TÍTULO:

CALIDAD DEL AGUA EN SUDÁFRICA

XVII Edición Master Data Science – KSchool

Elena González Gutiérrez

Índice

1. Introducción

La calidad del agua es un tema que cada vez preocupa más en todos los países del mundo por motivos como la salud de la población, la calidad ambiental de los ecosistemas y el desarrollo económico.

Sudáfrica depende principalmente de los recursos de agua superficiales para la mayoría de sus necesidades urbanas, industriales y de riego. El uso del agua está dominado por el riego que representa más del 60% del uso total de agua en el país. Los requerimientos de agua para uso urbano y doméstico representan casi el 30% y el resto se usa para minería, industrias a granel y como agua de refrigeración para la generación de energía.

La calidad del agua en muchas de las corrientes superficiales del país se ha visto gravemente comprometida por el tratamiento y control inadecuados de las descargas de efluentes y la escorrentía urbana/agrícola. Esto plantea serios riesgos ambientales, de salud y económicos en muchos lugares.

En este trabajo voy a centrarme en el estudio de la calidad del agua derivada de la química inorgánica a partir de los datos recopilados del monitoreo del agua en 1045 estaciones de ríos, embalses y lagos que cubren todo el país y que se reparten en 21 regiones de drenaje.

1.1. Objetivos

El trabajo está dividido en dos estudios diferentes:

- **Análisis y caracterización de las aguas superficiales**

Pretendo hacer un análisis de la calidad del agua basándome en el estudio de 3 parámetros muy importantes de las aguas superficiales del país:

- El **Índice de Calidad del Agua** *(Water Quality Index, **WQI**)*

  que es un valor numérico que califica en una de las 5 categorías, la calidad del agua de una corriente superficial, con base en las mediciones obtenidas para un conjunto de varias variables registradas en una estación de monitoreo concreta en un tiempo determinado.

- El **Índice de Química Inorgánica** *(Inorganic Chemistry Index, **ICI**)*

  que es un índice que determina el porcentaje de la química general del agua que se deriva de todas las fuentes, excluyendo la meteorización química de las rocas, es decir, el valor de este índice determina si el agua está dominada por la composición geológica de la zona o, por lo contrario, está dominada por fuentes alternativas como la industrialización, la minería o la agricultura. Además, a partir de este parámetro, se caracterizan las regiones de drenaje en 3 grupos diferentes, cada uno con su propia química característica del agua.

- La **Relación de Adsorción del Sodio** *(Sodium Adsorption Ratio, **SAR**)*

  que es un indicador de la idoneidad del agua para su uso en riego agrícola y que también diagnostica el peligro de sodicidad de un suelo.

- **Predicción del Índice de Calidad del Agua, del Índice de Química Inorgánica y de la Adsorción de Sodio utilizando Algoritmos de Regresión de Aprendizaje Supervisado**

Hay que tener en cuenta que, para los cálculos del WQI, del ICI y del SAR, se necesitan las mediciones de las concentraciones de 9, 5 y 3 elementos químicos, respectivamente.

La idea es hacer la mejor predicción posible de estas 3 variables a partir de únicamente 3 medidas económicas y muy sencillas de obtener: la **conductividad eléctrica**, el **PH** y la **latitud** de la estación de medida. Si se consiguiera este objetivo, se podría obtener información muy útil para el monitoreo de la calidad del agua a un coste económico y con una inversión de tiempo muy inferiores a lo que se requeriría al calcular matemáticamente dichas variables.

Para calcular el SAR, sólo se necesitan las concentraciones de 3 elementos que son el Sodio, el Calcio y el Magnesio, pero son más difíciles de obtener y requieren de un mayor tiempo de inversión que la medición de la conductividad eléctrica, el PH y la latitud.

1.2 Limitaciones

La principal limitación ha sido la **falta de datos** a partir del año 2013. Me puse en contacto con el Departamento del Agua y Sanidad de Sudáfrica para pedir acceso a estos datos y los dataset que me proporcionaron no contenían ningún análisis completo por lo que tuve que descartarlos.

Existe un **sesgo de datos** que hay que tener en cuenta. El sesgo geográfico se debe a que hay una mayor densidad de estaciones de monitoreo en unas regiones que en otras, y el sesgo temporal también existe al haber estaciones que contienen muchos datos durante un largo período de tiempo mientras que otras tienen conjuntos de datos muy limitados.

En este proyecto sólo he considerado las **características químicas y físicas del agua** y no he podido encontrar datos sobre la química orgánica que pudiesen completar la información sobre la calidad del agua. Evidentemente este sería un punto muy interesante a incluir puesto que la contaminación microbiológica provocada frecuentemente por indicadores fecales es la causante de miles de muertes principalmente en países en desarrollo como es el caso de Sudáfrica.

Hay múltiples **definiciones del Índice de Calidad del Agua**, no es un parámetro universal. He tenido que optar por una definición que sólo considerase la química inorgánica del agua pues únicamente dispongo de esos datos.

Las **Regiones de Drenaje**, las he tenido que crear utilizando el software http://geojson.io/ ya que no corresponden a regiones oficiales de Sudáfrica.

En Sudáfrica hay cerca de 4000 **licencias de minas** activas. A lo largo de las décadas este número ha ido cambiando y se han ido cerrando y abriendo nuevas. No me ha sido posible acceder al repositorio de las minas del país, por lo que sólo he posicionado las minas que podrían afectar a los puntos de muestreo del agua por su elevado contenido en sulfatos, así como aquellas que están próximas al río Vaal (el más importante de Sudáfrica).

2. Estructura del Proyecto

El proyecto está organizado con la siguiente estructura:

- **Memoria del proyecto**: Este documento en PDF en el cual se describe el proyecto, los pasos seguidos y las conclusiones.
- **Repositorio GitHub**: Contiene los dataset, notebooks (con los correspondientes códigos) y datos necesarios para ejecutar el proyecto.

La URL del repositorio a consultar de github es:

https://github.com/elenuskiya/TFM_EGG

2.1 Requisitos de instalación y librerías necesarias

Cualquier PC o portátil con al menos 8 GB de RAM es suficiente como para recrear el código.

Para la ejecución de los notebooks es necesario tener **Python** instalado y **Tableau** para ver las visualizaciones interactivas.

Las librerías necesarias para ejecutar los diferentes Notebooks son las siguientes:

- **Pandas**
- **Numpy**
- **Seaborn**
- **Matplotlib**
- **Altair**
- **Ternary**
- **Sklearn**

2.2 Estructura de las carpetas

Dentro del **repositorio** se encuentran las siguientes carpetas:

- **data**: Contiene todos los datos utilizados para la ejecución y desarrollo del proyecto, tanto los datos en bruto, como los modificados o creados.
- **Notebooks**: Contiene los notebooks del estudio.
- **Dashboard**: Se encuentra las visualizaciones realizadas en Tableau.
- **Complementary material**: Contiene la documentación utilizada para la obtención del Índice de Calidad del Agua y del Índice de Química Inorgánica, así como material útil para la comprensión del proyecto.

A continuación, describo las carpetas y sus contenidos:

Carpeta	Archivo	Extensión	Descripción
	Datos descargados		
	Dams and lakes 1999-2012 (A-X).xlsx	excel	Dataset con la información de la química del agua de ríos, lagos y embalses.
	Dams and lakes up to 1998 (A-X).xlsx		
	Rivers 1999-2012 (A-X).xlsx		
	Rivers up to 1998 (A-D).xlsx		
	Rivers up to 1998 (E-X).xlsx		
	Sample_stations.xlsx		Dataset que incluye una lista de todas las estaciones de muestreo con detalles como la localidad, tipo de superficie, regiones a las

data			que pertenecen o las coordenadas geográficas.
	southafrica.json	json	Mapa de Sudáfrica
	Dataset creados/modificados		
	chemical_stations.xlsx	excel	Dataset limpio y manipulado creado con la unificación de los dataset descargados de la química del agua
	chemical_WQI_ICI_stations.xlsx		Dataset final de la química del agua con la creación de las columnas del Índice de la Calidad del Agua y del Índice de la Química Inorgánica.
	cities.xls		Dataset creado con las 20 ciudades más pobladas de Sudáfrica
	mines.csv	csv	Dataset creado con las coordenadas, tipo de mineral y nombre de las minas que he considerado necesarias.
	dataset_tableau.xlsx	excel	Dataset con las columnas utilizadas para representar gráficamente con Tableau.
	Vaal_Tableau.xlsx	excel	Dataset con las coordenadas y nombres de los puntos de medida del río Vaal para representarlo gráficamente.
	Regions_SA.geojson	gjson	Mapa con las regiones de drenaje de Sudáfrica.
Notebooks	00_Cleaning_manipulation_and_unification.ipynb	Notebook Jupyter	Unificación, limpieza y manipulación de los datasets de la química del agua, creando el Dataset Final que se usará en los siguientes notebooks.
	01_Part_1_Analysis_and_Characterization_of_surface_waters.ipynb	Notebook Jupyter	Análisis y evolución de los índices de calidad del agua y química inorgánica, caracterización de las regiones por su índice de química inorgánica, diagramas ternarios y análisis del río Vaal.
	02_Part_2.1_Prediction_of_the_Water_Quality_Index.ipynb	Notebook Jupyter	Predicción del Índice de Calidad del Agua usando algoritmos de Regresión de Aprendizaje Supervisado.
	03_Part_2.2_Prediction_of_the_Inorganic_Chemistry_Index.ipynb	Notebook Jupyter	Predicción del Índice de Química Inorgánica usando algoritmos de Regresión de Aprendizaje Supervisado.
	04_Part_2.3_Prediction_of_the_Sodium_Adsorption.ipynb	Notebook Jupyter	Predicción de la Adsorción del Sodio usando algoritmos de Regresión de Aprendizaje Supervisado.
	Part_1_general_visualization.tbwx		Visualizaciones interactivas de las generalidades de Sudáfrica, por

Dashboard		(Libro Tableau, **interactivo**)	país, regiones y estaciones de medida.
	Part_1_chemical_characterization.tbwx	(Libro Tableau, **interactivo)**	Visualizaciones interactivas de la caracterización de las aguas por regiones clasificadas por su ICI.
	Part_1_Vaal_river.tbwx	(Libro Tableau, **interactivo**)	Visualizaciones interactivas del río Vaal.
	Part_1_effect_on_mines_on_water.tbwx	(Libro Tableau, **interactivo**)	Visualización del efecto de las minas sobre las aguas
	Part_1_general_visualizations.pdf	PDF	PDF con las visualizaciones generales más importantes
	Part_1_chemical_characterization.pdf	PDF	PDF con las visualizaciones de la caracterización de las aguas más importantes.
	Part_1_Vaal_river.pdf	PDF	PDF con las visualizaciones del río Vaal.
	Part_1_effect_on_mines_on_water.pdf	PDF	PDF con la visualización de la posición de las minas y las estaciones de medida.
Complementary material	An inorganic water chemistry dataset South Africa.pdf	PDF	PDF con la descripción de los datasets descargados.
	Calculation of the water quality index.pdf	PDF	PDF del que he obtenido la expresión a para calcular el Índice de Calidad del Agua.
	Characterisation of the inorganic chemistry of surface waters in SA.pdf	PDF	Estudio que determina el Índice de Química Inorgánica.

3. Datos

3.1 Obtención de los datos

Los datos que he utilizado en el proyecto los he obtenido de diferentes fuentes que enumero a continuación:

- **Datos de la química inorgánica del agua**

Los datos de la química inorgánica del agua se obtienen en el Departamento del Agua y Sanidad de Sudáfrica y se pueden descargar en el siguiente link:

https://www.waterscience.co.za/waterchemistry/data.html

- **Mapa de Sudáfrica**

Para la visualización con Altair he descargado el mapa de Sudáfrica en:

https://geojson-maps.ash.ms/

- **Coordenadas de las minas en Sudáfrica**

Las coordenadas geográficas de las minas que he utilizado las he obtenido de:

https://mine-alert.oxpeckers.org/

- **Coordenadas de las ciudades de Sudáfrica**

Las coordenadas de las 20 ciudades más pobladas de Sudáfrica las he obtenido directamente de Google.

3.2 Descripción de los datos

- **Datos de la química inorgánica del agua**

Estos datos se encuentran recopilados en los siguientes archivos:

- *Dams and lakes 1999-2012 (A-X).xlsx*
- *Dams and lakes up to 1998 (A-X).xlsx*
- *Rivers 1999-2012 (A-X).xlsx*
- *Rivers up to 1998 (A-D).xlsx*
- *Rivers up to 1998 (E-X).xlsx*

La estructura de estos 5 datasets se describe a continuación:

Columna	Descripción	Medido/Calculado
Station	Código de la estación de medida que hace referencia a la región a la que pertenece	-
POINT ID	Código de la estación de medida	-
DATE	Fecha en la que se recogió la muestra	-
YEAR	Año en que se recogió la muestra	-
EC (mS/m)	Conductividad eléctrica en mS/m	medido
EC (µS/cm)	Conductividad eléctrica medida en µS/cm	medido
EC (calc.) (µS/cm)	Conductividad eléctrica calculada a partir de la composición química	calculado
dEC (%)	Desequilibrio de conductividad	calculado
PH	PH	medido
TDS (mg/L)	Desequilibrio de sólidos disueltos totales	medido
Concentración de iones (mg/L): • **Na** • **Mg** • **Ca** • **F** • **Cl** • **NO2+NO3** • **SO4** • **PO4** • **CaCO3** • **Si** • **K** • **NH4**	Medida de las concentraciones de: • Sodio • Magnesio • Calcio • Flúor • Cloro • Nitratos y nitritos • Sulfato • Fosfato • Carbonato de calcio • Silicio • Potasio • Amonio	medido
TDS (calc.) /mg/L)	Suma de todos los iones en mg/L	calculado
dTDS (%)	Desequilibrio de los sólidos disueltos totales	calculado
TDS/EC (mg/L)/(mS/cm)	Cociente entre los sólidos disueltos totales medidos y la conductividad eléctrica medida	calculado
Sz+ (meq/L)	Carga de cationes corregidos	calculado
Sz- (meq/L)	Carga de aniones corregidos	calculado
Charge Balance (%)	Balance de carga estequiométrica	calculado

Concentración de iones (mmol/L)	Cálculo de las concentraciones de todos los iones medidos en mmol/L	calculado
Ionic Strength	Fuerza iónica calculada a partir de las principales concentraciones molares de iones	calculado
[HCO3]+2[CO3]	Alcalinidad total	calculado
2[SO4]	Contaminación por sulfato	calculado
[Cl]	Salinización con cloruro	calculado
SAR	Relación de adsorción del sodio	calculado
Adj. SAR	Relación de adsorción del sodio ajustado	calculado

La columna "**Medido/Calculado**" de la tabla anterior hace referencia a si el valor ha sido obtenido mediante la medición in situ o en laboratorio, o si lo han calculado a partir de fórmulas.

