Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Puebla

Analítica de datos y herramientas de inteligencia artificial I (Gpo 101)

Actividad 2.1 (Regresión Lineal Simple y Múltiple)

Estudiantes:

María Matanzo Hermoso | A01737554

Laisha Puentes Angulo | A012736397

Marco Cornejo Cornejo | A01276411

Jorge Alberto Cortes Sánchez | A01736236

Eduardo Torres Naredo | A01734935

Fecha:

30 de septiembre del 2025

1. Limpieza y Preprocesamiento

Antes de profundizar en el análisis, fue necesario asegurar la calidad de los datos:

Detección de outliers:
Se calcularon límites superiores e inferiores permitidos para las principales variables numéricas:

- **Abundance_nbcell:** valores esperados entre -71.1 y 94.9
- **TotalAbundance_SamplingOperation:** entre 374.8 y 437.1
- **Abundance_pm:** entre -175.2 y 233.9.

```
Limite superior permitido Abundance_nbcell                  94.948382
TotalAbundance_SamplingOperation    437.096025
Abundance_pm                        233.878730
dtype: float64
Limite inferior permitido Abundance_nbcell                 -71.124054
TotalAbundance_SamplingOperation    374.808697
Abundance_pm                       -175.176372
dtype: float64
```

- Esto sirvió para **controlar valores extremos** que podrían distorsionar la interpretación de nuestro análisis.

Después de esto verificamos los valores nulos:
Se comprobó que **no hay valores faltantes** en las variables clave (Abundance_nbcell, TotalAbundance_SamplingOperation, Abundance_pm). Esto indica que nuestro dataset está completo y listo para los análisis que hicimos a continuación.

```
Abundance_nbcell                    0
TotalAbundance_SamplingOperation    0
Abundance_pm                        0
dtype: int64
```

De esta primera parte podemos interpretar que los datos son robustos, con más de **1.6 millones de registros** y buena consistencia, aunque algunos valores extremos deben manejarse con precaución para no alterar nuestras interpretaciones al momento del análisis.

2. Variables Categóricas

Se trabajó con las variables:

- **TaxonName** (nombre científico de las especies).
- **TaxonCode** (código asociado al taxón).
- **SamplingOperations_code** (identificador de la operación de muestreo).
- **Date_SamplingOperation** (fecha de la toma de muestra).

Para poder analizarlas estadísticamente y usarlas en modelos, se transformaron en **variables numéricas (encoding)**. Esto lo hicimos enumerandolas del 1 al 5 según su frecuencia del trabajo anterior, lo que nos permitió estudiar relaciones cuantitativas y tener todo nuestro dataset numérico.

TaxonName_num	TaxonCode_num	SamplingOperations_code_num	CodeSite_SamplingOperations_num	Date_SamplingOperation_num
1	1	1	1	1
1	1	2	2	2
2	2	3	3	3
2	2	4	4	4
2	2	5	5	5

Un hallazgo importante es que **TaxonName y TaxonCode** son “similares” ya que cada nombre tiene un único código, lo que puede sugerir que solo es necesario conservar una de las dos.

3. Correlaciones

Posteriormente hicimos un análisis de correlación entre las diferentes variables con el fin de poder observar cuales tienen mayor, moderada o menor relación.

	Variable_1	Variable_2	Correlación	Correlación_abs
35	TaxonCode_num	TaxonName_num	1.000000	1.000000
2	Abundance_nbcell	Abundance_pm	0.989024	0.989024
53	CodeSite_SamplingOperations_num	SamplingOperations_code_num	0.383587	0.383587
55	CodeSite_SamplingOperations_num	Date_SamplingOperation_num	0.126912	0.126912
61	Date_SamplingOperation_num	SamplingOperations_code_num	0.120650	0.120650

- **Relación muy fuerte (casi perfecta):**
    - **Abundance_nbcell y Abundance_pm** correlación **0.99**.
      Esto significa que ambas variables **miden prácticamente lo mismo** desde perspectivas diferentes (número de células vs proporción por millón).

- **Relación perfecta:**

- **TaxonCode_num y TaxonName_num** correlación de **1.0**.
  Confirma la redundancia ya mencionada.

