INFORME TÉCNICO – Exploración de datos

Archivo: Marine_Microplastics.csv

Fuente: https://www.kaggle.com/datasets/william2020/marine-microplastics

About:

The Marine Microplastics database contains information on microplastics concentrations, date of collection, latitude and longitude where data was collected, sampling instruments used for collection, and links to publications on the data. The information in this database can be used to improve water quality and protect the ecosystem, especially coastal ecological habitats such as salt marshes and mangrove forests that help recycle nutrients, serve as breeding grounds for fingerlings, and permanent homes for oysters and other coastal marine wildlife.

The dataset includes the following key **columns**:

OBJECTID: Unique identifier for each record.
Oceans: Name of the ocean where the sample was collected.
Regions: Specific region within the ocean.
SubRegions: Subregion within the region.
Sampling Method: Method used to collect the sample.
Measurement: Microplastic density measurement
Unit: Unit of the measurement.
Density Range: Range of the density measurement.
Density Class: Classification of the density (e.g., Very Low, Low, Medium, High, Very High).
Short Reference: Reference for the study.
Organization: Organization responsible for the data collection.
Keywords: Keywords related to the study.
Accession Number: Accession number for the data record.
Accession Link: Link to the data record.
Latitude: Latitude of the sample location.
Longitude: Longitude of the sample location.
Date: Date when the sample was collected.
GlobalID: Global identifier for the record.
x: X-coordinate of the sample location.
y: Y-coordinate of the sample location.

EDA:

1.-Comprobación de nulos y duplicados: Se realiza con una función que engloba ambas herramientas, y devuelve como resultado un pequeño informe detallado del % de nulos por columna en el conjunto de datos, y si existen o no duplicados, en el caso de existir, devuelve el número de duplicados. En este caso:

========== Informe de Datos ==========

Porcentaje de Nulos por Columna:

OBJECTID	0.000000
Oceans	1.326805
Regions	56.499388
SubRegions	93.600979
Sampling Method	0.000000
Measurement	28.455324
Unit	0.000000
Density Range	0.000000
Density Class	0.000000
Short Reference	0.000000
Long Reference	0.000000
DOI	0.000000
Organization	0.000000
Keywords	0.088127
Accession Number	0.000000
Accession Link	0.000000
Latitude	0.000000
Longitude	0.000000
Date	0.000000
GlobalID	0.000000
x	0.000000
y	0.000000

Duplicados:

No hay duplicados

==========

1.1.-Gestión de duplicados: No procede

1.2.-Gestión de nulos: Oceans, Regions, SubRegions, Measurement, y Keywords

- Oceans(1,32): Para abordar los valores nulos en la columna "Oceans", se ha seguido un proceso de imputación utilizando el algoritmo KNN (K-Nearest Neighbors). A continuación, se describen los pasos realizados:

- **Conversión de la columna "Oceans" a valores numéricos**:
  La columna original "Oceans" contenía los siguientes valores únicos: 'Atlantic Ocean', 'Pacific Ocean', 'Arctic Ocean', nan, y 'Indian Ocean'. Se creó una nueva columna llamada "Ocean_cat" mediante la herramienta .map(), utilizando un diccionario que asigna un número a cada océano de forma única:
  - 'Atlantic Ocean' → 1
  - 'Pacific Ocean' → 2
  - 'Arctic Ocean' → 3
  - 'Indian Ocean' → 4
  - nan (valor nulo) → NaN
  - De esta forma, los nombres de los océanos fueron reemplazados por números, y los valores nulos se mantuvieron como NaN.
- **Imputación utilizando KNN**:
  Se utilizó el algoritmo KNN para imputar los valores nulos de la columna "Ocean_cat". La imputación se realizó basándose en dos características: la latitud y la nueva columna "Ocean_cat", que ahora contiene valores numéricos. KNN utiliza los valores más cercanos en función de estas características para predecir los valores faltantes.
- **Conversión de valores numéricos de nuevo a nombres de océanos**:
  Una vez realizada la imputación y obtenida una columna "Ocean_cat" sin valores nulos, se creó un diccionario "inverso" que asigna nuevamente los números a sus respectivos nombres de océano:
  - 1 → 'Atlantic Ocean'
  - 2 → 'Pacific Ocean'
  - 3 → 'Arctic Ocean'
  - 4 → 'Indian Ocean'
  - Usando este diccionario "inverso", se volvió a mapear la columna "Ocean_cat" a los nombres de océano correspondientes.
- **Obtención de la columna sin nulos**:
  Al final del proceso, la columna "Oceans" fue restaurada, y los valores nulos fueron reemplazados por los océanos correspondientes, obteniendo una columna sin valores nulos.