Por último, hay un archivo que contiene la información relacionada con los puntos de muestreo:

- *Sample_stations.xlsx*

Columna	Descripción
Station	Código de la estación de medida que hace referencia a la región a la que pertenece
POINT ID	Código de la estación de medida
River/Lake/Dam	Tipo de superficie
Brief Locality description of sample station	Localización del punto de muestreo
South Lat. degr/min/sec	Latitud en grados minutos y segundos
South Lon. deg/min/sec	Longitud en grados minutos y segundos
South Lat decimal degr	Latitud decimal
South Long decimal degr	Longitud decimal
Total number of samples	Número de muestras totales medidas en el punto de muestreo
First sample date	Fecha de la primera muestra recogida en ese punto de muestreo
Last sample date	Fecha de la última muestra recogida en ese punto de muestreo
Sample drainage region	Región de drenaje a la que pertenece el punto de muestreo
Sample type code	Código referente al tipo de muestra
Sample type description	Tipo de muestra

- **Coordenadas de las minas en Sudáfrica**

El archivo que contiene la información de las minas es:

- *mines.csv*

Las columnas que forman este dataset:

Columna	Descripción
Region	Región de drenaje a la que pertenece la mina
Mineral	Mineral que explota
Mine Name	Nombre de la mina
lon	Longitud de la mina en decimal
lat	Latitud de la mina en decimal

- **Coordenadas de las ciudades de Sudáfrica**

Este archivo es:

- *Cities.xls*

Contiene las siguientes columnas:

Columna	Descripción
City	Nombre de la ciudad
Region	Región de drenaje a la que pertenece
lon	Longitud de la mina en decimal
lat	Latitud de la mina en decimal
Population	Población (datos 2020)

4. Secuencia de ejecución del proyecto

En esta sección voy a enumerar los pasos a seguir para la ejecución del proyecto.

1. Descargar los dataset:
    a. *Dams and lakes 1999-2012 (A-X).xlsx*
    b. *Dams and lakes up to 1998 (A-X).xlsx*
    c. *Rivers 1999-2012 (A-X).xlsx*
    d. *Rivers up to 1998 (A-D).xlsx*
    e. *Rivers up to 1998 (E-X).xlsx*
    f. *Sample_stations.xlsx*
2. Ejecutar el Notebook ***00_Cleaning_manipulation_and_unification.ipynb***
3. Descargar los datasets:
    a. *chemical_WQI_ICI_stations.xlsx*
    b. *southafrica.json*
    c. *cities.xls*
    d. *mines.csv*
4. Ejecutar el Notebook *01_Part_1_Analysis_and_Characterization_of_surface_waters.ipynb*
5. Para visualizar los **gráficos interactivos** abrir los libros de los archivos:
    a. ***Part_1_general_visualizations.twbx***
    b. ***Part_1_chemical_characterization.twbx***
    c. ***Part_1_Vaal_river.twbx***

    Nota: Si no se dispone de Tableau se pueden visualizar los archivos en PDF (sin gráficos interactivos).

6. Ejecutar el Notebook ***02_Part_2.1_Prediction_of_the_Water_Quality_Index.ipynb***
7. El orden de ejecución de los siguientes Notebooks es indiferente:
    a. ***03_Part_2.2_Prediction_of_the_Inorganic_Chemistry_Index.ipynb***
    b. ***04_Part_2.3_Prediction_of_the_Sodium_Adsorption.ipynb***

5. Metodología

5.1 Limpieza y manipulación de los datos

(datos: Los archivos del punto 1.a de la sección anterior)

*(Notebook: **00_Cleaning_manipulation_and_unification.ipynb**)*

Los **objetivos** de esta fase son:

- **Generar un archivo único con toda la información de la química inorgánica del agua**

No conlleva mayor dificultad que la de cambiar el nombre de una de las columnas del archivo *Dams and lakes 1999-2012 (A-X).xlsx* para poder concatenar en un solo paso los 4 archivos de la química del agua. Se renombran las columnas y se cambia el tipo de "*POINTID*" a objeto.

- **Filtrar, limpiar el archivo de outliers y modificarlo para obtener los datos de interés**

Para que los análisis químicos sean precisos, es necesario realizar un filtrado que cumpla unas determinadas condiciones:

  - El balance de carga estequiométrica debe estar entre el ±5%.
  - Los valores aceptables de δTDS para que las muestras de agua sean precisas deben estar entre 0% y 15%.
  - Los valores de δEC se consideran exactos si están entre ±20%.

Los valores no medidos vienen dados por el número ***-9999***. Los he sustituido por valores nulos ***NaN*** de manera que se tiene un total del **12,5%** de valores nulos en todo el conjunto de datos. El tratamiento de estos datos lo realizaré más adelante.

La manera que he utilizado para visualizar posibles **outliers** ha sido la realización de **histogramas** de los principales elementos químicos, así como del PH y la conductividad eléctrica. De esta forma es fácil ver que todas las variables tienen valores atípicos.

El agua dulce no excede de 10000 µS/cm (valor que ya corresponde a aguas industriales), por lo que he utilizado esta referencia para eliminar análisis erróneos. También he establecido un límite para el fosfato, los nitritos y nitratos, el amonio y el flúor, ya que son elementos que se encuentran en muy pequeñas proporciones y cuya medida es muy sensible. A pesar de que de esta manera siguen quedando valores atípicos para algunos elementos, no he querido eliminarlos para no sobreajustar los datos, pues no es imposible que existan estos valores en el agua superficial.

- **Unificar dicho archivo con el que contiene la información de las estaciones de medida**

Las modificaciones que he hecho en el archivo que contiene la información de los puntos de muestreo han sido las de cambiar el signo de las latitudes, pues no era el correcto, renombrar las columnas y cambiar el tipo de "*POINTID*" a objeto.

Las columnas "*POINTID*" y "*Station*" son códigos que representan las mismas estaciones de medida, pero en algunos casos se carece de la información de una de las dos o de ambas. Por ello he unificado ambos dataset usando por un lado "*POINT ID*" y por el otro "*Station*" y después he eliminado duplicados. Finalmente hay análisis que están realizados el mismo día y en el mismo punto y para obtener un solo valor se ha realizado la media.

- **Crear las columnas con los valores y clasificaciones del WQI, del ICI y de la clasificación del SAR, que serán necesarias para todo el proyecto**

Dos de las variables objetivo las he tenido que calcular a partir de expresiones matemáticas que se encuentran en artículos científicos.

El **Índice de Calidad del Agua** lo he obtenido utilizando la siguiente expresión[2]:

$$WQI = \sum W_i \cdot Q_i$$

donde:

$$W_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

donde:

- $\mathbf{W_i}$ representa el peso relatico para cada parámetro medido
- $\mathbf{W_i}$ es el peso de cada parámetro

Y:

$$Q_i = \frac{C_i x 100}{S_i}$$

donde:

- $\mathbf{Q_i}$ es un parámetro normalizado
- $\mathbf{C_i}$ es la concentración de cada parámetro en (mg/L)
- $\mathbf{S_i}$ es el parámetro estándar de calidad dado por la OMS.

Los parámetros estándar de calidad que se necesitan, son:

Parámetro	Parámetro estándar (S_i)
TDS	500
HCO_3	120
Cl	250
$NO_2 + NO_3$	45
F	1
Ca	75
Mg	30
Na	200
K	10

La clasificación del agua según su valor del WQI es:

Clasificación del WQI	Valores
Excelente	0-50
Buena	50-100
Pobre	100-200
Muy pobre	200-300
No apta	>= 300

El Índice **de la Química Inorgánica** viene dado por[3]:

$$ICI = \frac{[Cl]+2[SO4]+[NO3]+3[PO4]}{[Cl]+2[SO4]+[NO3]+3[PO4]+[HCO3]} x100$$

La clasificación del ICI es:

Clasificación del ICI	Valores
No dominada por la intemperie	>= 70
Mezcla	30-70
Dominada por las rocas	<=30

La **Relación de Adsorción de Sodio** no he tenido que calcularla porque viene dada en los datasets iniciales, pero sí que he creado la columna con su la clasificación de la calidad del agua según su valor[4]:

Clasificación del SAR	Valores
Muy buena	< 2
Buena	2-8
Pobre	8-15
Mala	>=15

- **Terminar de ajustar los datos a partir de las distribuciones de las variables objetivo.**

Finalmente he representado las distribuciones de las 3 variables objetivo para terminar de detectar **valores atípicos** que puedan afectar al desarrollo del proyecto, de forma que he eliminado los valores que desajustaban mucho las distribuciones.

- **Guardar el archivo final.**

El dataset completo, limpio y modificado que se va a utilizar a lo largo del trabajo, se guarda bajo el nombre de *chemical_WQI_ICI_stations.xlsx*.

5.2 Predicciones

Pautas generales aplicadas independientemente del parámetro a predecir:

- **Escalar las características**

Es necesario escalar las características porque están en diferentes unidades y quiero normalizarlas para que las unidades no afecten al algoritmo. La **regresión lineal** y el **Random Forest** no requieren de escala de características, pero otros métodos, como el **SVR** y el **KNN** sí, puesto que tienen en cuenta la distancia euclidiana entre variables. Por eso, como voy a comparar múltiples algoritmos, escalar características es lo más apropiado.

- **Métricas**

He seguido el consejo de **Andrew Ng** de utilizar una única métrica de rendimiento de valor real para comparar los modelos porque simplifica el proceso de evaluación. Voy a utilizar el **Error Absoluto Medio** (M*ean Absolute Error, **MAE***) y el **Error Porcentual Absoluto Medio** (*Mean Absolute Percentage Error, **MAPE***) que son métricas apropiadas para una regresión (son dos errores diferentes, pero representan lo mismo).

- **Tiempo de ejecución**

Calcular el tiempo que lleva entrenar los modelos también puede ayudar a elegir el modelo más adecuado. Los modelos con mejores precisiones normalmente tienen unos tiempos de ejecución más elevados, por lo que hay que tenerlo en cuenta a la hora de la elección final.

- **Establecimiento de una línea de base**

Se debe establecer una línea de base antes de comenzar a hacer modelos de aprendizaje automático. Si los modelos que se construyen no pueden superar esta línea, entonces se tendrá que admitir que el aprendizaje automático no es el más adecuado para el problema.

Como voy a utilizar modelos de regresión, una buena línea de base es predecir el valor medio del objetivo en el conjunto de entrenamiento para todos los ejemplos en el conjunto de prueba. Esto es muy simple de implementar y establece una barra relativamente baja para nuestros modelos: si no pueden hacerlo mejor que adivinar el valor de la mediana, entonces habrá que repensar el enfoque para afrontar el problema.

- **Modelos de Aprendizaje Automático**

Voy a comparar 5 modelos diferentes de aprendizaje automático para cada una de las variables objetivo utilizando la biblioteca Scikit-Learn:

- **Linear Regression**
- **Support Vector Machine Regression (SVR)**
- **Random Forest Regression**
- **Gradient Boosted Regression**
- **K-Nearest Neighbors Regression (KNN)**

- **Ajuste de hiperparámetros con búsqueda aleatoria y validación cruzada**

Ya que estos modelos vienen dados con valores predeterminados, en un primer momento, voy a determinar el rendimiento de referencia de cada modelo y después voy a intentar optimizarlos mediante el ajuste de los hiperparámetros. De esta manera intentaré encontrar la mejor configuración para cada uno de los 3 casos de predicción.

Para elegir los mejores hiperparámetros para cada modelo voy a usar la **Búsqueda Aleatoria** (*Random Search*) y la **validación cruzada** (*Cross Validation*). Con la **búsqueda aleatoria** se define un rango de opciones y luego se selecciona al azar combinaciones para probar. Después con la **validación cruzada** se evalúa el rendimiento de los hiperparámetros. En lugar de dividir el conjunto de entrenamiento en conjuntos de entrenamiento y validación, lo que reduciría la cantidad de datos de entrenamiento que se pueden usar, se utiliza el **K-Fold Cross Validation** que divide los datos de entrenamiento en K pliegues y luego pasa por un proceso iterativo donde primero se entrena en K-1 de los pliegues y luego se evalúa el rendimiento en el k-ésimo pliegue. El proceso se repite K veces. Al final de la **K-Fold Cross Validation**, se toma el error promedio en cada una de las K iteraciones como la medida de rendimiento final y luego se entrena el modelo en todos los datos de entrenamiento a la vez.

La **búsqueda aleatoria** va a reducir los posibles parámetros a intentar ya que reduce el rango de opciones. Los resultados de esta búsqueda se podrían usar para hacer una **búsqueda de cuadrícula** (*Grid Search*) creando una cuadrícula con los hiperparámetros que mejor han funcionado en la

búsqueda aleatoria. En lugar de evaluar todos estos ajustes nuevamente, voy a enfocarme en: el **número de estimadores** (para Random Forest y Gradient Boosted), **número de vecinos** (KNN) y **parámetro de regulación C y de potencia p** (SVR).

- **Estudio del sobreajuste (*overfitting*)**

El modelo elegido para las 3 predicciones es el **Gradient Boosted**. Para cada caso, se ha representado el MAE en función del número de estimadores (árboles de decisión para el Gradient Boosted) manteniendo constante el resto de los parámetros.

En las tres predicciones, se observa que, a medida que aumenta el número de árboles utilizados por el modelo, disminuyen tanto el error de entrenamiento como el de prueba. Sin embargo, el error de entrenamiento disminuye mucho más rápidamente que el error de prueba y, por tanto, el **modelo está sobreajustado**: funciona muy bien en los datos de entrenamiento, pero no es capaz de lograr el mismo rendimiento en el conjunto de pruebas.

Para abordar el sobreajuste, he disminuido la complejidad del modelo reduciendo la profundidad máxima de cada árbol y en algún caso, aumentando también el número mínimo de muestras en un nodo de hoja.

6. Principales resultados

En esta sección se muestran los principales resultados que se han obtenido en las dos partes del proyecto. Los resultados detallados están disponibles al ejecutar el código en el repositorio.

6.1 Análisis y Caracterización de las aguas superficiales

(datos: chemical_WQI_ICI_stations.xlsx, mines.csv, cities.xls)

*(Notebook: **01_Part_1_Analysis_and_Characterization_of_surface_waters.ipynb**)*

- **Generalidades**

Las regiones de drenaje y los puntos de muestreo se representan en la siguiente figura:

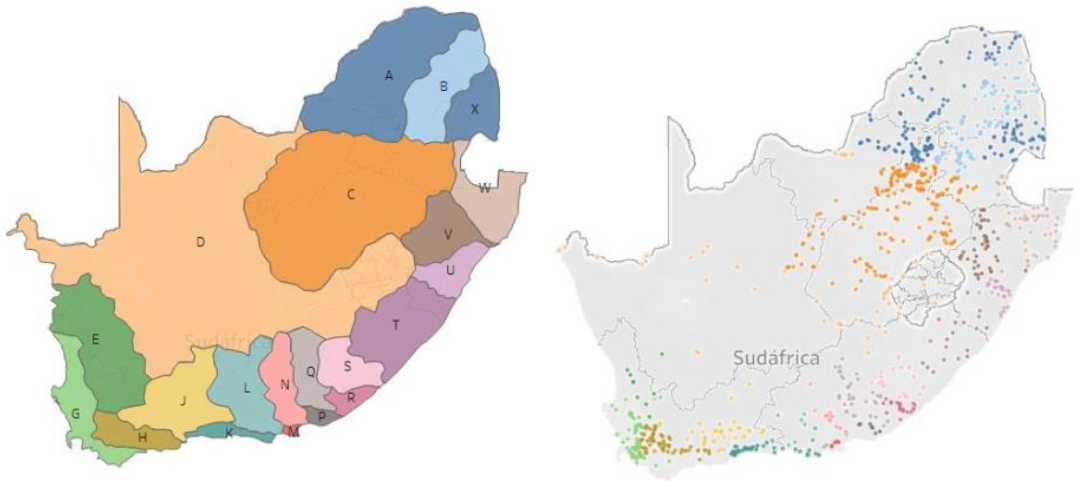
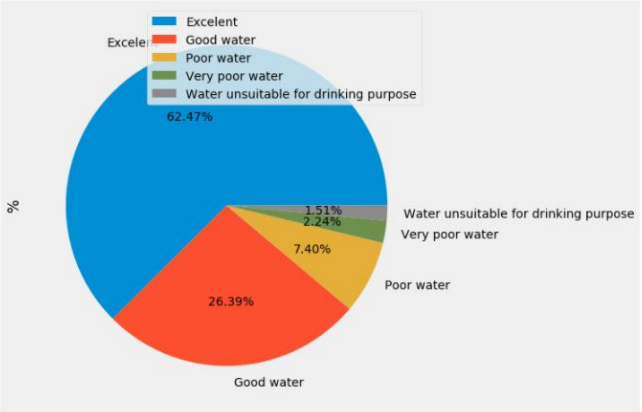
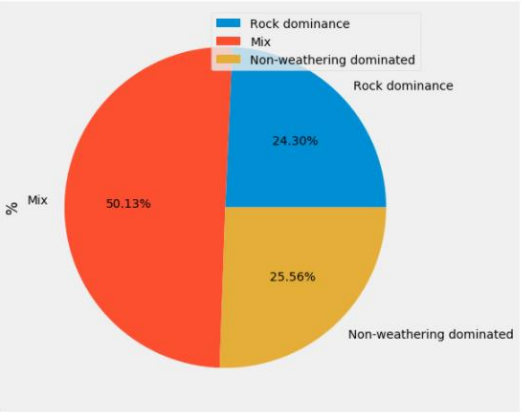


Figura 1. Regiones de drenaje (izquierda) y estaciones de muestreo (derecha)

Hay 21 regiones de drenaje y 1045 estaciones de ríos, embalses y lagos.