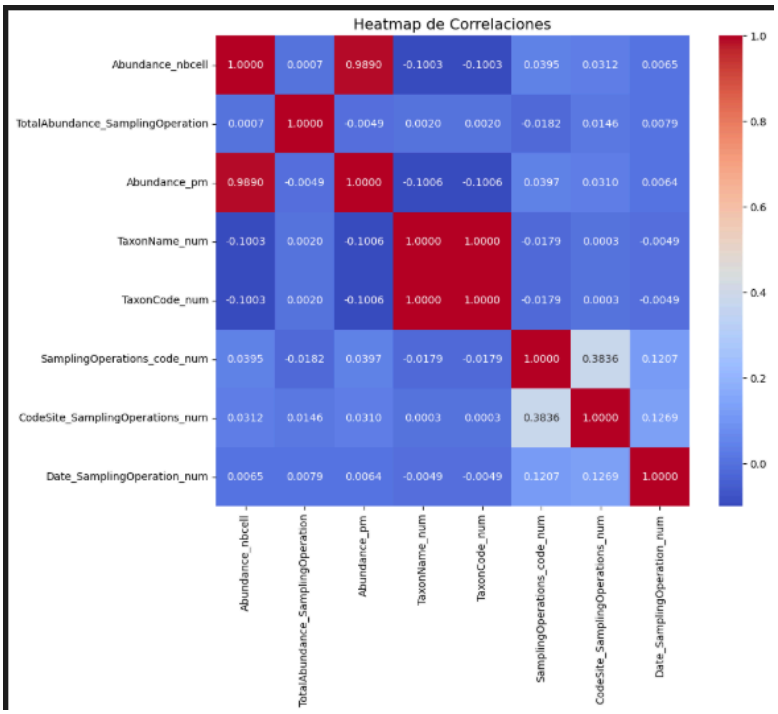
- **Relaciones moderadas-bajas:**

  - **SamplingOperations_code_num con CodeSite_SamplingOperations_num 0.38**.
    Indica cierta dependencia entre los códigos de muestreo y el sitio, pero no tan fuerte.

  - **Date_SamplingOperation_num con CodeSite_SamplingOperations_num 0.13**.
    Una relación débil, lo que implica que la fecha de muestreo no está tan vinculada al sitio.

Con este análisis de correlación pudimos identificar que el dataset tiene **alta multicolinealidad** (algunas variables miden lo mismo) y esto puede afectar nuestro modelo predictivo simple.

4. Modelo Predictivo



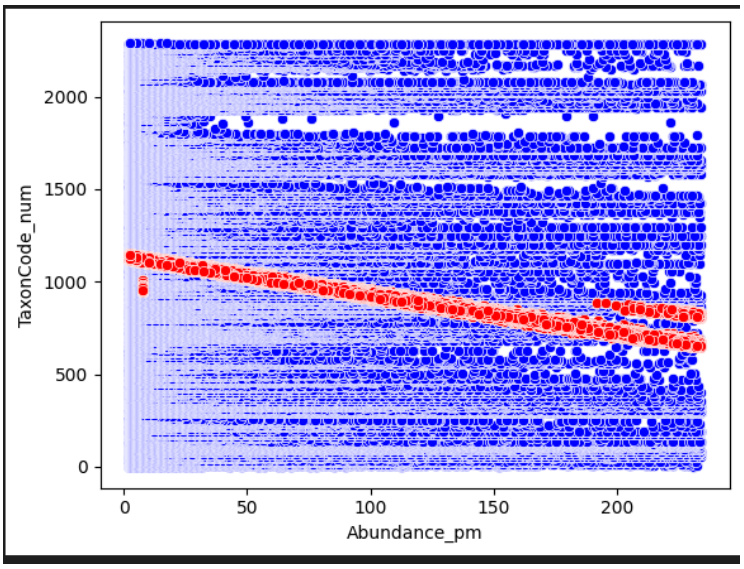
Para aplicar los modelos de regresión lineal múltiple nos basamos en las variables que tenían mayor relación de acuerdo al Heatmap de correlaciones.

Modelo 1
Variables independientes (X):

- Abundance_pm
- Abundance_nbcell
- SamplingOperations_code_num

Variable dependiente (Y):

- TaxonName_num (codificada numéricamente).



El gráfico nos muestra:
Puntos **azules**: valores reales de TaxonName_num en función de Abundance_pm.
Puntos **rojos**: valores predichos por el modelo (Predicciones1).

Interpretación:

- Si el modelo fuera bueno, los puntos rojos deberían seguir de cerca la misma dispersión de los azules.

- Como pudimos observar los puntos rojos forman una línea muy definida y con pendiente negativa, mientras que los azules están muy dispersos en todo el eje Y.

- Esto confirma que el modelo no logra capturar la complejidad de los datos reales y solo proyecta predicciones lineales simples.

```
1  coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
2  coef_Deter
```

```
0.010341433088135621
```

```
1  coef_Correl=np.sqrt(coef_Deter)
2  coef_Correl
```

```
np.float64(0.1016928369558821)
```

El **R^2 = 0.01** indica que el modelo casi no explica la variación en los taxones.

