[**Entrega 2: Descripción del Dataset y Origen**](https://politecnico.ar/campus/mod/assign/view.php?id=4470)

**Introducción Dataset**

En esta segunda etapa presentaremos los data sets que será utilizado para el proyecto de Aprendizaje Automático. El objetivo es clasificar la calidad del agua mediante variables físico-químicas y microbiológicas.

**Descripción del Dataset**

El conjunto de datos contiene valores obtenidos por el laboratorio sobre diversos parámetros de calidad del agua, a partir de muestras recolectadas a lo largo de la costanera del Río de la Plata, en la provincia de Buenos Aires, Argentina. La recolección y el análisis de los datos fueron llevados a cabo por el Ministerio de Ambiente y Desarrollo Sostenible, en colaboración con los municipios que integran la Red de Intercambio de Información de los Gobiernos Locales (RIIGLO).

Los datos utilizados en este proyecto abarcan el período comprendido entre los años 2013 y 2024, están distribuidos en doce archivos CSV, cada uno correspondiente a diferentes campañas de monitoreo. Estos archivos fueron obtenidos de plataformas públicas como Kaggle y CiAM. Cabe destacar hasta el momento, este conjunto de datos no ha sido explorado por la comunidad de Kaggle, lo que representa una oportunidad valiosa para realizar un análisis exploratorio de datos.

**Diccionario del dataset**

Los seis data sets fueron unificados en un único DataFrame que contiene 30 columnas y 1186 registros. Dado que el diccionario de datos no estaba disponible en las páginas, fue necesario realizar un análisis detallado de cada columna para interpretar correctamente su significado y tipo de variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **DICCIONARIO DE DATOS** | | | |
| **Data Set** | agc\_y\_riodelaplata\_2013.csv | **Periodo** | 2013 al 2024 |
| agc\_y\_riodelaplata\_2014.csv |
| agc\_y\_riodelaplata\_2015.csv |
| agc\_y\_riodelaplata\_2016.csv |
| agc\_y\_riodelaplata\_2017.cvs |
| agc\_z\_riodelaplata\_2018.csv |
| agc\_z\_riodelaplata\_2019.csv |
| agc\_z\_riodelaplata\_2020.csv |
| agc\_z\_riodelaplata\_2021.csv |
| agc\_y\_riodelaplata\_2022.csv |
| agc\_y\_riodelaplatal\_2023.csv |
| agc\_z\_riodelaplata\_2024.csv |
| **Descripción:** | Archivos que contienen muestras de calidad de agua del Río de la Plata. | | |
| **Columnas** | **Variable** | **Tipo de valor** | **Descripción** |
|  |
| Año | Numérica discreta | int64 | Año de la campaña de monitoreo. |  |
| Tem\_agua | Numérica continua | float64 | Temperatura del agua (°C). |  |
| Tem\_aire | Numérica continua | float64 | Temperatura del aire (°C). |  |
| OD | Numérica continua | float64 | Oxígeno disuelto (mg/L). |  |
| pH | Numérica continua | float64 | Nivel de acidez o alcalinidad del agua. |  |
| Olores | Binaria | booleano | Presencia o ausencia de olores perceptibles. |  |
| Color | Binaria | booleano | Presencia o ausencia de color anormal. |  |
| Espumas | Categórica | object | Presencia o ausencia de espumas visibles. |  |
| Mat\_susp | Binaria | booleano | Presencia o ausencia de materiales en suspensión. |  |
| colif\_fecales\_ufc\_100ml | Numérica discreta | int64 | Coliformes totales por 100 ml. |  |
| escher\_coli\_ufc\_100ml | Numérica discreta | int64 | Escherichia coli por 100 ml. |  |
| enteroc\_ufc\_100ml | Numérica discreta | int64 | Enterococos por 100 ml. |  |
| Nitrato\_mg\_l | Numérica continua | float64 | Nitratos (mg/L). |  |
| NH4\_mg\_l | Numérica continua | float64 | Amonio (mg/L). |  |
| P\_total\_l\_mg\_l | Numérica continua | float64 | Fósforo total (mg/L). |  |
| Fosf\_ortofos\_mg\_l | Numérica continua | float64 | Fosfato ortofosfato (mg/L). |  |
| DBO\_mg\_l | Numérica continua | float64 | Demanda Biológica de Oxígeno (mg/L). |  |
| DQO\_mg\_l | Numérica discreta | int64 | Demanda Química de Oxígeno (mg/L). |  |
| Turbiedad\_NTU | Numérica continua | float64 | Turbidez del agua (NTU). |  |
| Hidr\_Deriv\_del\_Petroleo\_ug\_l | Numérica discreta | int64 | Hidrocarburos derivados del petróleo (µg/L). |  |
| Cr\_total\_mg\_l | Numérica discreta | int64 | Cromo total (mg/L). |  |
| Cd\_total\_mg\_l | Numérica discreta | int64 | Cadmio total (mg/L). |  |
| Clorofila\_a\_ug\_l | Numérica discreta | int64 | Clorofila-a (µg/L). |  |
| Microcistina\_ug\_l | Numérica discreta | int64 | Microcistina (µg/L). |  |
| ICA | Numérica discreta | int64 | Índice de Calidad del Agua. |  |
| calidad\_de\_agua | Numérica discreta | int64 | Categoría numérica |  |
| campaña\_invierno | Numérica discreta | int64 | Estación del año |  |
| campaña\_otono | Numérica discreta | int64 | Estación del año |  |
| campaña\_primavera | Numérica discreta | int64 | Estación del año |  |
| campaña\_verano | Numérica discreta | int64 | Estación del año |  |

**Fuente de datos**

Existen 12 data sets que fueron descargados desde la plataforma Kaggle y CiAM, específicamente de los siguientes enlaces.

<https://www.kaggle.com/datasets/palomachiacchiara/muestreos-de-calidad-de-agua-de-la-riiglo/data>

[https://ciam.ambiente.gob.ar/repositorio.php?tid=1&stid=105&did=408#](https://ciam.ambiente.gob.ar/repositorio.php?tid=1&stid=105&did=408)

**Github - Cookiecutter**

Para la organización del proyecto, se seguirá la estructura generada con Cookiecutter en GIT:

Data sets.

* Los archivos originales del dataset se almacenarán en la carpeta: Cookiecutter\_Proyecto\_Final\data\raw
* Los archivos procesados en: Cookiecutter\_Proyecto\_Final\data\interim
* El dataset final unificado y será el insumo principal para el modelo de análisis: Cookiecutter\_Proyecto\_Final\data\processed

Documentos.

* Reportes en formato PDF, Cookiecutter\_Proyecto\_Final\reports
* Documentos respaldo, Cookiecutter\_Proyecto\_Final\docs

Link de Git.

<https://github.com/casescas/Ciencia_datos_2A1C/tree/main>