Estadísticas del WQI y del ICI



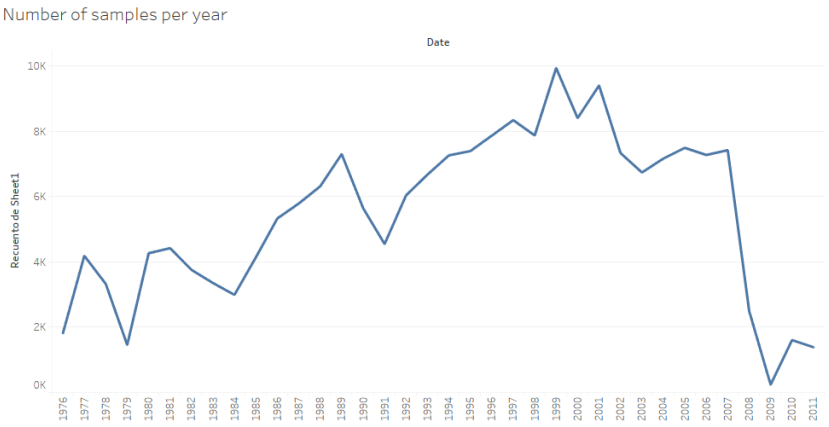


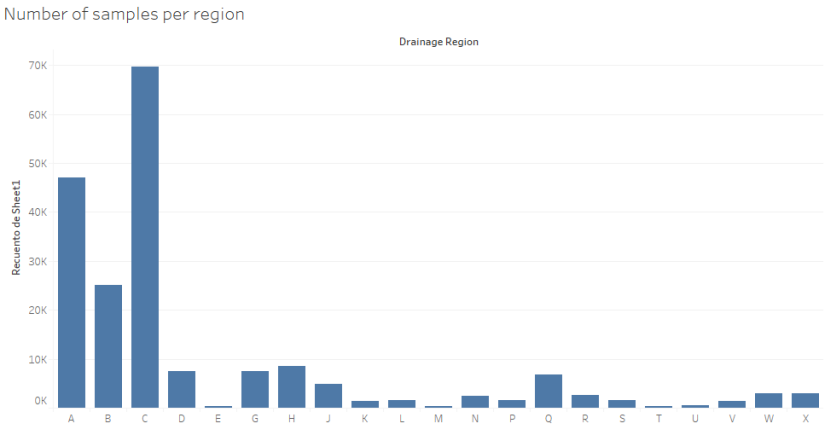
Apenas el **10%** de las muestras no tienen una buena calidad del agua.

En el **25%** de las muestras recogidas, la química está dominada por la intemperie de las rocas, el mismo porcentaje para aquellas cuya química está dominada por las fuentes alternativas. El **50%** restante es una combinación de ambas.

Sesgo temporal y espacial

A continuación, se muestran los sesgos:





Evoluciones temporales

Las siguientes gráficas muestran las evoluciones temporales en Sudáfrica de los parámetros de interés:

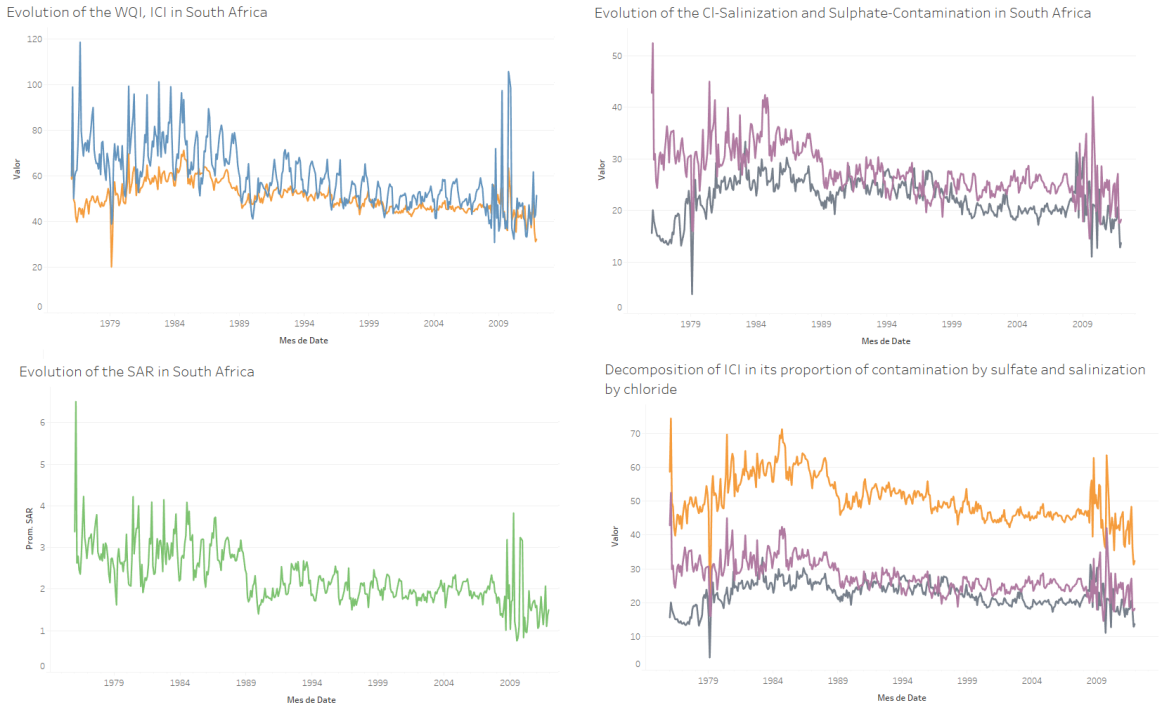


Figura 2. *Evolución de los parámetros en Sudáfrica: del WQI y del ICI (arriba a la izquierda), de la salinización por cloruro y la contaminación por sulfato (arriba a la derecha), del SAR (abajo a la izquierda) y descomposición del ICI (abajo a la derecha)*

Debido al sesgo temporal y espacial es difícil sacar conclusiones. La importancia de este sesgo se muestra en, por ejemplo, febrero de 1979, donde todas las evoluciones marcan el valor mínimo. Esto se debe a que en este mes sólo se realizó una medida, no se puede considerar correcto este valor. Del mismo modo, entre los años 2008 y 2010 las evoluciones de todos los parámetros tienen picos muy altos y muy bajos. Es un periodo en el que se realizaron muy pocas medidas, tampoco es posible sacar conclusiones fiables.

La conclusión que se puede sacar de estos gráficos es que **todos los parámetros siguen la misma evolución** a lo largo de los años: hay una relación evidente entre ellos.

Esta misma relación ocurre en cada región individualmente y en cada punto de muestreo.

Nota: Para completar las visualizaciones interactivas de las evoluciones de los parámetros por regiones y/o estaciones de medida, abrir el libro empaquetado de Tableau P*art_1_ general_visualizations*.

Promedio en las estaciones de medida del WQI, ICI y Salinización por cloruro y Contaminación por sulfato

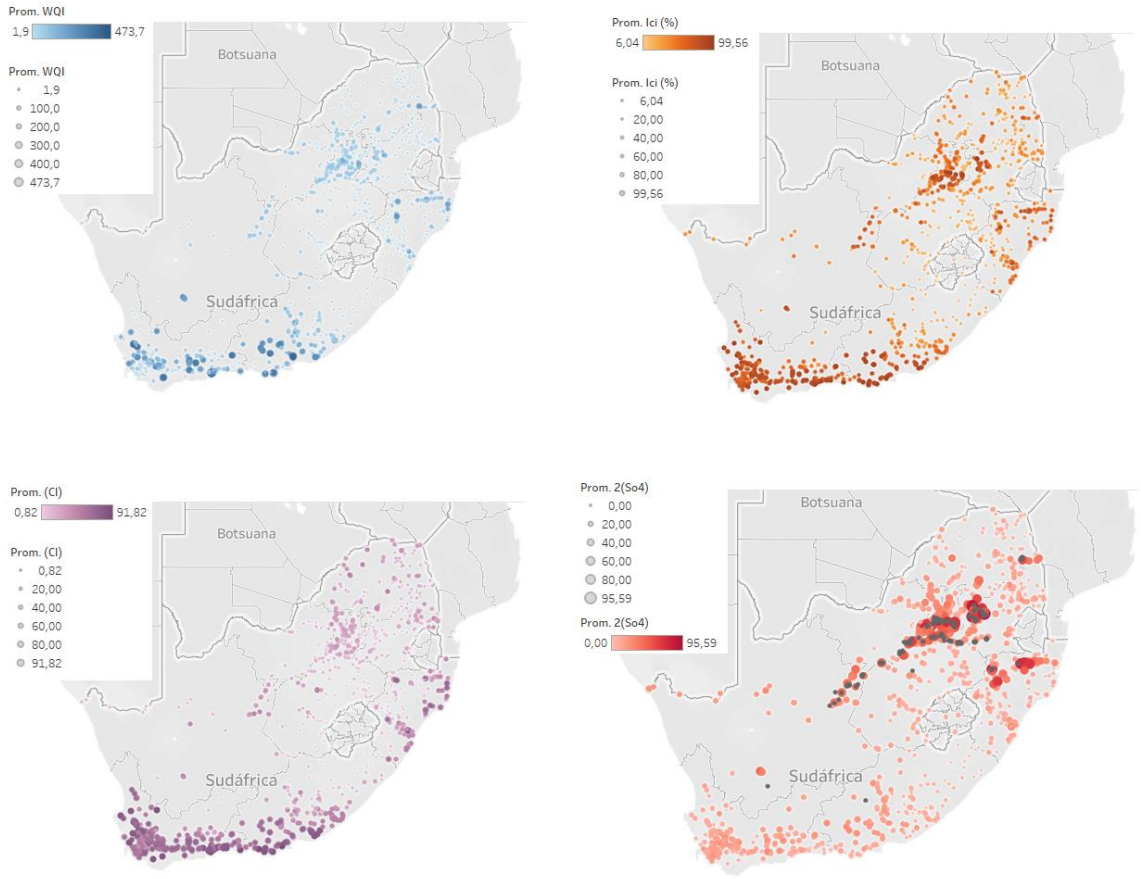


Figura 3. *Promedio de los parámetros en las diferentes estaciones de medida: del WQI (arriba a la izquierda), del ICI (arriba a la derecha), de la salinización por cloruro (abajo a la izquierda) y de la contaminación por sulfato (abajo a la derecha)*

A partir de la figura 3, se sacan las siguientes conclusiones:

- Las estaciones con peor media de WQI están situadas en el sur del país.
- Las estaciones más dominadas por fuentes alternativas a la meteorización de las rocas, están en el sur principalmente, y en la región de Guateng (región en la que siempre ha habido tradición de explotación minera y recoge numerosas ciudades con gran número de población).
- Las estaciones con mayor salinización por cloruro se encuentran en el sur del país.
- Las estaciones con mayor contaminación por sulfato están en la región de Guateng. En el mapa de abajo a la derecha de la figura se puede observar la posición de las minas que se encuentran muy próximas a los puntos de medida con mayor concentración de sulfatos.

- **Caracterización de las aguas**

Las regiones de Sudáfrica se clasifican en 3 grupos dependiendo de su Índice de Química Inorgánica:

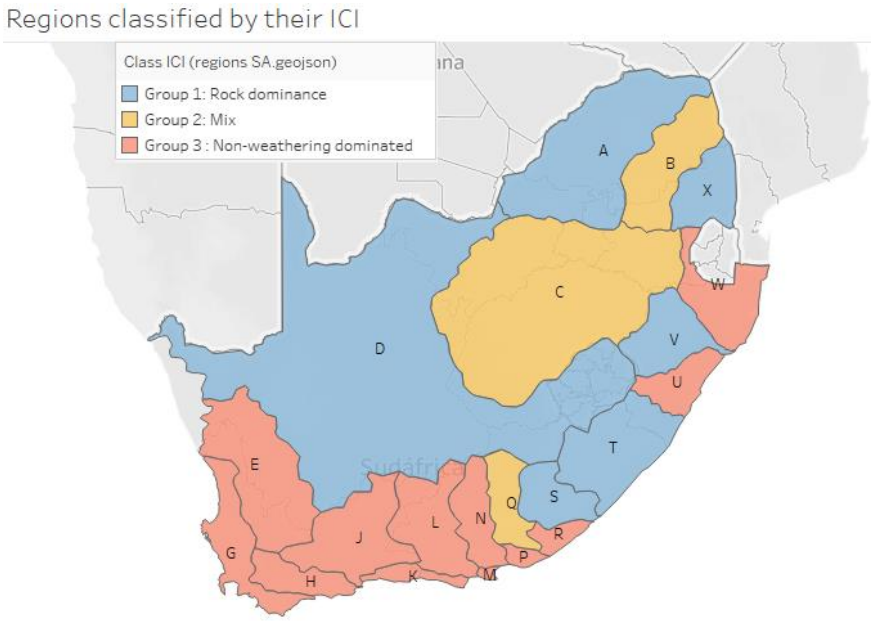
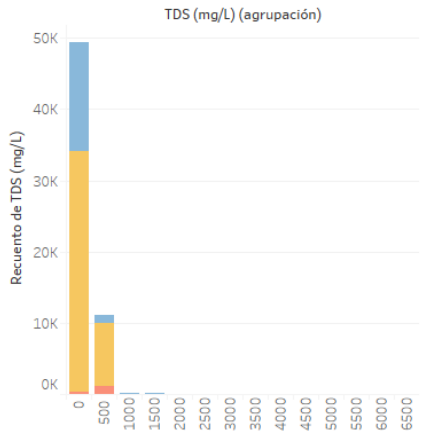


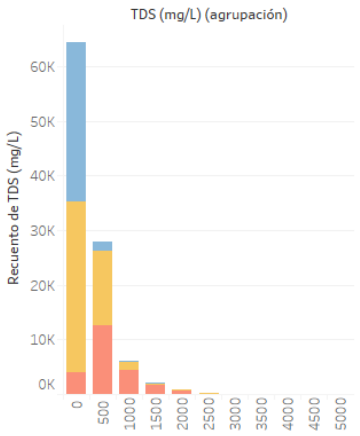
Figura 4. *Regiones de drenaje coloreadas por la clasificación según su ICI.*

