Tecnológico de Monterrey

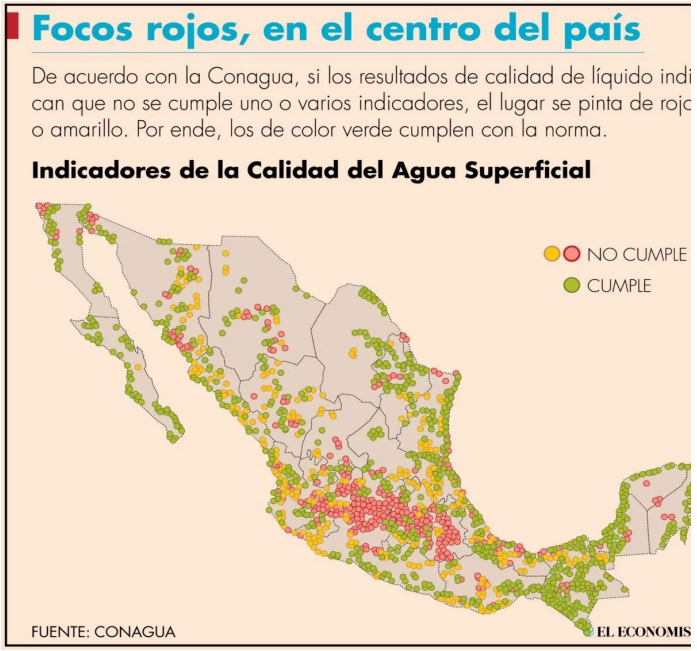
# Proyecto Final

Equipo 3:
Julio Osvaldo Hernández Bucio
Juan Antonio Melendres Villa

# Base de datos

Aguas superficiales que se consideraron:

- 3493 registros
- 53 columnas
- Algunas de las columnas son de información general, mientras que otras muestran datos relevantes del registro
- Variables de salida es la columna de semaforo (Verda, Amarillo, Rojo)

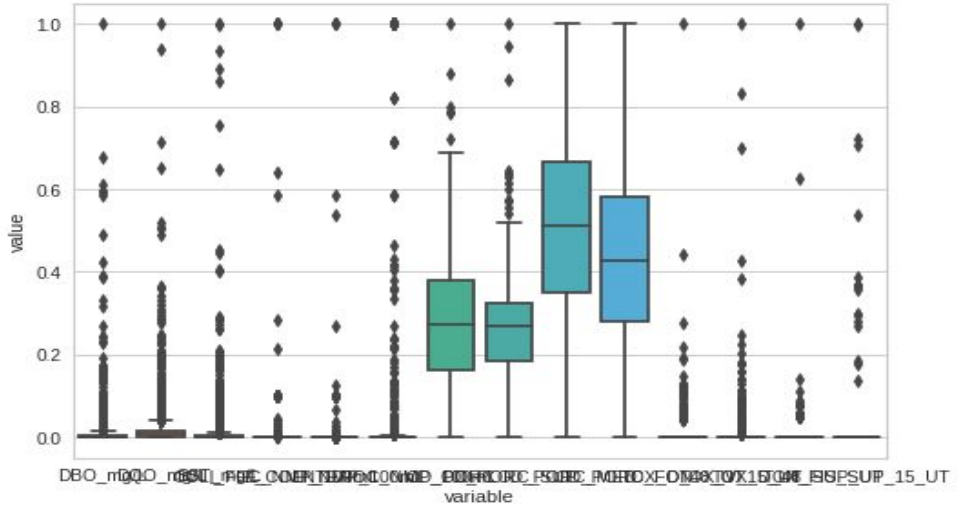



El Economista. (2022, 1 agosto). *Contaminada, 59.1% del agua superficial de México*.