- SubRegions tiene un elevado porcentaje de nulos (93%). Debido a la falta de información relevante que puede aportar, se ha decidido excluir esta columna del análisis, ya que no contribuirá significativamente al estudio.

- Regions (56%): Se intentó imputar utilizando el valor único más frecuente respecto a “Oceans”, pero no tienen relevancia. Se utilizo la librería geopy a través de la latitud y la longitud encontrar a que Región pertenecía, asignando rangos de margen de tolerancia y error, pero no se logró imputar de forma coherente para este estudio. Se decidió no gestionar nulos, pese al alto % de los mismos.
- Measurement(28%): Utilizando
  ```
  df_mp.groupby("Sampling Method")["Measurement"].apply(lambda x: x.isnull().sum())
  ```
  Se comprobó, que todos los nulos de “Measurement” coincidían con la categoría “Hand Picking” de “Sampling Method”. Para tener constancia de ello en el estudio.
- Keywords: No relevante

2.-Transformaciones para optimizar el uso de los datos

- La columna *Density Range* contiene intervalos de valores. Para facilitar los análisis posteriores, se ha creado una nueva columna con el valor central (promedio) de cada rango o si no lo contiene, el valor junto a >= o >, sin eliminar la columna original. Esta nueva variable permite trabajar con un valor numérico representativo de la densidad
- Se convierte Date a formato “datetime”, y se generar 3 nuevas columnas a partir de ‘Date’:
  - ‘Year’
  - ‘Month’
  - ‘Hour’

  La columna “Hour” tiene un solo valor único, 12 AM para todas las observaciones, se puede eliminar.

  Se conserva la columna ‘Date’ Original sin la hora, ‘Year’ y ‘Month’
- Se ajusta el nombre de los valores únicos en “Regions” (eliminando espacios y normalizando el texto a minúsculas, ya que “Rio de la Plata” aparecía duplicado. Se soluciona
- Eliminación de columnas irrelevantes: ‘x’, ‘y’, ‘Hour` ‘Date_Only’(generada en la tranformación Date, que contiene la misma información que Date, pero en formato object) y ‘SubRegions’

3.- Resultados:

El nuevo archivo “Clean_Microplastic.csv” tiene una forma de : 20425 filas, 22 columnas

Informe final:

========== Informe de Datos ==========

Porcentaje de Nulos por Columna:

OBJECTID	0.000000
Oceans	0.000000
Regions	56.499388
Sampling Method	0.000000
Measurement	28.455324
Unit	0.000000
Density Range	0.000000
Density Class	0.000000
Short Reference	0.000000
Long Reference	0.000000
DOI	0.000000
Organization	0.000000
Keywords	0.088127
Accession Number	0.000000
Accession Link	0.000000
Latitude	0.000000
Longitude	0.000000
Date	0.000000
GlobalID	0.000000
Date_Only	0.000000
Density_center	0.000000
Year	0.000000
Month	0.000000

Duplicados:

No hay duplicados

==========

The CLEAN_dataset includes the following key columns:

OBJECTID: Unique identifier for each record.
Oceans: Name of the ocean where the sample was collected.
Regions: Specific region within the ocean.
Sampling Method: Method used to collect the sample.

Measurement: Microplastic density measurement (pieces/m3).
Unit: Unit of the measurement.
Density Range: Range of the density measurement.
Density Class: Classification of the density (e.g., Very Low, Low, Medium, High, Very High).
Short Reference: Reference for the study.
Organization: Organization responsible for the data collection.
Keywords: Keywords related to the study.
Accession Number: Accession number for the data record.
Accession Link: Link to the data record.
Latitude: Latitude of the sample location.
Longitude: Longitude of the sample location.
Date: In the correct datetime format (YYYY-MM-DD).
GlobalID: Global identifier for the record.
Density_center: It is generated to represent the central value of a density range.
Year: int.
Month: int.