A partir de esta clasificación se representan las características químicas para los 3 grupos diferentes:

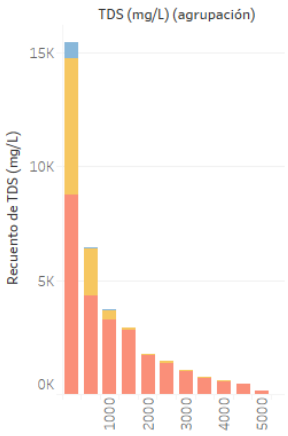
Group 1 TDS



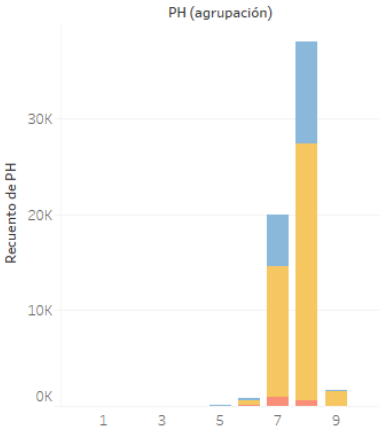
Group 2 TDS



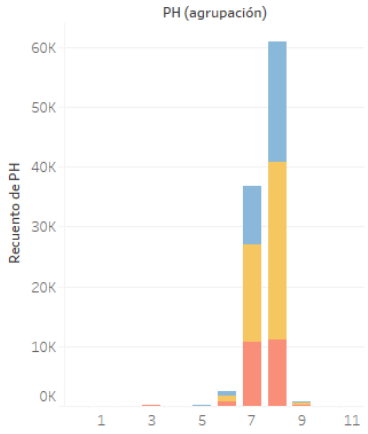
Group 3 TDS



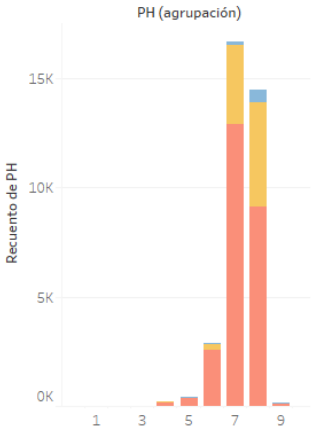
Group 1 PH



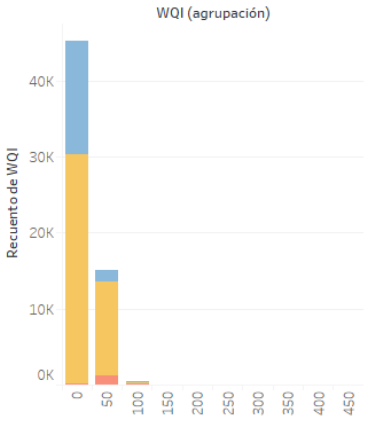
Group 2 PH



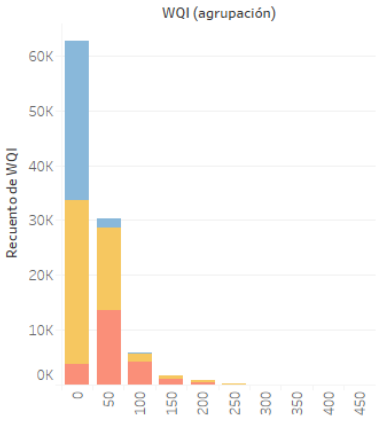
Group 3 PH



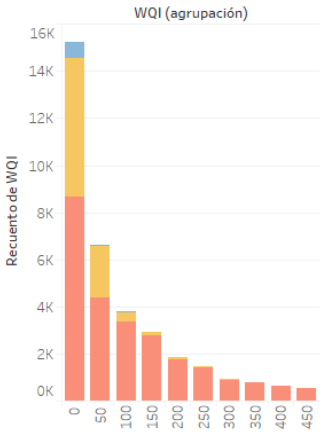
Group 1 WQI



Group 2 WQI



Group 3 WQI



Grupo 1

Las propiedades químicas de sus aguas se caracterizan por estar **dominadas por la intemperie química** o por combinación de geología y fuentes alternativas. Mayoritariamente las muestras recogidas tienen buena calidad y pocas sobrepasan los límites establecidos.

Grupo 2

Aparecen un número considerable de medidas con valores superiores a los ideales de TDS y WQI que corresponden con aquellas muestras dominadas por fuentes alternativas a la meteorización química. La concentración de sulfato es mayor que la de cloruro en comparación con el grupo 1, y aporta más contribución al ICI.

En estas regiones es frecuente la extracción de carbón y oro, y las **altas concentraciones de sulfato** están asociadas precisamente al **drenaje ácido de las minas** de estos minerales.

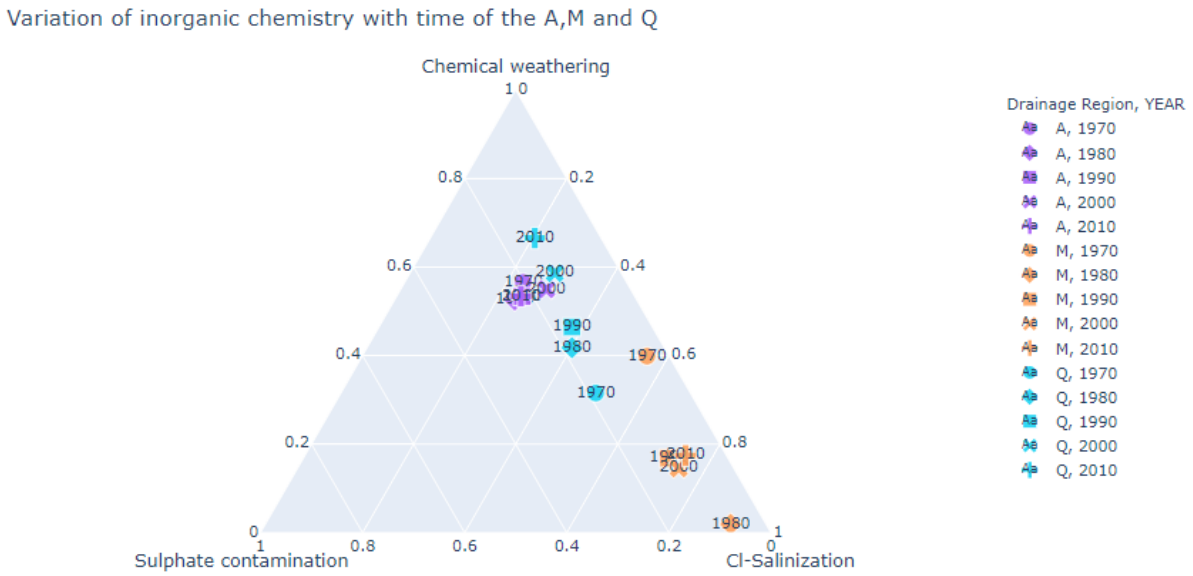
Grupo 3

El número de muestras dominadas por **fuentes alternativas** es mayoritario. Un gran número de medidas superan los límites establecidos para la óptima calidad del agua. El contenido de TDS y los valores de WQI son muy variables. El pH muestra una tendencia a valores más bajos (aguas más ácidas). Los valores de ICI son generalmente altos y están relacionados con una alta c**ontribución de cloruro**.

Las concentraciones elevadas de cloruro son causadas principalmente por la salinización natural (todas son regiones costeras) pero también es causada por el riego y la eliminación de vegetación natural.

- **Diagramas ternarios**

Estos diagramas son interesantes para ver la variación química en el tiempo de las regiones de drenaje. A continuación, tres ejemplos, la variación de las regiones A, M y Q:

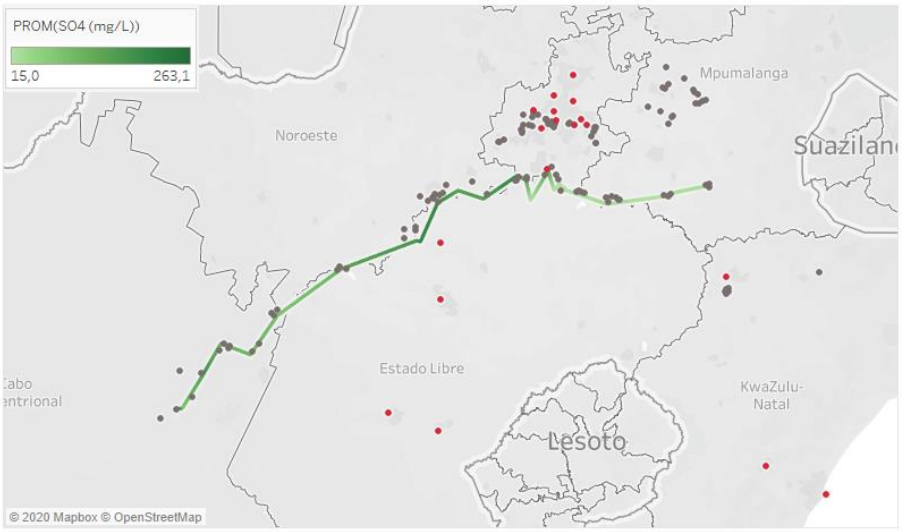


A partir de este diagrama, se puede concluir que:

- La química de las aguas superficiales del área de captación A permaneció relativamente contante.
- En la región M, la química de las aguas superficiales cambió significativamente: muestra un aumento de salinización por cloruro en la década de 1980, seguida de una disminución de la salinización en las siguientes décadas hasta quedar más o menos constante.
- La química de las aguas superficiales del área de captación Q ha ido evolucionando progresivamente desde un estado de salinización por cloruro hacia un estado dominado por la meteorización de las rocas.

- **Río Vaal**

El río Vaal es el más importante de Sudáfrica ya que suministra agua al corazón económico del país. Principalmente recorre la región de drenaje C, perteneciente al grupo 2, y cuyas aguas tienden a tener concentraciones de sulfato más altas de las permitidas. Como se ha visto anteriormente, esto se debe en gran parte al drenaje ácido de las minas. La siguiente representación, muestra las estaciones de medida en este río, las posiciones de las minas que extraen actualmente y la evolución del sulfato en el río:



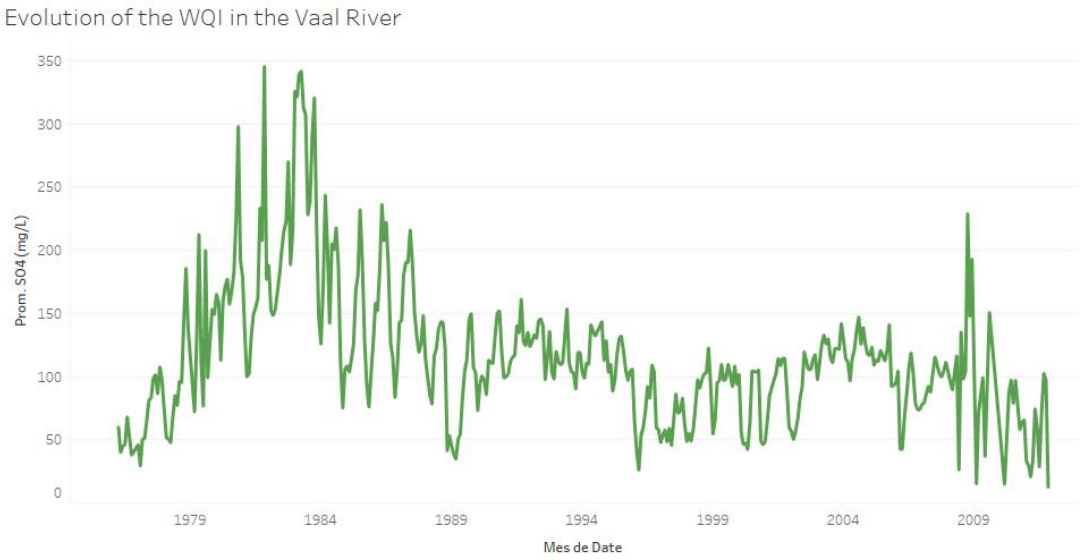


Figura 5. *Río Vaal con tramos coloreados según su promedio en sulfato y posiciones de las minas (gris) y ciudades más importantes (rojo) (arriba) y evolución del sulfato en el río (abajo).*

Se esperaba obtener unas concentraciones mucho más elevadas de sulfato a lo largo del río, ya que hay numerosas minas en sus alrededores. Sin embargo, no es así, lo que podría deberse, por un lado, al sesgo de datos (temporal y espacial), y, por otro lado, al hecho de que las minas representadas son las actuales (datos del 2019) y los datos con los que se trabaja son de décadas anteriores. Hay más información gráfica del río Vaal en el archivo *Part_1_Vaal_river.tbwx.*

6.2 Predicción del Índice de Calidad del agua

(datos: chemical_WQI_ICI_stations.xlsx)

*(Notebook: **02_Part_2.1_Prediction_of_the_Water_Quality_Index.ipynb**)*

En esta sección, he realizado 3 predicciones utilizando diferentes variables predictivas:

1. Conductividad eléctrica, PH, Fosfato (PO_4), Silicio (Si), Amonio (NH_4) y la latitud.
2. Conductividad eléctrica, PH y Latitud.
3. Conductividad eléctrica.

Las predicciones con las variables de punto 1., las he realizado para saber si la introducción de los elementos químicos en los modelos aporta mejora en los mismos. Para el cálculo del WQI se necesitan los valores de las concentraciones de 9 elementos, de esta manera utilizaríamos la concentración de 4, por lo que, a priori, incluso utilizando todas estas variables, supondría una reducción de costos (a falta de saber si la obtención de las concentraciones de estos elementos es más fácil y económica que de los otros 9 necesarios) y sería útil siempre y cuando las predicciones que obtengamos sean suficientemente buenas.

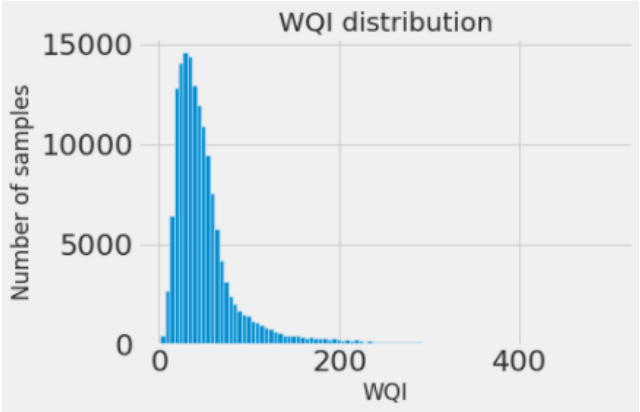
1. Tratamiento de variables

- **Elección de la variable predictiva de posición**

Hay 4 variables en el dataset que contienen información sobre las posiciones de los puntos de muestreo: "*Drainage Region*", "*Subdrainage Region*", "*POINT ID*", "*Latitude (degrees)*" y "*Longitude (degrees)".* Introducir en los modelos predictivos más de una de estas variables no tendría sentido pues se estaría dando información redundante y las variables serían colineales entre sí.

La elección de las coordenadas geográficas como variables predictivas para representar la posición de las estaciones de medida la he hecho basándome en que son variables numéricas y no categóricas como las demás, que supondría codificarlas (proceso *one-hot encoding*) e introducir en el modelo muchas más variables (una para cada categoría diferente) y por tanto ralentizaría el proceso.