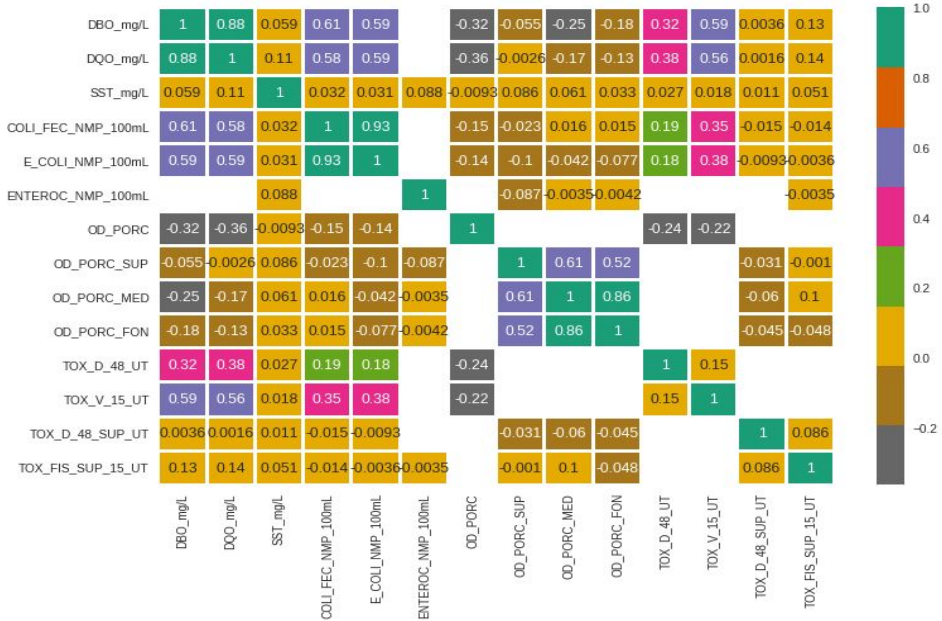
https://www.eleconomista.com.mx/politica/Contaminada-59.1-del-agua-superficial-de-Mexico-20220801-0005.html

# Análisis y Limpieza

- Para el análisis de variables, primero se buscaron las columnas que no tenían un valor significativo para el desafío, como las columnas clave, estado, ubicación, piscina, cuerpo de agua y tipo porque no representan información útil y se puede inferir.
- En esta gráfica nos basamos para determinar qué columnas son más significativas que otras.



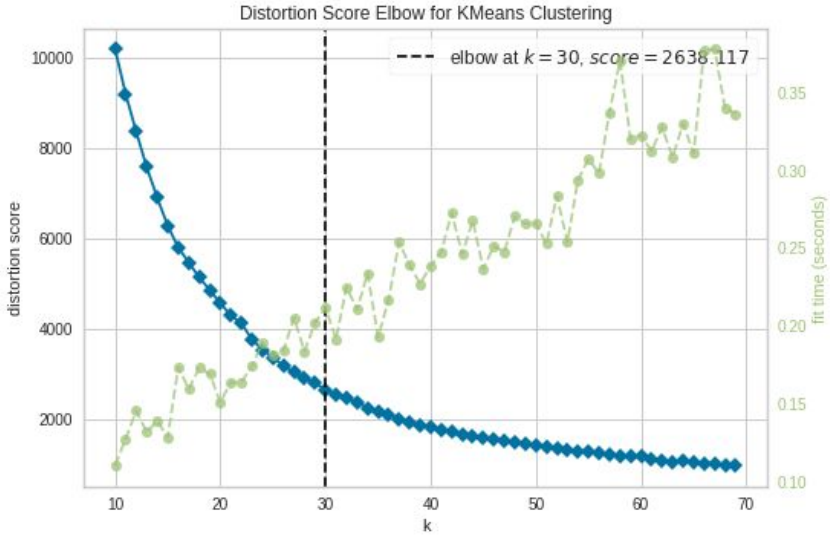
# Anàlisis de variables (Correlación)



