# Procesamiento de datos y manejo de datos faltantes

**Objetivo y alcance del trabajo**

El objetivo de este componente práctico, es explorar métodos para preparación y preprocesamiento de datos con Python, aplicando todos los contenidos revisados durante toda la asignatura, es decir, obtención del conjunto de datos, análisis del conjunto de datos, tratamiento de datos, y evaluación de datos faltantes Para esto, se recomienda utilizar Google Colab que permite escribir y ejecutar código Python en la nube, disponible mediante el navegador web, e integrarlo con bloques de texto en formato Markdown para documentación complementaria. Se debe presentar un solo documento tipo notebook en formato "ipynb" y adjuntar los archivos que se soliciten en el desarrollo. El nombre del notebook a entregar debe tener el siguiente formato: "AMGD_CP_W3_G#", en donde '#' es el número de grupo.

## Fase I

Dado el conjunto de datos oil-spill.csv:

(https://www.kaggle.com/datasets/ashrafkhan94/oil-spill/data)

- Importa el dataset y has un análisis exploratorio de los datos muestra el resultado de este análisis en el notebook.
- Identifica cuántas y cuáles columnas tienen valores únicos. Elimínalas e imprime el tamaño del dataset antes y después.
- Analiza el porcentaje de valores únicos por columna. Define un límite de incidencia (por ejemplo, 1%), identifica columnas por debajo de ese umbral y elimínalas, mostrando el porcentaje respectivo.
- Analiza y grafica cuántas columnas serían eliminadas usando diferentes umbrales de incidencia.
- Visualiza la distribución de los valores en las columnas filtradas mediante gráficos (barplot, histograma o similar).
- Realiza un análisis de varianza por columna y comenta los resultados.

## Fase II

Dado el conjunto de datos Pima-indians-diabetes.csv

(https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database)

- Importa el dataset y has un análisis exploratorio de los datos muestra el resultado de este análisis en el notebook.
- Determina, consultando la documentación, qué columnas consideran los ceros como datos faltantes.
- Reemplaza estos ceros por NaN y genera dos datasets: uno eliminando los registros faltantes y otro manteniendo los registros reemplazados.
- Imprime el tamaño antes y después de la limpieza.

- En un nuevo DataFrame imputa los valores faltantes con una estrategia de tu elección y realiza una comparación de los resultados respecto al dataset sin imputación.
- Obtén la matriz de sombras del set de datos que tenías reemplazado los 0 por NaN
- Visualiza el patrón de datos faltantes.
- Realiza el dendograma de estos datos faltantes

## Fase III

Aplicando todo lo revisado en esta asignatura, accede al conjunto de datos Auto MPG:

https://archive.ics.uci.edu/dataset/9/auto+mpg

- Importa el dataset y has un análisis exploratorio de los datos muestra el resultado de este análisis en el notebook.
- Investiga sobre el contenido del conjunto de datos y definición de las columnas útiles
- Realiza un análisis descriptivo del set de datos
- Detecta, visualiza y analiza los valores faltantes con al menos dos técnicas de visualización diferentes
- Evalúa la correlación de nulidad
- Obtén la matriz de sombras
- Realiza la imputación de los datos
- Escala los datos utilizando al menos una técnica (MinMaxScaler, StandardScale) y discute las diferencias en los resultados.

## Fase IV

**Reflexión y buenas prácticas**

Redacta una sección final de reflexiones:

- ¿Qué desafíos encontraste al tratar los datos faltantes y al escalar los datos?
- ¿Cuál método de imputación te pareció más efectivo y por qué?
- ¿Qué importancia tiene la visualización en el tratamiento de datos faltantes?
- Indica al menos dos buenas prácticas para futuros proyectos de ciencia de datos respecto al preprocesamiento y manejo de datos faltantes.