- **Distribución de la variable objetivo**



- **Gráficas de densidad**

Gráfica de densidad del WQI en función de las coordenadas geográficas

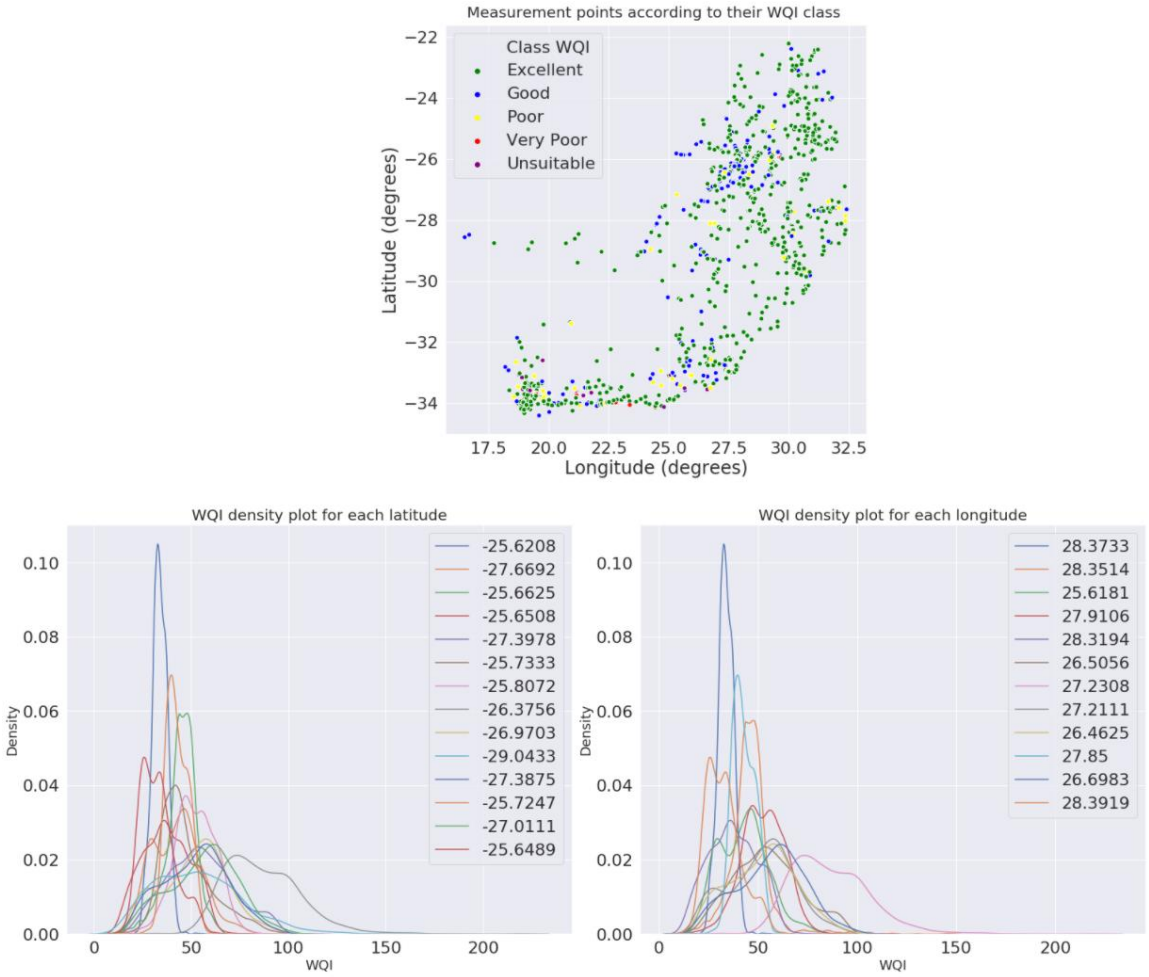


***Figura 6**. Puntos de medida coloreados según su clase de WQI en el mapa (arriba), función de densidad del WQI en función de las latitudes y las longitudes de los puntos de muestreo con más muestras realizadas (izquierda y derecha respectivamente)*

En la figura 6 se muestra que la gráfica de densidad del Índice de Calidad del Agua sí que cambia según la latitud o la longitud de las estaciones de medida. Por tanto, introducir estas variables puede aportar información y mejora en los modelos predictivos y voy a considerarlas como variables de entrada.

Gráfica de densidad del WQI en función de las clases del ICI

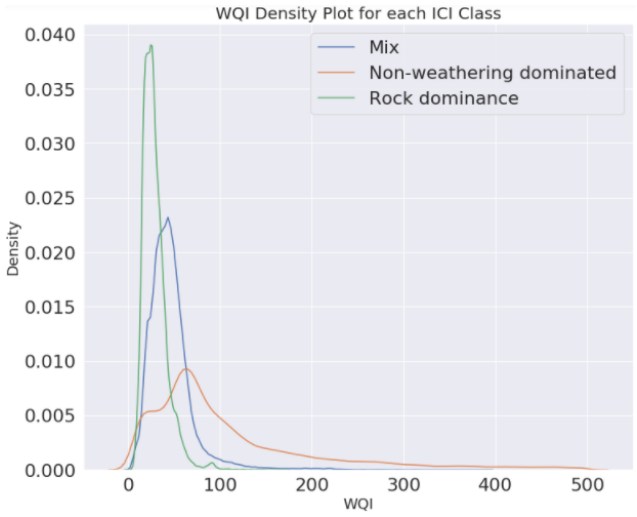


Figura 7. *Gráfica de densidad del WQI en función de las clases del ICI.*

En la figura 7 se ve la relación entre el WQI y el ICI. Cuanto peor es la calidad del agua según su valor del Índice de Calidad, mayor influencia tienen las fuentes alternativas a la meteorización de las rocas en el dominio de la química del agua.

Sin embargo, introducir esta variable categórica en los modelos no tiene sentido, puesto que precisaríamos de las medidas de los 5 elementos químicos a partir de los cuales se calcula el ICI.

- **Correlación entre las variables predictivas y la variable objetivo**
  - Correlación entre WQI y EC

La correlación entre el WQI y la conductividad eléctrica es del **0.989628**. Esta correlación tan alta se muestra gráficamente en la figura 5 y evidencia la fuerte relación lineal que hay entre ambas variables.

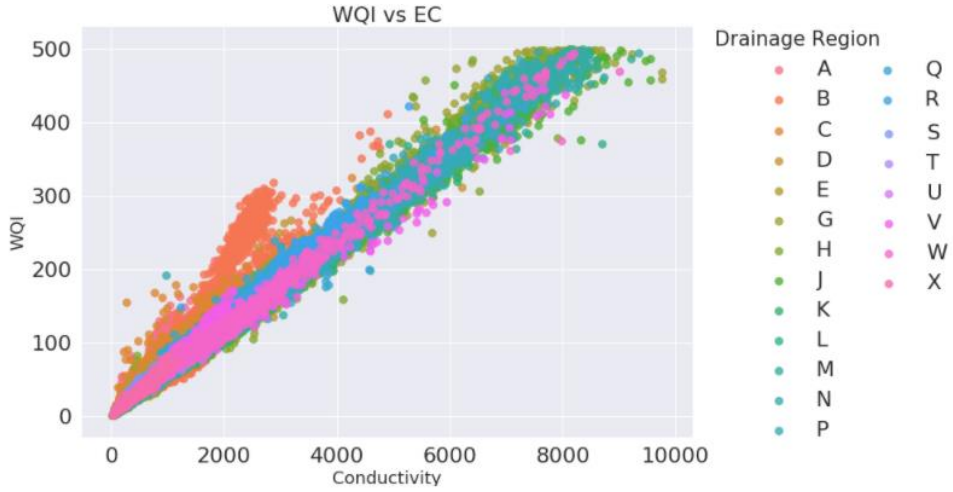
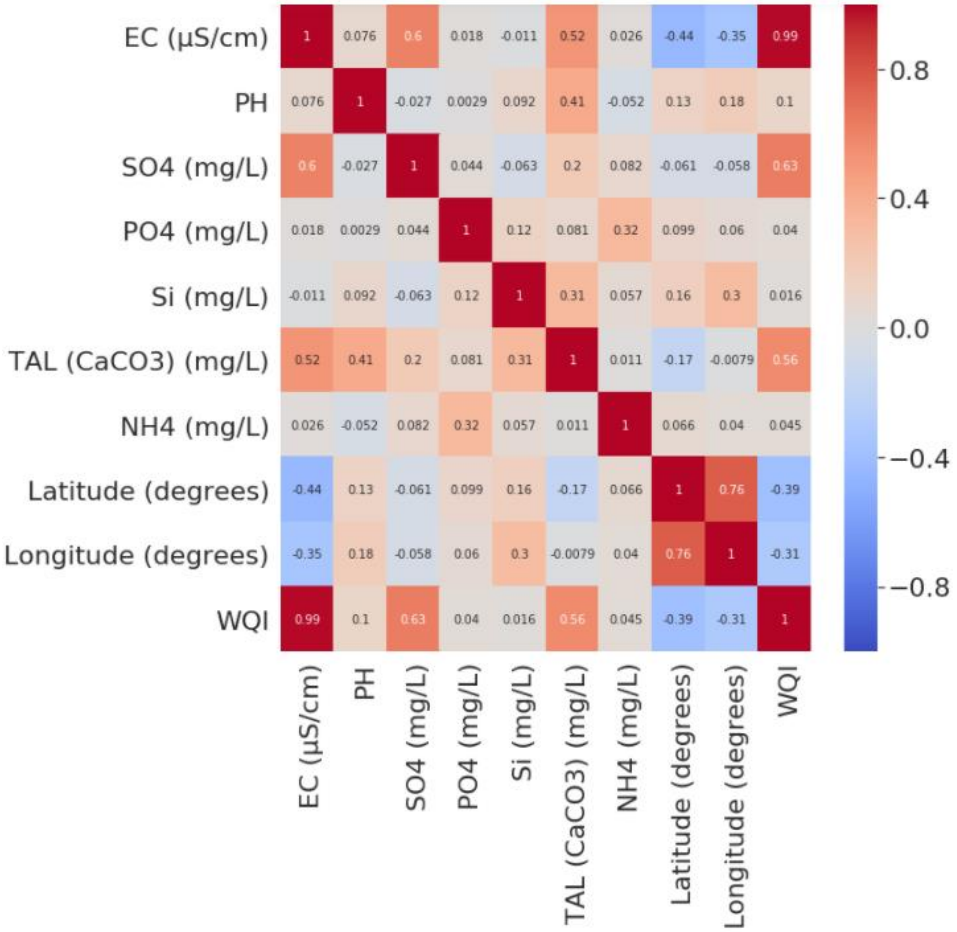


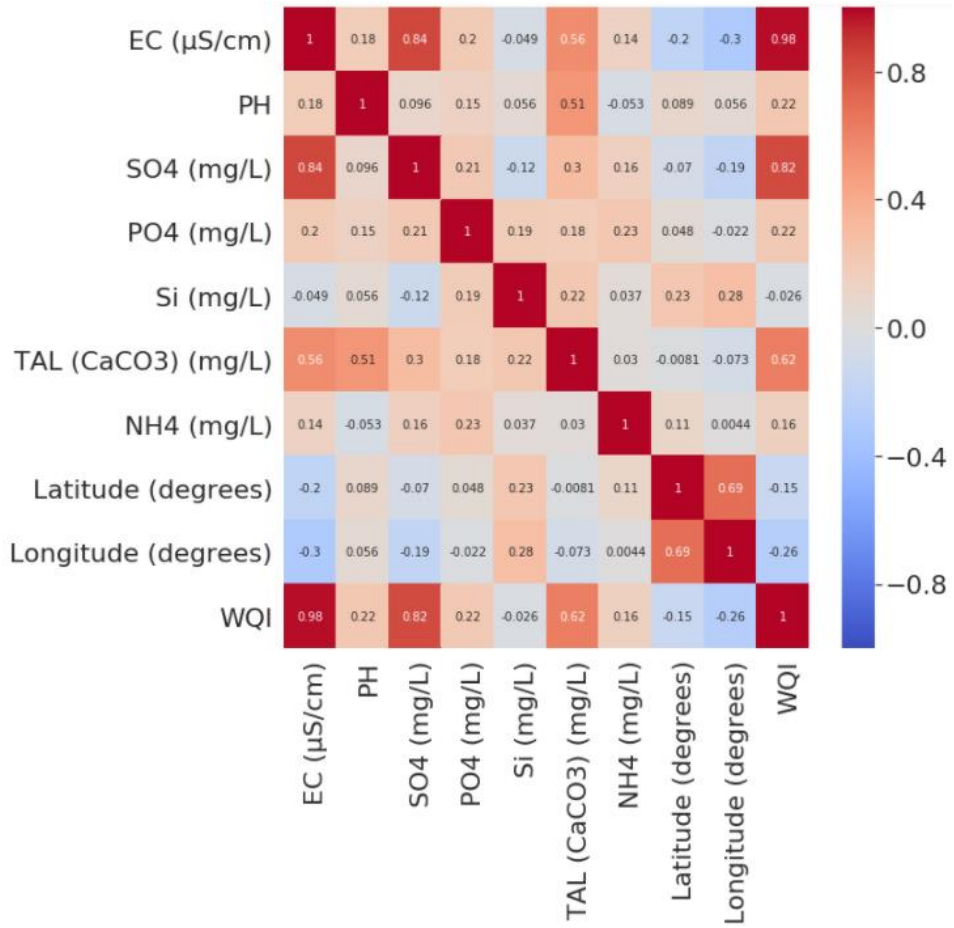
Figura 8. *Representación gráfica entre el Índice de Calidad del Agua y la conductividad eléctrica.*

- Matrices de correlación

Matriz de correlación lineal



Matriz de correlación de "spearman"

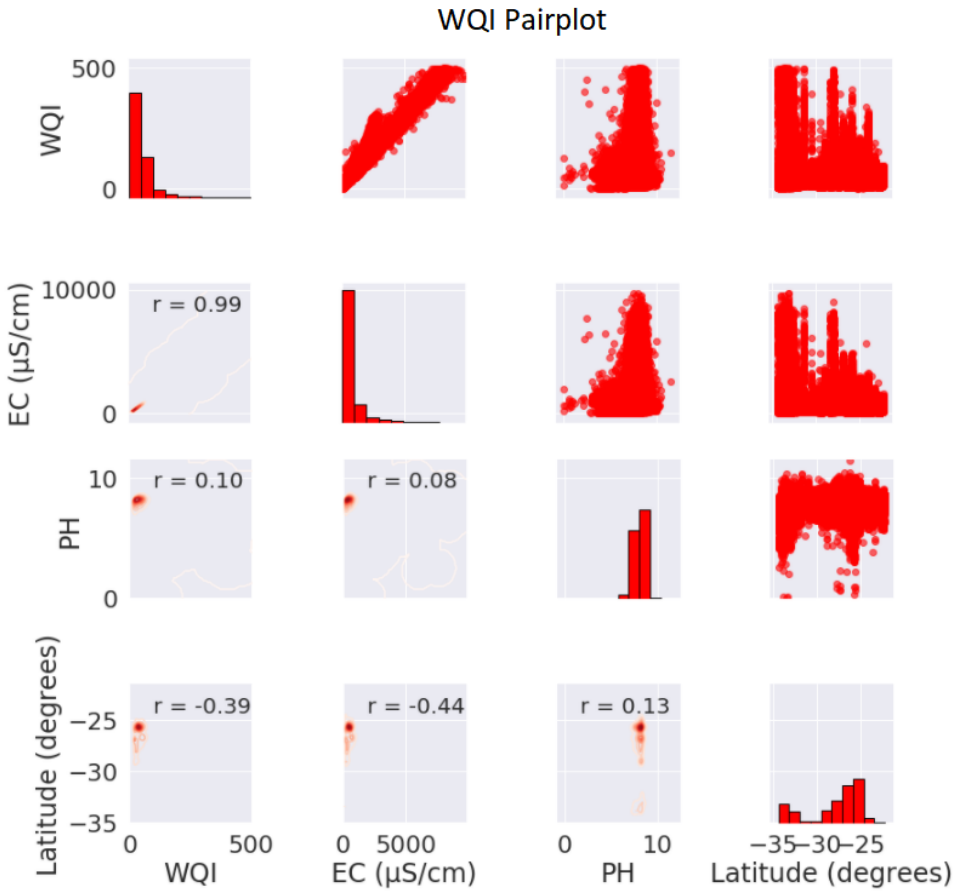


Las matrices de correlación muestran que el WQI tiene una **relación lineal muy fuerte** con la conductividad eléctrica y relaciones no lineales altas con el "SO_4" y el "$CaCO_3$". También mejoran las correlaciones no lineales con el "*PH*", el "PO_4" y el "NH_4".