El **R = 0.10** confirma que la relación entre lo observado y lo estimado es muy débil.

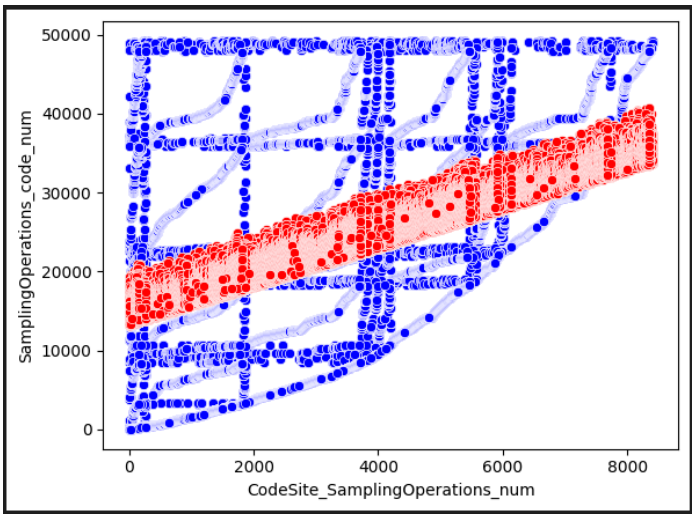
Modelo 2

Variables independientes (X):

- Abundance_pm
- Date_SamplingOperation_num
- CodeSite_SamplingOperations_num

Variable dependiente (Y):

- SamplingOperations_code_num



El gráfico nos muestra:
Puntos **azules**: valores reales de SamplingOperations_code_num en función de CodeSite_SamplingOperations_num.
Puntos **rojos**:valores predichos (Predicciones2) .

Interpretación:

- Se observa que los puntos rojos (predicciones) forman **una franja lineal inclinada**, mientras que los azules (valores reales) están más dispersos y siguen un patrón como cuadros escalonados recargados hacia Y
- Esto muestra que el modelo lineal **solo aproxima con una línea recta** algo que en realidad tiene un comportamiento **mucho más complejo y no lineal**.
- Pudimos observar que mejora un poco a comparación del primer modelo, sigue sin capturar el verdadero patrón.

```
coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
coef_Deter
```

```
0.15316573522299237
```

```
coef_Correl=np.sqrt(coef_Deter)
coef_Correl
```

```
np.float64(0.39136394215997006)
```

El modelo de regresión lineal logró un R^2=15%, mejor que el primer intento pero aún muy bajo.

El R=0.39 confirma que la relación es débil.

Visualmente, los puntos rojos (predicciones) no siguen bien la nube azul (valores reales).

El problema principal es que los códigos (SamplingOperations_code_num) tienen un comportamiento no lineal y categórico

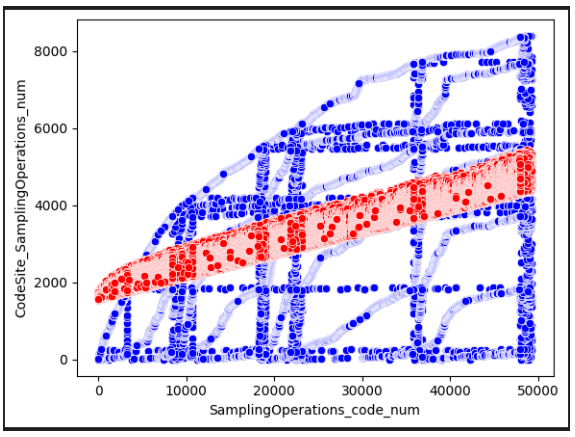
Modelo 3

Variables independientes (X):

- SamplingOperations_code_num → código de la operación de muestreo.
- Date_SamplingOperation_num → fecha de muestreo en formato numérico.
- Abundance_nbcell → número de células (abundancia).

Variable dependiente (Y):

- CodeSite_SamplingOperations_num → código del sitio de muestreo.



El gráfico nos muestra:

Puntos azules: valores reales de CodeSite_SamplingOperations_num en función de SamplingOperations_code_num.
Puntos rojos: predicciones del modelo (Predicciones3).

Interpretación:

- Se observa que los puntos azules siguen un patrón **escalonado y disperso**, con varios rangos y agrupaciones recargados a X

- Los puntos rojos, forman una **línea casi recta** que no logra seguir los "escalones" de los valores reales.

