Aprendizaje Automático – Entrega 2: Descripción del Dataset y Origen
Alumno: Diego Estrada

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

Este conjunto de datos tiene valores obtenido por el laboratorio sobre diversos parámetros de calidad del agua, a partir de muestras recolectadas a lo largo de la costera 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).

El dataset que se utilizará en este proyecto abarca el período comprendido entre los años 2017 y 2022, el dataset completo estará compuesto por seis archivos en formato csv, cada uno correspondiente a diferentes campañas de monitoreo. Estos archivos fueron obtenidos de la plataforma Kaggle, donde se encuentran disponibles públicamente. Cabe destacar que hasta el momento este conjunto de datos no ha sido analizado por la comunidad de dicha plataforma, lo que representa una oportunidad valiosa para realizar un estudio exploratorio y aplicar técnicas de aprendizaje automático sobre información ambiental aún no explorada.

Para la organización del proyecto, los archivos originales del dataset se almacenarán en la carpeta data/raw dentro de la estructura del proyecto generada con Cookiecutter en GIT. Por otro lado, se mantendrá una copia de los archivos en la carpeta data/interim, donde se llevará a cabo el procesamiento de datos, ETL (Extracción, Transformación y Carga).

Diccionario del dataset

Los seis datasets fueron unificados en un único DataFrame que contiene 39 columnas y 902 registros. Dado que el diccionario de datos no estaba disponible en la página de Kaggle, fue necesario realizar un análisis detallado de cada columna para interpretar correctamente su significado y tipo de variable.

Es importante señalar que, a medida que se avance en el desarrollo del proyecto, el dataset podrá ser modificado con el objetivo de mejorar el rendimiento del modelo de aprendizaje automático que se aplicará.

Aprendizaje Automático – Entrega 2: Descripción del Dataset y Origen
Alumno: Diego Estrada

DICCIONARIO DE DATOS			
Nombre de Archivos:	agc_y_riodelaplata_2017.cvs agc_y_riodelaplata_2022.csv agc_z_riodelaplata_2018.csv agc_z_riodelaplata_2019.csv agc_z_riodelaplata_2020.csv agc_z_riodelaplata_2021.csv	**Fecha de descarga**	17-may-25
Descripción:	Archivos que contienen muestras de calidad de agua del Río de la Plata.		

Columnas	Variable	Tipo de valor	Descripción
Orden	Numérica discreta	int64	Número secuencial del registro.
Sitios	Categórica	object	Nombre del lugar donde se tomó la muestra.
codigo	Categórica	object	Código identificador del sitio de muestreo.
Fecha	Categórica (fecha como texto)	object	Fecha en que se realizó la medición.
Año	Numérica discreta	int64	Año de la campaña de monitoreo.
Campaña	Categórica	object	Estación del año o periodo de la campaña (ej. Verano).
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), indicador de la capacidad del agua para sustentar vida acuática.
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_totales_ufc_100 ml	Numérica discreta	int64	Unidades formadoras de colonias de coliformes totales por 100 ml.
escher_coli_ufc_100 ml	Numérica discreta	int64	Unidades formadoras de colonias de Escherichia coli por 100 ml.
enteroc_ufc_100ml	Numérica discreta	int64	Unidades formadoras de colonias de enterococos por 100 ml.
Nitrato_mg_l	Numérica continua	float64	Concentración de nitratos (mg/L), indicador de contaminación por fertilizantes o aguas residuales.
NH4_mg_l	Numérica continua	float64	Amonio (mg/L), otro indicador de contaminación orgánica.
P_total_l_mg_l	Numérica continua	float64	Fósforo total (mg/L), relacionado con la eutrofización.

Fosf_ortofos_mg_l	Numérica continua	float64	Fosfato ortofosfato (mg/L), forma biodisponible del fósforo.
DBO_mg_l	Numérica continua	float64	Demanda Biológica de Oxígeno (mg/L), mide la materia orgánica biodegradable.
DQO_mg_l	Numérica discreta	int64	Demanda Química de Oxígeno (mg/L), mide la materia orgánica total.
Turbiedad_NTU	Numérica continua	float64	Turbidez del agua (NTU), indica la cantidad de partículas suspendidas.
Hidr_Deriv_Petr_ug_l	Categórica (valores como "<0.10")	object	Hidrocarburos derivados del petróleo (µg/L).
Cr_total_mg_l	Categórica (valores como "<0.005")	object	Cromo total (mg/L), metal pesado.
Cd_total_mg_l	Categórica (valores como "<0.001")	object	Cadmio total (mg/L), metal pesado tóxico.
Clorofila_a_ug_l	Categórica (valores como "<10")	object	Clorofila-a (µg/L), indicador de biomasa de fitoplancton.
Microcistina_ug_l	Categórica (valores como "<0.20")	object	Microcistina (µg/L), toxina producida por cianobacterias.
ICA	Numérica discreta	int64	Índice de Calidad del Agua (valor numérico).
Calidad_de_agua	Categórica	object	Clasificación cualitativa de la calidad del agua (ej. Muy deteriorada, Extremadamente deteriorada).

Origen del Dataset

El dataset fue descargado desde la plataforma Kaggle, específicamente del siguiente enlace: https://www.kaggle.com/datasets/palomachiacchiara/muestreos-de-calidad-de-agua-de-la-riiglo/data

Muestreos de calidad de agua de la Rio de La Plata – Kaggle

- Fuente: Kaggle – Publicado por el usuario Paloma Chiacchiara
- Fecha de adquisición: [2017 - 2022]
- Licencia: Datos de dominio público (según lo indicado en la plataforma)

Link de Github - Cookiecutter

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