Se deben eliminar aquellas variables colineales para evitar el "***overfitting***" de la predicción. El "SO_4" tiene una relación muy alta con la conductividad eléctrica y también la "*Longitud*" con la "*Latitud*", por lo que descartamos estas dos variables como variables de entra en los modelos predictivos.

- **Pairplots**

Utilizando el diagrama de pares se pueden examinar varias variables a la vez: **diagramas de dispersión** (triángulo superior), **histogramas** (diagonal) y el **coeficiente de correlación** entre dos variables como una estimación de la densidad del núcleo en 2D de las dos variables (triángulo inferior). He representado los diagramas de pares entre la variable objetivo y las variables de entrada: "*PH*", "*Conductividad eléctrica*" y "*Latitud*".



2. Línea de base

La línea de base que se obtiene es de una puntuación de **41.47**, y la estimación promedio en el conjunto de prueba es de **31.4240**. Los valores de la variable objetivo están entre 0 y 300, lo que significa que el error promedio inicial es de aproximadamente del **10.5%**. Los modelos deben superar esta línea de base que es bastante baja.

3. Resultados de las métricas

A continuación, se resume en una tabla los resultados obtenidos en los modelos:

Variables de entrada: Conductividad eléctrica, PH, PO4, Si, NH4 y Latitud			
Modelo predeterminado	**MAE**	**MAPE**	**Tiempo de ejecución**
Regresión Lineal	4.5624		
SVR	3.9411		
Random Forest	**2.9030**	**91.49%**	**1min 30s**
Gradient Boosted	3.5914		
KNN	3.7404		
Modelo ajustado			
Random Forest	2.9249	91.43%	13min 34s
Variables de entrada: Conductividad eléctrica, PH y Latitud			
Modelo predeterminado	**MAE**	**MAPE**	**Tiempo de ejecución**
Regresión Lineal	4.6112	86.49%	29.6ms

SVR	4.1336	87.89%	28min 43s
Random Forest	3.0983	90.93%	47.5s
Gradient Boosted	3.6533	89.3%	11s
KNN	3.4475	89.9%	213ms
Modelo ajustado			
Random Forest	3.0245	91.14%	2min 42s
Gradient Boosted	**2.7006**	**92.09%**	**4min 15s**
KNN	3.3872	90.08%	239ms
SVR	3.7406	89.04%	48min 1s
Variables de entrada: Conductividad eléctrica			
Modelo predeterminado	**MAE**	**MAPE**	**Tiempo de ejecución**
Regresión lineal	5.0711		
Random Forest	4.7301	86.14%	13.5s
Gradient Boosted	3.6533	89.3%	12.1s
KNN	5.1846	84.81%	80.5ms
SVR			
Modelo ajustado			
Random Forest	4.6238	86.45%	2min 11s
Gradient Boosted	3.4962	89.67%	36.4s
KNN	4.7714	86.02%	85ms

Todos los modelos superan a la línea de base, por lo que el aprendizaje automático sí que es apropiado para el problema.

4. Elección del modelo final

Considero que, para predecir el Índice de Calidad del Agua, las variables de entrada y el modelo más adecuados son:

- **Variables: EC, PH y Latitud**

  Porque son variables con un método de medición rápido y sencillo que mejora los resultados que utilizando únicamente la conductividad eléctrica, y aunque se obtenga resultados un poco peores que introduciendo la concentración de los elementos químicos (PO4, Si, NH4), compensa frente al tiempo y coste que supondría tener que medirlos.

- **Modelo: Gradient Boosted**

  El modelo Gradient Boosted ajustado obtiene el mejor MAE y MAPE (2.7006 y 92.9% respectivamente). Supone una mejora del **0.95%** con respecto al siguiente mejor modelo, el Random Forest, pero a coste de un tiempo de ejecución significativamente mayor, ya que es aproximadamente el doble (4min 15s frente a 2min 42s). Sin embargo, considero que la magnitud absoluta del tiempo de entrenamiento no es significativa, por lo que voy a elegir como modelo más adecuado el **Gradient Boosted**.

5. Estudio del sobreajuste

La representación de la figura 8 muestra que **el modelo está sobreajustado**. La siguiente tabla, recoge los resultados obtenidos reduciendo la profundidad máxima de cada árbol:

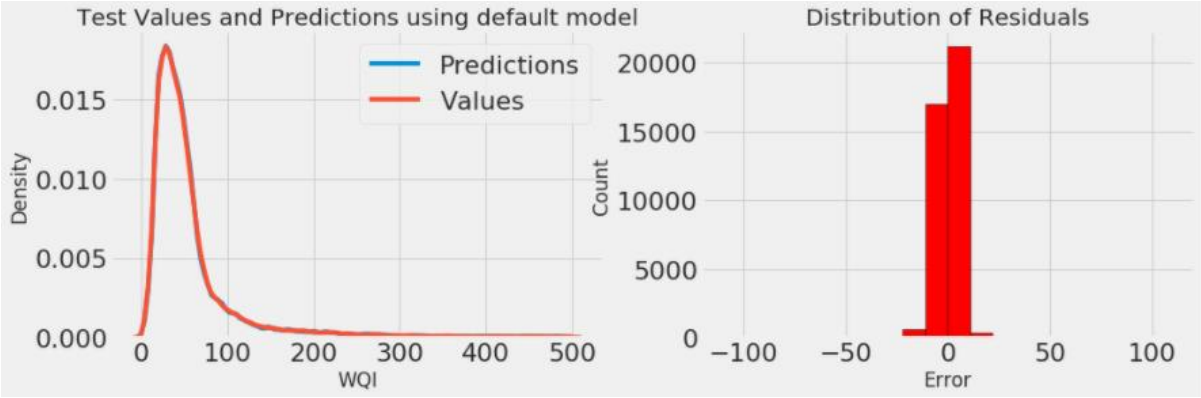
Modelos Gradient Boosted a diferentes profundidades						
Opción	**Max_depth**	**Trees**	**Min_samples_leaf**	**MAE**	**MAPE**	**Tiempo de ejecución**
Elegido en el apartado anterior	10	550	6	2.7006	92.09%	5min 15s
1	5	550	10	2.8085	91.77%	2min 17s
2	3	550	10	3.1229	90.85%	1min 41s
3	7	550	10	2.6951	92.1%	3min 23s

Figura 9. *Representaciones del MAE en función del número de árboles, de derecha a izquierda: modelo elegido en el punto anterior, opción 3, opción 2 y opción 1.*

La opción 1 puede ser la más adecuada. Pierde un **0.32%** respecto al modelo Gradient Boosted elegido en el apartado anterior, pero se ha reducido mucho el sobreajuste (la diferencia entre el error de entrenamiento y el error de prueba no es significativa).

6. Distribución de los valores reales y predicciones

Para tener una idea de las **predicciones**, voy a trazar la distribución de los valores reales y los valores pronosticados en el conjunto de prueba en el modelo final (Gradient Boosted opción 1 del apartado anterior) y la distribución de los residuos:



La distribución de valores pronosticados es la misma a la de los valores verdaderos del conjunto de prueba. Las predicciones son más que correctas.

El **histograma de residuos** no tiene una distribución normal (que sería lo ideal, al significar que el modelo se equivoca la misma cantidad de veces en ambas direcciones (alta y baja)). En este caso, el modelo se equivoca más pronosticando valores más bajos a los verdaderos.

6.2 Predicción del Índice de Química Inorgánica

(datos: chemical_WQI_ICI_stations.xlsx)

*(Notebook: **03_Part_2.2_Prediction_of_the_Inorganic_Chemistry_Index.ipynb**)*

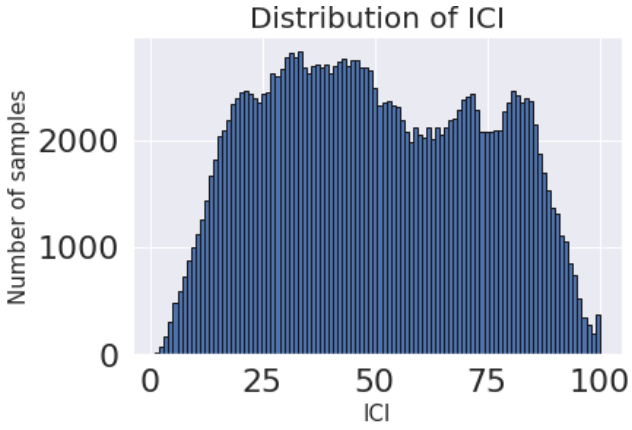
Siguiendo el mismo procedimiento que en el punto 5.4:

1. Tratamiento de variables

- **Elección de variables de entrada**

En este caso directamente voy a tratar con las variables de entrada: **Conductividad eléctrica**, **PH** y **latitud**.

- **Distribución de la variable objetivo**



- **Gráficas de densidad**

Gráfica de densidad del ICI en función de la latitud

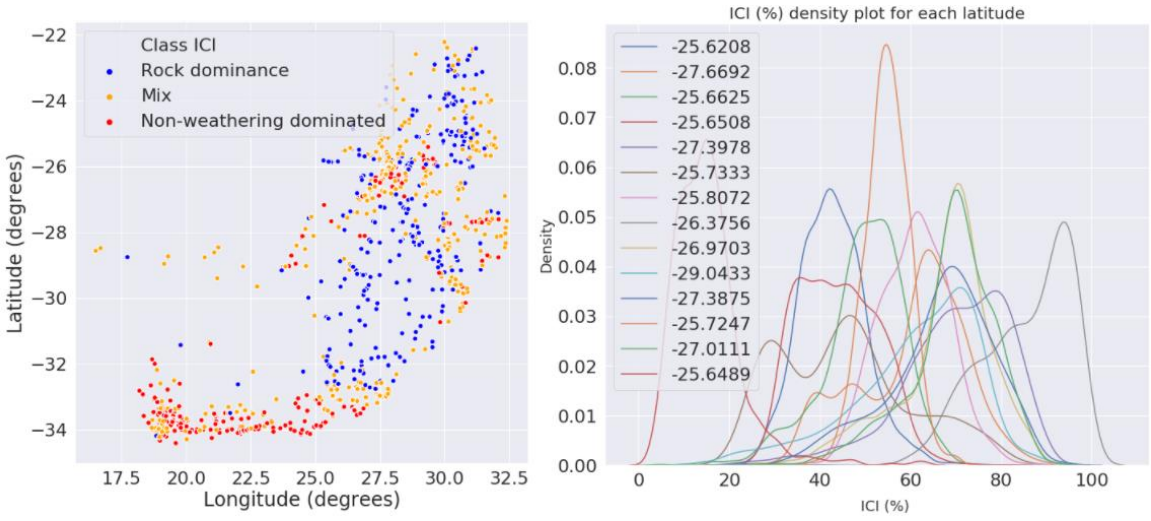
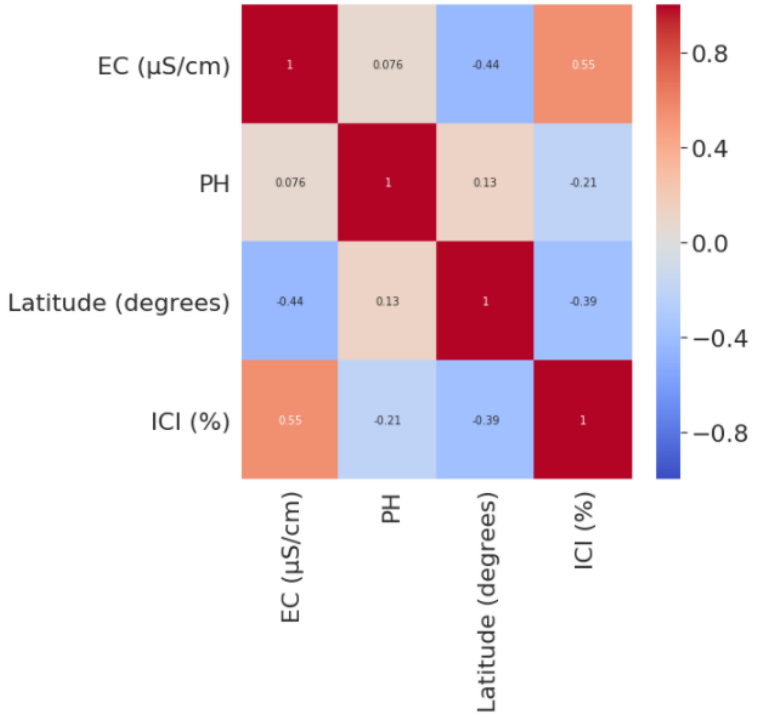


Figura 10. *Mapa con las estaciones de muestreo coloreadas por su clase ICI (izquierda) y gráficas de densidad del ICI en función de la latitud en las estaciones con más medidas (derecha)*

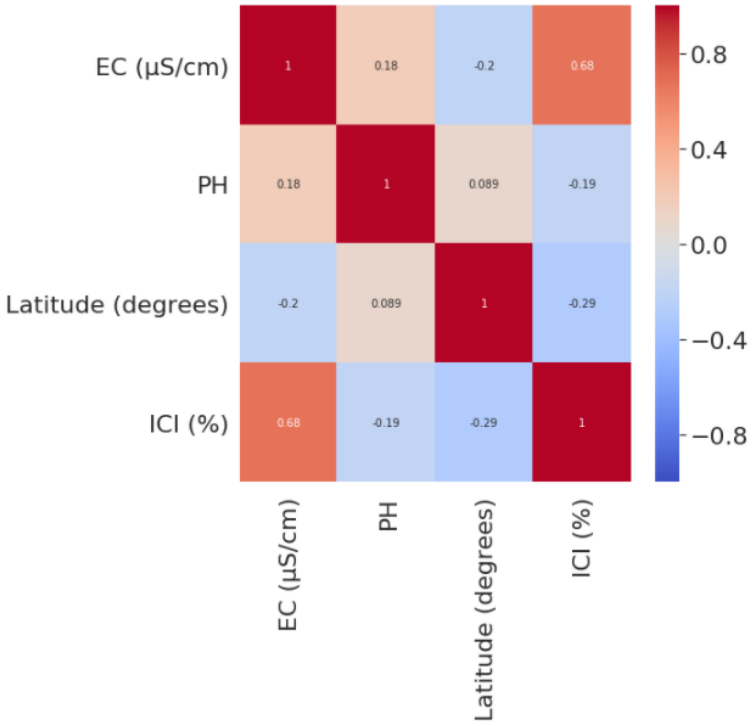
Las gráficas de densidad del ICI muestran que la latitud tiene un efecto significativo sobre el ICI, por lo que es una buena variable de predicción que puede aportar información útil para determinar el ICI.

