# Informe De Análisis De Datos

## Metodología

Para el desarrollo de esta actividad se aplicaron técnicas estadísticas de análisis de correlación y regresión no lineal, con el objetivo de explorar las relaciones existentes entre las variables cuantitativas del conjunto de datos.

1. **Selección y preparación de datos:** Se trabajó con el dataframe: *df_numeric*, compuesto por variables como: TaxonName, TaxonCode, Abundance_nbcell, Abundance_pm, TotalAbundance_SamplingOperation, SamplingOperations_code, CodeSite_SamplingOperations

   Se eliminaron valores nulos y se normalizaron los datos numéricos para asegurar la comparabilidad entre variables.

2. **Cálculo de la matriz de correlación**
   - Se generó una matriz de correlación utilizando el método de **Pearson**.
   - Posteriormente, se obtuvieron los valores absolutos de la correlación para identificar las relaciones más fuertes sin considerar la dirección (positiva o negativa).
   - Se extrajeron los cinco pares de variables con mayor correlación.

3. **Selección de modelos de regresión:** Con base en los resultados de correlación, se implementaron tres modelos de regresión no lineal:
   - **Modelo 1: Polinómico racional**

     $$f(x) = \frac{ax^2 + b}{cx^2}$$

     Variables: Abundance_pm (independiente) y TaxonName (dependiente).

- **Modelo 2: Logarítmico**

$$f(x) = a\log(x) + b$$

Variables: Abundance_pm (independiente) y TaxonCode (dependiente).

**Modelo 3: Cuadrático simple**

$$f(x) = ax^2 + bx + c$$

Variables: CodeSite_SamplingOperations (independiente) y SamplingOperations_code (dependiente).

4. **Evaluación del desempeño**
    - Se aplicó la función curve_fit() del paquete **scipy.optimize** para ajustar los parámetros.
    - Se calculó el **coeficiente de determinación R2R^2R2** y su valor absoluto para evaluar el grado de ajuste del modelo.

## Hallazgos

- Los **pares con correlación perfecta (1.00)** correspondieron a variables equivalentes o directamente derivadas unas de otras:
    - TaxonName y TaxonCode
    - Abundance_nbcell y Abundance_pm
    - TotalAbundance_SamplingOperation con Abundance_nbcell y Abundance_pm
- La correlación **moderada (0.24)** entre SamplingOperations_code y CodeSite_SamplingOperations sugiere cierta relación jerárquica o de codificación compartida.
- Los modelos de regresión arrojaron los siguientes resultados:
    - **Modelo 1:** Ajuste no significativo, alta dispersión de los datos.

    - **Modelo 2:** R2=0.0906R^2 = 0.0906R2=0.0906, relación débil entre Abundance_pm y TaxonCode.
    - **Modelo 3:** Presentó bajo nivel de ajuste y comportamiento errático en los valores predichos.
- El **coeficiente de correlación absoluto R=0.301R = 0.301R=0.301** confirma que la relación entre las variables modeladas es débil.

**Conclusiones**

- La correlación moderada entre SamplingOperations_code y CodeSite_SamplingOperations representa la única relación parcialmente significativa, pero aún insuficiente para construir un modelo robusto.
- Los resultados de las regresiones no lineales indican que las variables seleccionadas no presentan una relación funcional clara; los modelos simples no capturan adecuadamente la estructura subyacente de los datos.
- Se recomienda:
    - Eliminar variables redundantes antes del modelado.
    - Aplicar **modelos multivariados** o métodos de aprendizaje no lineal (como regresión polinómica de mayor orden o random forest).
      Incorporar nuevas variables contextuales que expliquen mejor las variaciones observadas.
- En general, el análisis permitió identificar los límites de las correlaciones lineales y la necesidad de enfoques más avanzados para comprender la dinámica de las variables biológicas estudiadas.