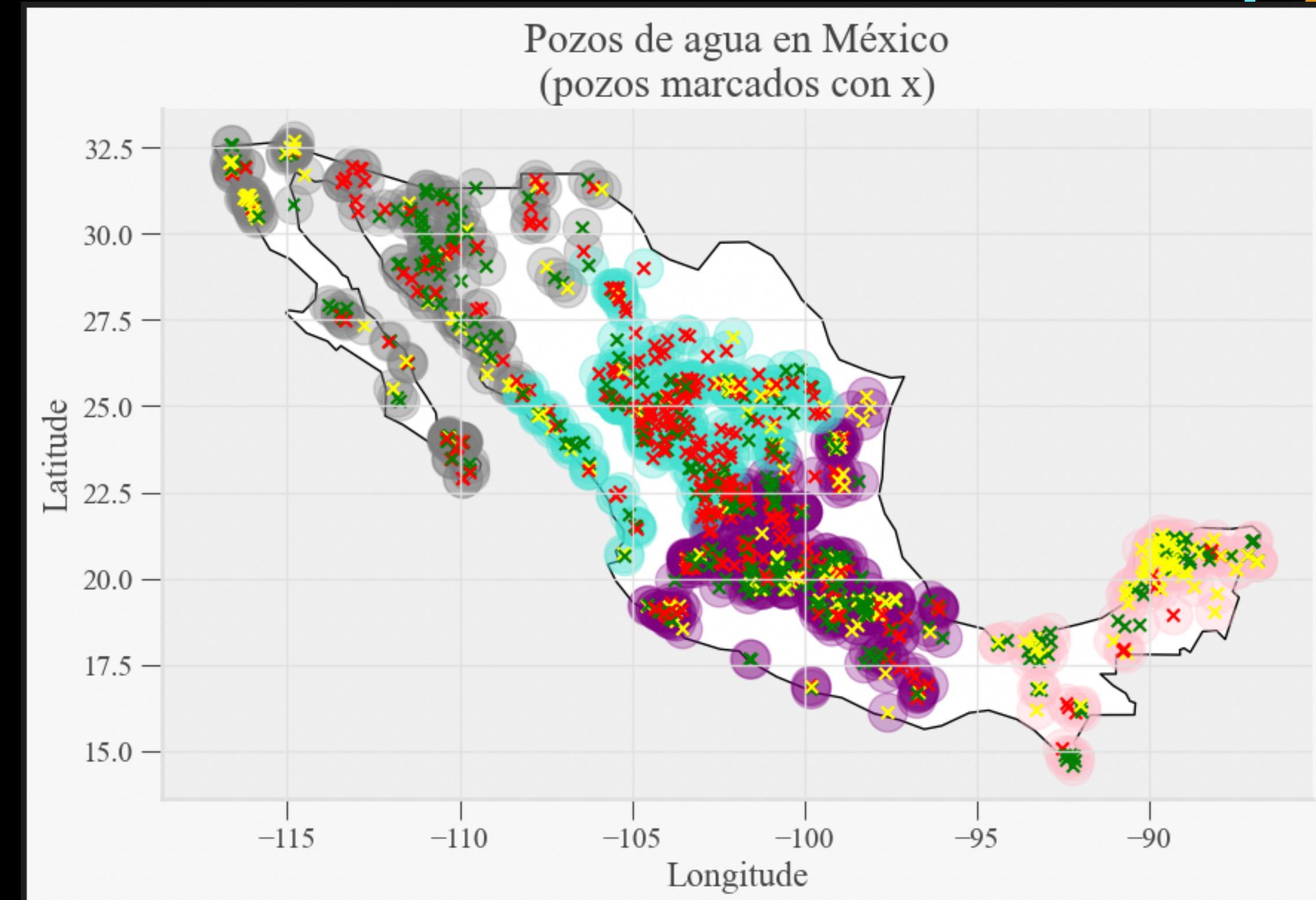
Longitud

0.5 %

Resultados K-means

Parece ser que existe una correlación entre la calidad del agua y su ubicación geográfica:

- Se puede apreciar que en el sureste de México prevalecen mas las aguas con semáforo amarillo.
- El centro superior de la república la gran mayoría de los pozos tienen un semáforo rojo.
- El centro inferior de la república y la península de Baja California (y una parte del norte) tienen una combinación de semáforos verdes con amarillo (con un porcentaje más bajo de semáforos rojos).





¡Gracias!

César Iván Pedrero Martínez
A01366501