- **Correlación entre las variables predictivas y la variable objetivo**

Matriz de correlación lineal

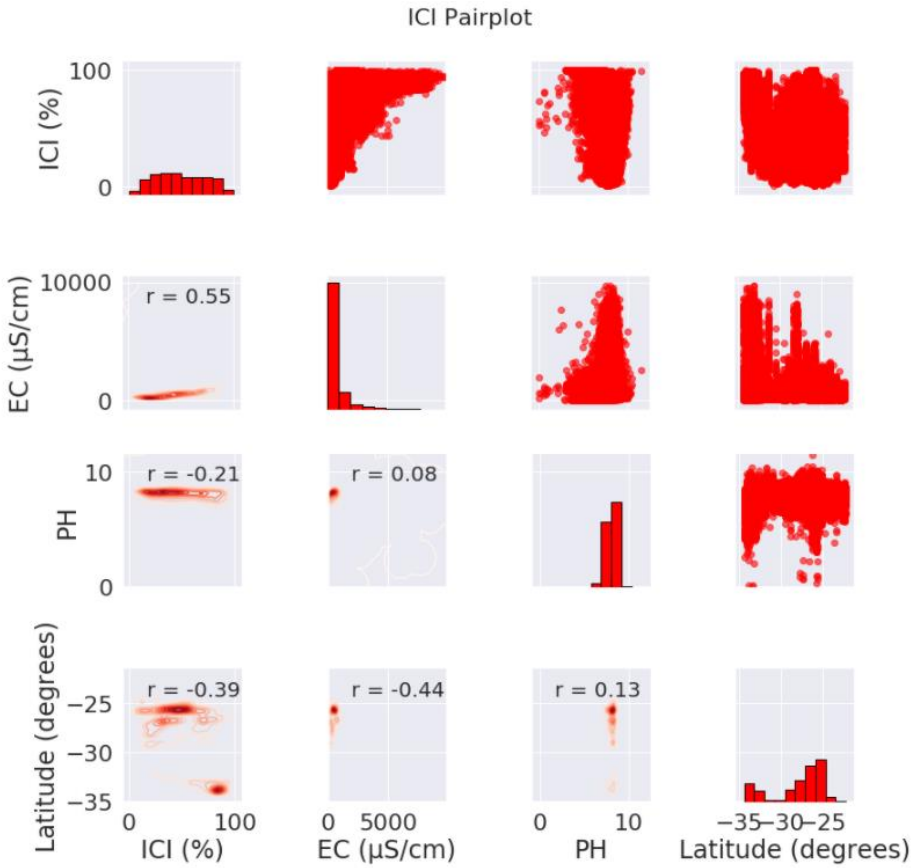


Matriz de correlación de "spearman"



La variable objetivo, ICI, tiene una correlación no lineal alta con la conductividad eléctrica, y está correlacionada linealmente con el PH y la latitud.

- **Pairplots**



2. Línea de base

La línea de base es **48.39** y la estimación promedio en el conjunto de prueba tiene una puntuación de **20.3886**. Los valores de la variable objetivo están entre 0 y 100, por lo que el error promedio inicial es del **20%**. Esta línea de base no es demasiado baja, pero veremos si los modelos la superan.

3. Resultados de las métricas

Los resultados de los modelos se recogen en la siguiente tabla:

Variables de entrada: Conductividad eléctrica, PH y Latitud			
Modelo predeterminado	**MAE**	**MAPE**	**Tiempo de ejecución**
Regresión Lineal	15.7531	55.14%	45.7ms
SVR	11.2654		
Random Forest	4.4235	87.4%	46.7s
Gradient Boosted	8.5268	75.72%	12.1s
KNN	6.7137	80.88%	133ms
Modelo ajustado			
Random Forest	4.3580	87.59%	6min 41s
Gradient Boosted	**4.0875**	**88.36%**	**6min 40s**
KNN	6.3910	81.8%	122ms

Aunque todos los modelos superan a la línea de base, la regresión lineal no obtiene buenos resultados.

Nota: No se ha calculado el SVR ajustado porque se consume mucho tiempo en lanzarlo y no es muy probable que mejore a los otros modelos teniendo en cuenta que el MAE de su modelo predeterminado está lejos de ser más pequeño.

4. Elección del modelo final

El mejor modelo para el problema es el **Gradient Boosted ajustado** ya que supera en 0.77% al modelo Random Forest ajustado y los tiempos de ejecución son los mismos. Con el modelo **Random Forest predeterminado** el MAPE disminuye en **0.96%** a coste de un tiempo de ejecución significativamente menor (8.5 veces más rápido). Sería otro modelo a considerar.

5. Estudio del sobreajuste

El sobreajuste del modelo se representa en la figura 11. La tabla que se muestra a continuación, recoge los resultados obtenidos al reducir la profundidad máxima de cada árbol, incluso en la opción 4 he aumentado el número mínimo de muestras en un nodo de hoja:

Opción	Max_depth	Trees	Min_samples_leaf	MAE	MAPE	Tiempo de ejecución
Elegido en el apartado anterior	10	750	6	4.0875	88.36%	6min 40s
1	5	750	10	4.5595	87.02%	5min 19s
2	3	750	10	5.8179	83.43%	2min 18s
3	4	750	12	5.0313	85.67%	2min 48s
4	7	750	10	4.1902	88.07%	4min 28s

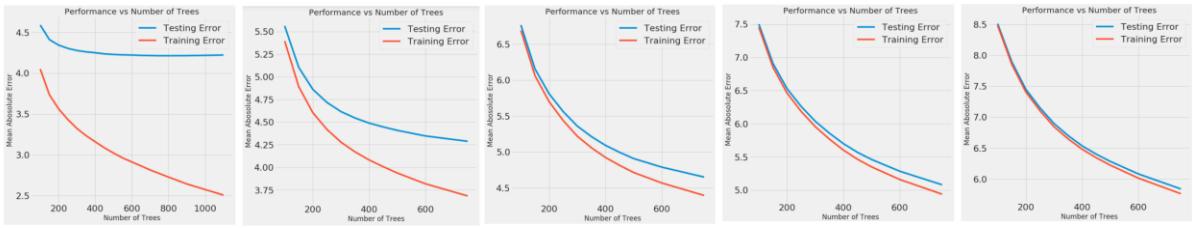
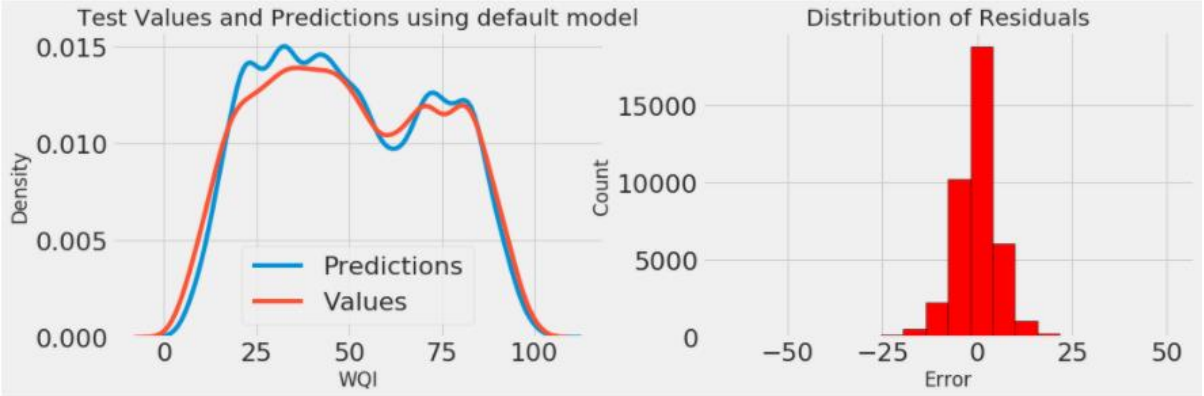


Figura 11. *Representación del MAE en función del número de árboles de derecha a izquierda: modelo elegido en el apartado anterior, opción 4, opción 1, opción 3 y opción 2.*

El modelo con los parámetros de la opción 4 tiene un MAPE **0.29%** más bajo que el modelo elegido en el apartado anterior. Se pierde precisión, pero se disminuye considerablemente el sobreajuste.

6. Distribución de los valores reales y predicciones



La distribución de valores pronosticados es muy similar a la de valores verdaderos del conjunto de prueba. El modelo podría ser menos preciso para predecir valores intermedios pronosticando valores más altos o más bajos a los que corresponden, y en cambio, las predicciones son más precisas para valores más cercanos a los extremos, aunque en esta ocasión los valores pronosticados son más pequeños a los valores verdaderos.

6.3 Predicción de la Relación de Adsorción del Sodio

(datos: chemical_WQI_ICI_stations.xlsx)

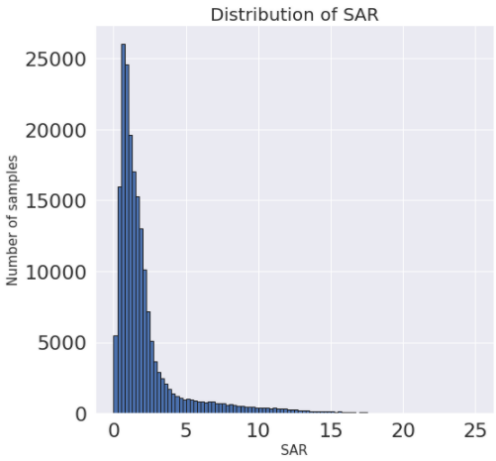
*(Notebook: **04_Part_2.3_Prediction_of_the_Sodium_Adsorption_Ratio.ipynb**)*

1. Tratamiento de variables

- **Elección de variables de entrada**

Las variables de entrada que se usan para la predicción: **Conductividad eléctrica**, **PH** y **latitud**.

- **Distribución de la variable objetivo**



- **Gráficas de densidad**

Gráfica de densidad del SAR en función de la latitud

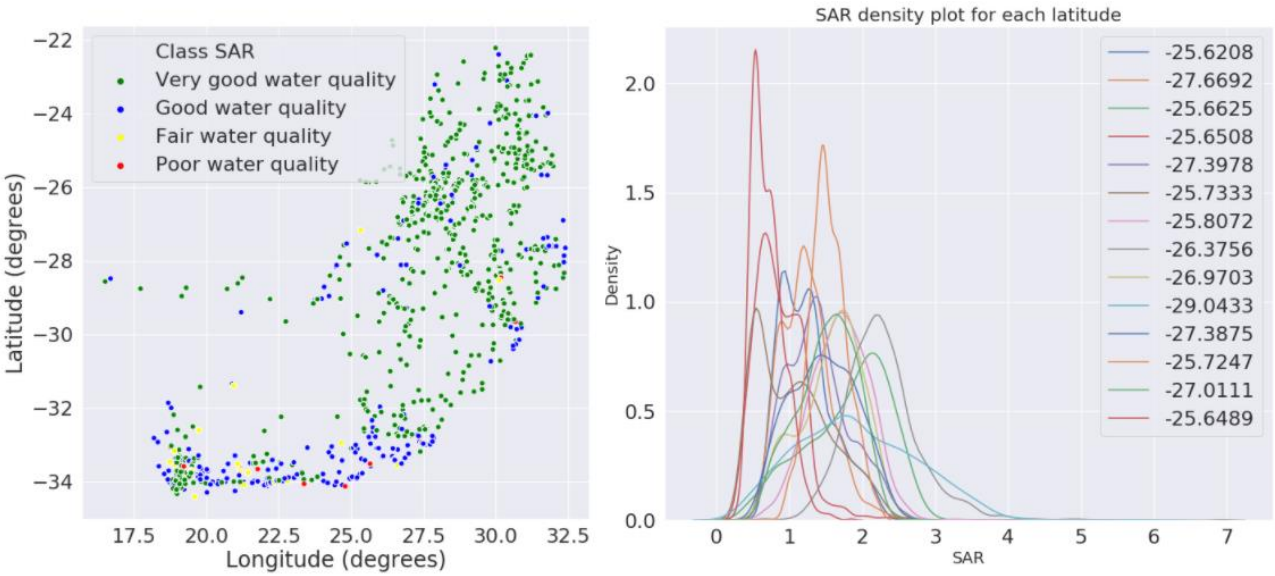


Figura 12. *Mapa con las estaciones de muestreo coloreadas por su clase SAR (izquierda) y gráficas de densidad del SAR en función de la latitud en las estaciones con más medidas (derecha)*

La latitud puede aportar información útil para la predicción del SAR según se muestra en las gráficas de densidades de la figura 12.

- **Correlación entre las variables predictivas y la variable objetivo**
  - Correlación entre SAR y EC

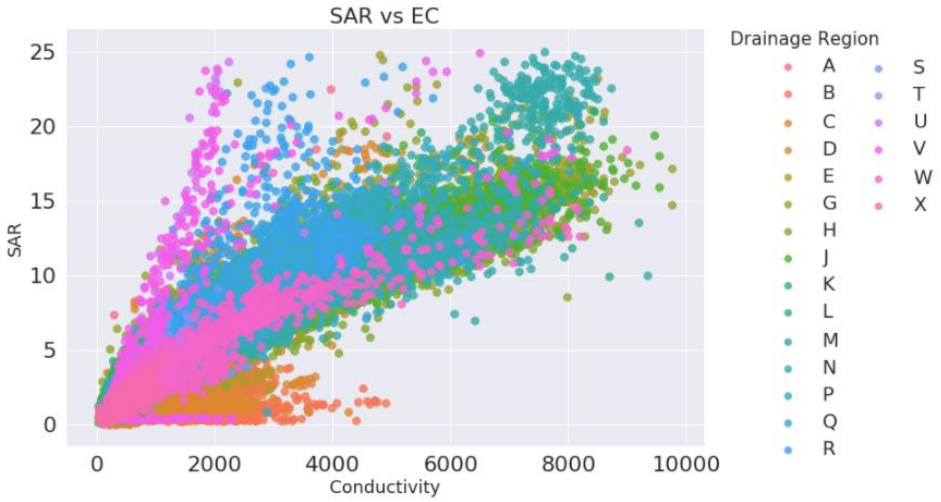
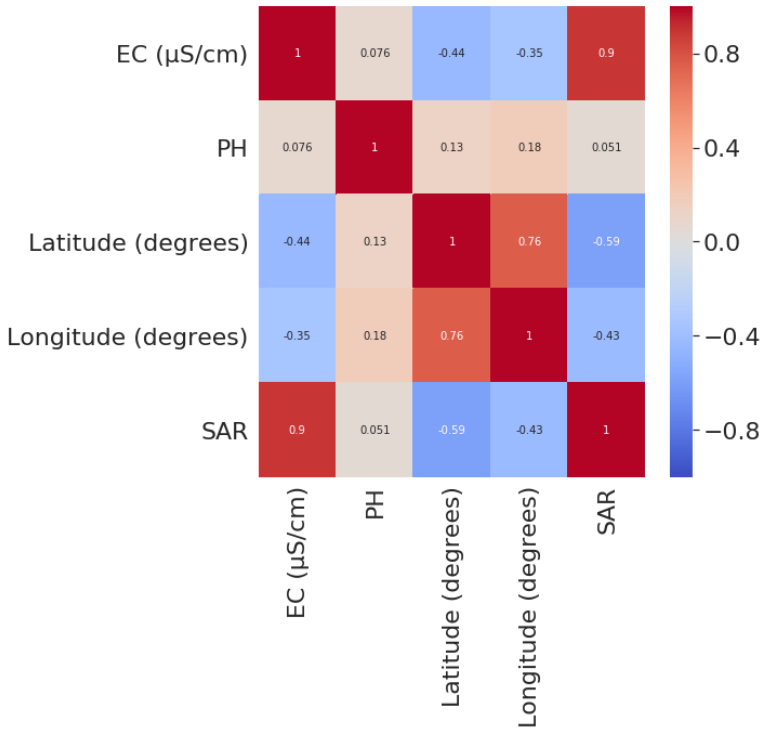