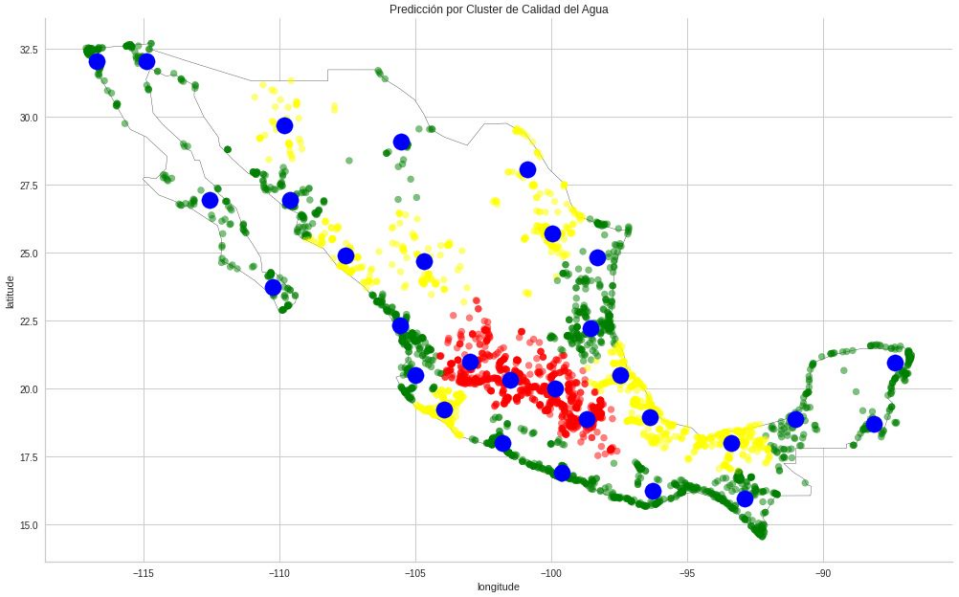

Se construyó una matriz de correlación simplemente para determinar si era necesario aplicar alguna técnica de reducción de dimensionalidad, aunque algunos valores mostraron bajas correlaciones, se decidió no reducir el número de componentes para no afectar el resultado final. No se muestra correlaciones significativas más allá de lo que realmente querían entender cuáles de estas columnas eran más importantes para los siguientes pasos.