Figura 13. *Representación gráfica entre la Relación de Adsorción del Sodio y la conductividad eléctrica.*

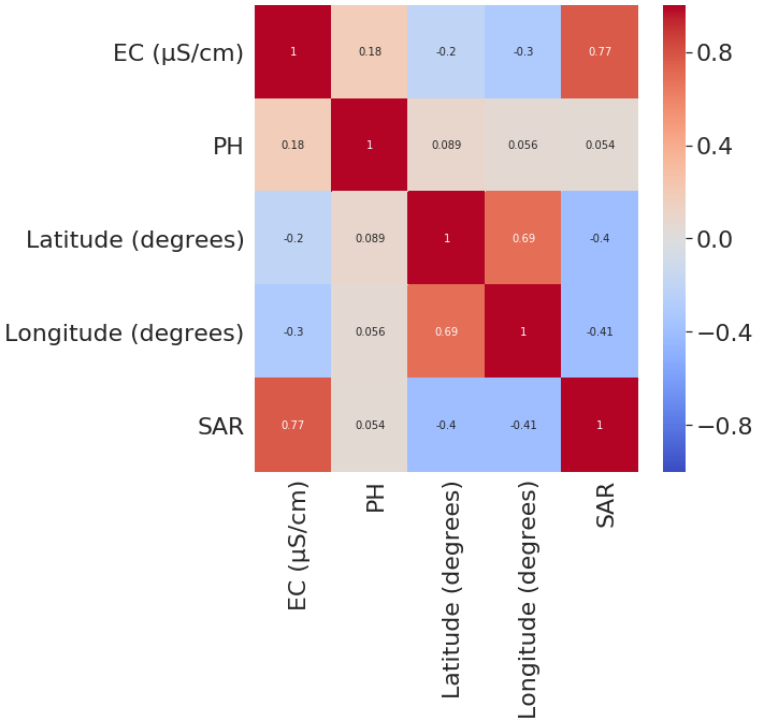
La correlación lineal entre el SAR y la conductividad eléctrica es muy alta: **0.898258**. Se puede ver esta relación en la figura 13.

- Matrices de correlación

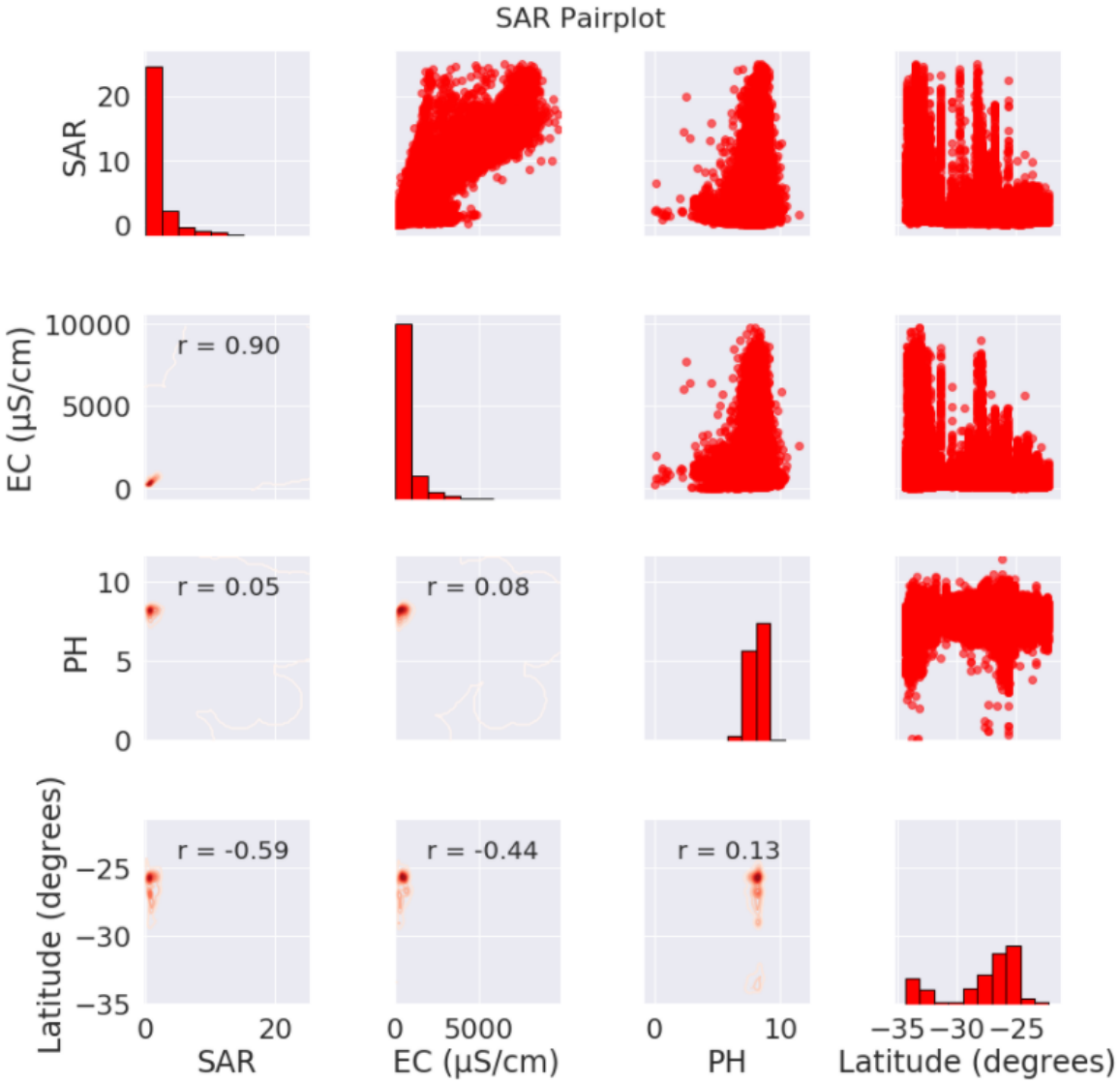
Matriz de correlación lineal



Matriz de correlación de "spearman"



- **Pairplots**



2. Línea de base

La línea de base tiene una puntuación de **1.35** y la estimación promedio en el conjunto de prueba es **1.4766**. El SAR tiene unos valores comprendidos entre 0 y 25, por lo que el error promedio inicial es del **6%**. Es una línea de base muy baja.

3. Resultados de las métricas

Los resultados para los diferentes modelos se muestran en la siguiente tabla:

Variables de entrada: Conductividad eléctrica, PH y Latitud		
Modelo predeterminado	**MAE**	**Tiempo de ejecución**
Regresión Lineal	0.6513	88.1ms
SVR	0.5361	
Random Forest	0.2498	57.1s
Gradient Boosted	0.4322	11.2s
KNN	0.3670	130ms

Modelo ajustado		
Random Forest	0.2462	4min 38s
Gradient Boosted	**0.2263**	**6min 51s**
KNN	0.3489	121ms
SVR		

Todos los modelos superan a la línea de base.

Nota: No se han incluido valores del MAPE porque al ser errores de tan pequeña magnitud, su resultado es –inf%.

Nota: En este caso tampoco se ha calculado el modelo SVR ajustado por el mismo motivo que en la predicción anterior.

4. Elección del modelo final

Una vez más voy a elegir el modelo **Gradient Boosted** ajustado que tiene el MAE más bajo pues las diferencias en los tiempos de ejecución no las considero significativas. Otro modelo a considerar sería el Random Forest predeterminado, ya que, aunque no tiene un MAE tan bajo, el tiempo de ejecución es mucho menor.

5. Estudio del sobreajuste

En este caso, no sólo he reducido la profundidad de cada árbol, si no que he también he aumentado el número mínimo de muestras en un nodo de hoja:

Opción	Max_depth	Trees	Min_samples_leaf	MAE	Tiempo de ejecución
Elegido en el apartado anterior	10	900	6	0.2263	6min 51s
1	7	900	10	0.2412	4min 1s
2	5	900	10	0.2461	3min 15s
3	3	900	10	0.3087	2min 13s

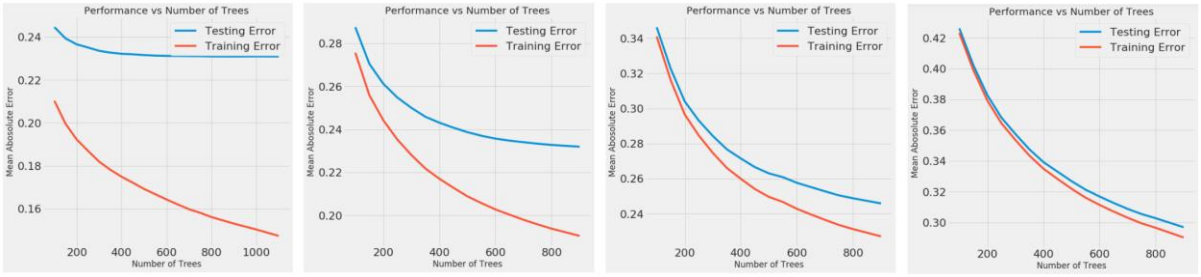
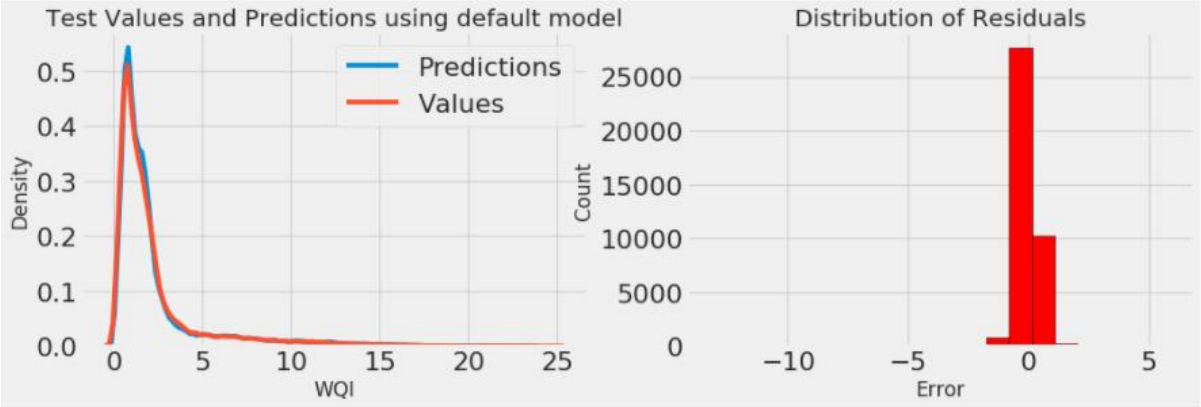


Figura 13. *Representación del MAE en función de los árboles, de derecha a izquierda: modelo elegido en el punto anterior, opción 1, opción 2 y opción 3.*

El modelo con los parámetros de la opción 2 reduce drásticamente el sobreajuste a cambio de empeorar el MAE en un **9%**. Aun así, considero que es la mejor opción.

6. Distribución de los valores reales y predicciones



La distribución de los valores pronosticados tiene la misma forma que la distribución de los valores verdaderos en el conjunto de prueba. Sin embargo, el modelo parece ser menos preciso para los valores cercanos a la mediana que predice valores mayores a los reales.

7. Conclusiones

Los puntos más destacados de la **primera parte** del proyecto son:

- Basándonos en el Índice de Calidad del Agua calculado, la calidad del agua es buena en la mayor parte de las muestras recogidas.
- Las estaciones con peores WQI y salinización por cloruro, se encuentran en el sur del país.
- La región de Guateng contiene los puntos de medida con niveles de contaminación por sulfatos más altos.
- Las estaciones más dominadas por fuentes alternativas a la meteorización química, es decir, a la mano del hombre, están situadas al sur del país y en la región de Guateng.
- Las regiones de captación primaria se pueden clasificar según un Índice de Química Inorgánica en **3 grupos**. Las regiones pertenecientes a un mismo grupo, tiene unas características químicas comunes.
- Los parámetros: Índice de Calidad del Agua, Índice de Química Inorgánica, Relación de Adsorción de Sodio, Contaminación por Sulfato y Salinización por Cloruro, tienen las **mismas evoluciones a lo largo del tiempo**, tanto en el conjunto del país, como en las diferentes regiones y estaciones de muestreo.
- Los **diagramas ternarios** ayudan a monitorear la variación de la calidad del agua en el espacio tiempo. Hay regiones que han mantenido la química de las aguas superficiales relativamente constante, y otras, han sufrido grandes cambios.
- Debido al **sesgo temporal y espacial**, y a la falta de datos más recientes, no es posible sacar conclusiones sobre el efecto de las fuentes alternativas a la meteorización de las rocas sobre las aguas superficiales del país.

En la **segunda parte** del proyecto se plantea si es posible crear modelos de aprendizaje automático para inferir con precisión los parámetros WQI, ICI y SAR. A partir de los resultados obtenidos, se puede concluir que sí en los 3 casos. Los aspectos más relevantes de esta segunda parte son:

- Las distribuciones del WQI y del SAR están sesgadas hacia valores altos.
- La **conductividad eléctrica** es la variable más útil para determinar los 3 parámetros.
- La conductividad eléctrica, el PH y la Latitud están correlacionados positivamente con el WQI.
- El ICI tiene una correlación positiva con la conductividad eléctrica y una correlación negativa con el PH y la Latitud.
- El SAR está correlacionado positivamente con la conductividad eléctrica y el PH y negativamente con la Latitud.
- Se puede construir un modelo **Gradient Boosted** sin que esté sobreajustado que predice:
    - el WQI con un MAE de **2.8085** y un MAPE del **91.77%.**
    - el ICI con un MAE de **4.1902** y un MAPE del **88.07%**.
    - el SAR con un MAE de **0.2461.**

El siguiente paso del proyecto sería introducir los datos de la química inorgánica del agua de los últimos años y crear una interfaz en la que el usuario únicamente tuviera que introducir las mediciones de las 3 variables predictivas: conductividad eléctrica, PH y Latitud y que devolviera la predicción del Índice de Calidad del Agua, del Índice de Química Inorgánica y de la Relación de Adsorción del Sodio.

7. Referencias

[1] Jan Marten Huizenga, Michael Silberbauer, Rainier Dennis, Ingrid Dennis (2013) " *An inorganic water chemistry dataset (1972–2011) of rivers, dams and lakes in South Africa* " . Water SA vol.39 n.2 Pretoria Jan.2013.

[2] P. Krishnakumar, C. Lakshumanan, V. Pradeep Kishore, V. Pradeep Kishore, G. Santhiya, G. Santhiya (2013). "*Assessment of groundwater quality in and around Vedaraniyam, South India*". Springer-Verlag Berlin Heidelberg 2013.

[3] JM Huizenga (2011). "*Characterisation of the inorganic chemistry of surface waters in South Africa*". African Journals Online.

[4] Hohls, BC, Silberbauer, MJ, K¸hn, AL, Kempster (2002) "*National water resource quality status report: inorganic chemical water quality of surface water resources in SA*". Departmen of Water Affairs and Forestry, Institute for Water Quality Studis. Private Bag X313, Pretoria 0001.

Softwares utilizados:

Mapa Regiones de Drenaje: http://geojson.io/

Coordenadas geográficas de las minas: https://mine-alert.oxpeckers.org/