# K means

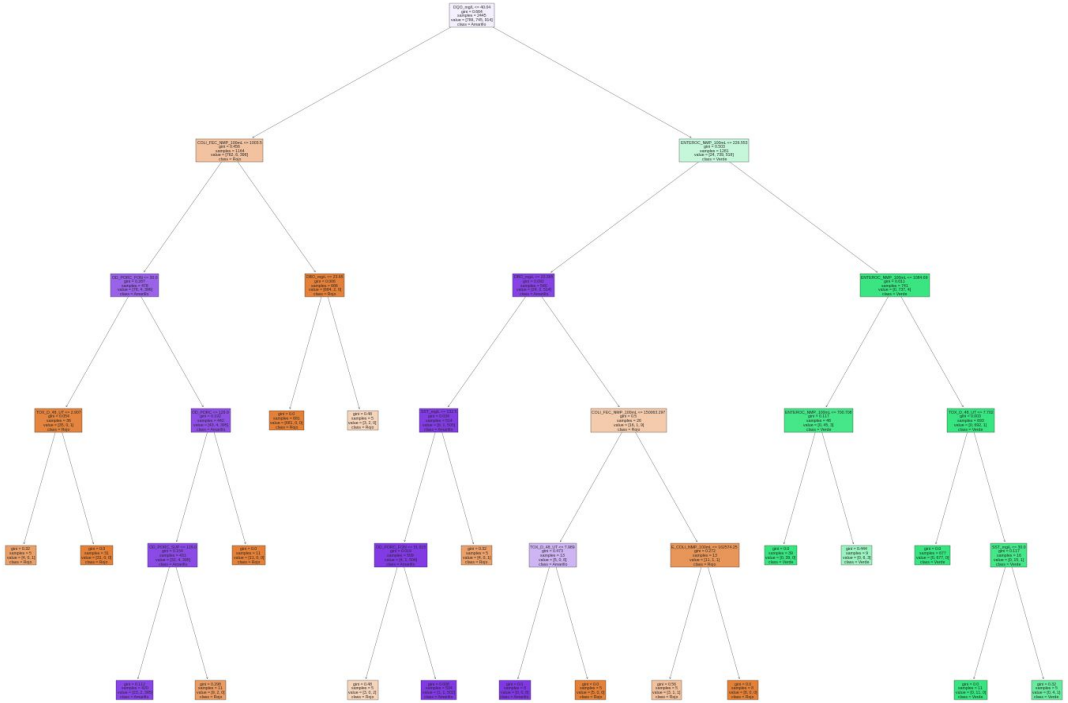
- Prueba del codo = 30 Clusters




- Clusters representados en el mapa



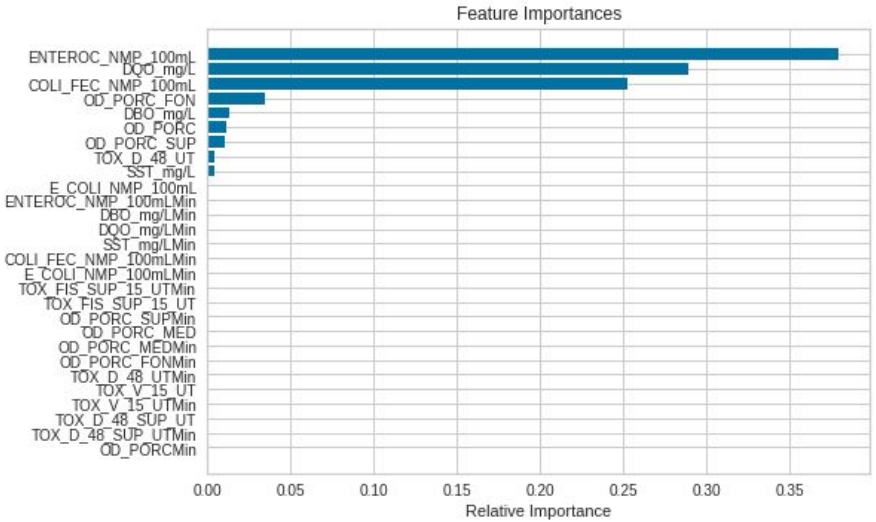
# Árbol de decisión



Árbol

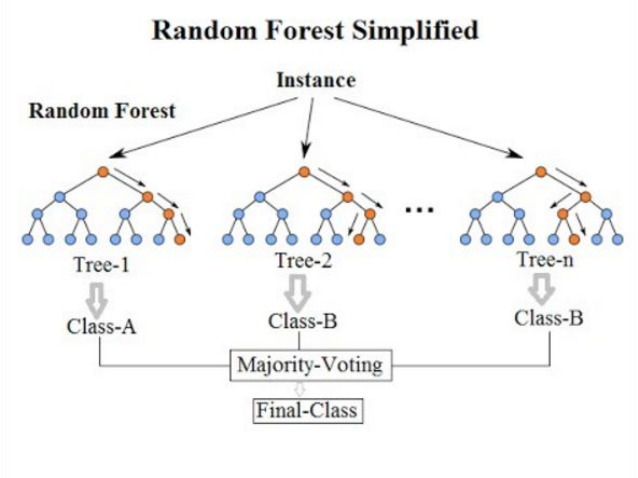
# Árbol de decisión

Con base en el gráfico de barras de importancia del atributo, se concluye que las sustancias más importantes son ENTEROC_NMP_100mLMin, COD_mg/L, COLI_FEC_NMP_100mL y OD_PORC_FON. Reconocemos que su presencia es un factor determinante en el resultado.



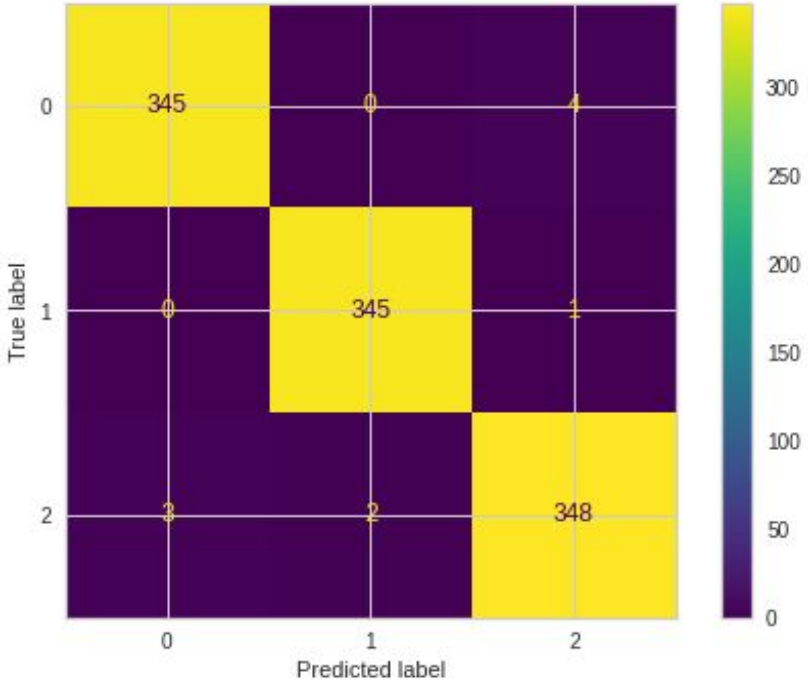
# Random Forest vs otros modelos

Los modelos fueron entrenados y probados con sus respectivos parámetros, extrayendo las predicciones a partir de los valores extraídos durante la distribución inicial. Las mediciones de rendimiento arrojaron los siguientes resultados.




**LogisticRegression** 0.7309160305343512
**RandomForestClassifier** 0.982824427480916
**SVC** 0.5152671755725191
**VotingClassifier** 0.9446564885496184

# Matriz de confusión




Una matriz de confusión es una técnica para resumir el rendimiento de un algoritmo de clasificación. La precisión de la clasificación por sí sola puede ser engañosa si tiene un número desigual de observaciones en cada clase o si tiene más de dos clases en su conjunto de datos. Calcular una matriz de confusión puede darle una mejor idea de qué está haciendo bien su modelo de clasificación y qué tipos de errores está cometiendo.

# Conclusiones

La actividad fue muy interesante porque nos ayudó a obtener una comprensión integral de los modelos de clasificación y el significado de K-means. Sería interesante analizar otro tipo de etiquetas para trabajos futuros, y también podríamos confirmar los valores importantes para la clasificación. Comprender los factores  que afectan la calidad del agua es el primer paso para crear estrategias de mejora.