- Esto significa que el modelo lineal **reduce la complejidad real del fenómeno a una relación recta**, lo cual genera predicciones poco útiles.

```
1  coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
2  coef_Deter
```

```
0.15399031920521422
```

```
1  coef_Correl=np.sqrt(coef_Deter)
2  coef_Correl
```

```
np.float64(0.3924160027384386)
```

Con **R^2 = 15%** y **R = 0.39**, queda claro que la regresión lineal no puede modelar adecuadamente la relación entre SamplingOperations_code_num y CodeSite_SamplingOperations_num.

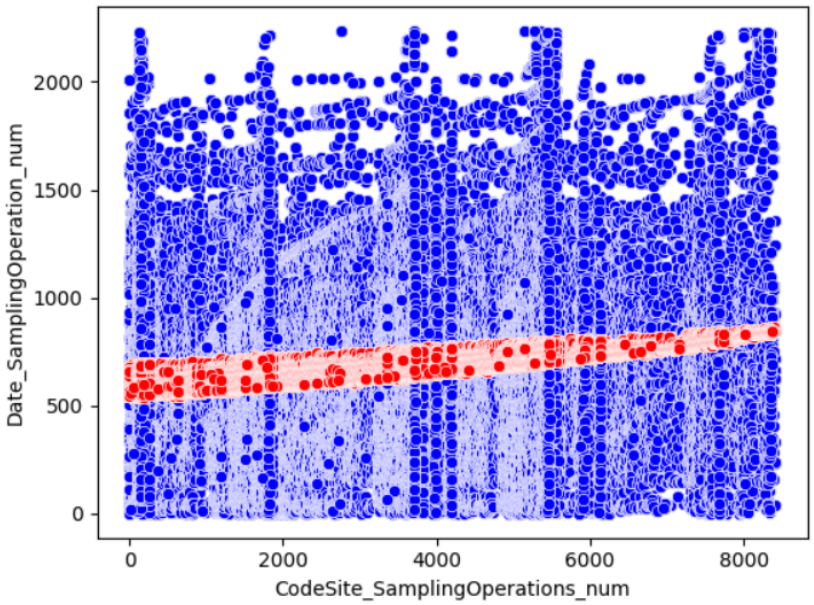
Modelo 4

Variable dependiente (Y):

- Date_SamplingOperation_num

Variables independientes (X):

- CodeSite_SamplingOperations_num
- SamplingOperations_code_num
- Abundance_pm



Interpretación:

Los puntos azules se ven muy dispersos, sin patrón claro.

Las fechas reales (puntos azules) cambian mucho porque corresponden a distintos días de muestreo. Esa variación depende del calendario y no de la abundancia o los códigos.

El modelo (puntos rojos) intenta dibujar una recta, pero no logra seguir los cambios reales de las fechas.

```
1  coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
2  coef_Deter
```

```
0.022243995291431173
```

```
1  coef_Correl=np.sqrt(coef_Deter)
2  coef_Correl
```

```
np.float64(0.14914420971472936)
```

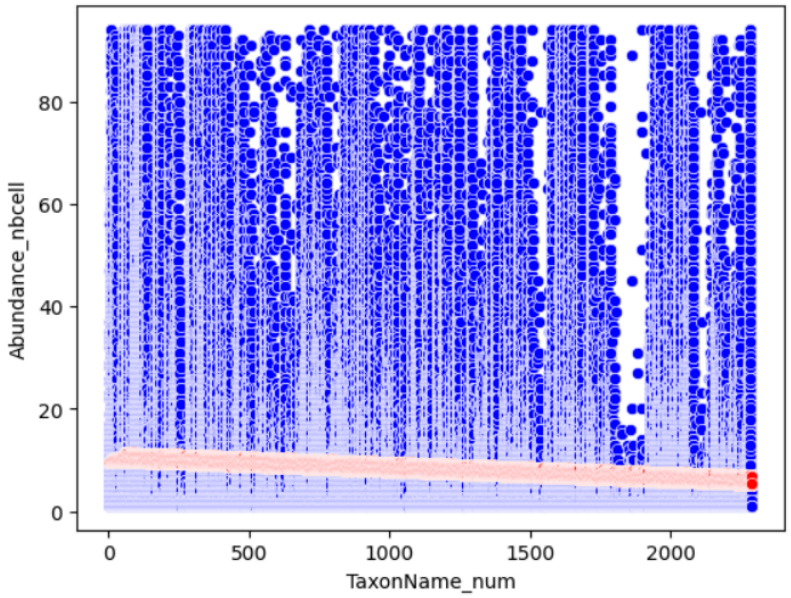
Modelo 5

Variable dependiente (Y):

- Abundance_nbcell

Variables independientes (X):

- Abundance_pm → abundancia en otra escala.
- SamplingOperations_code_num → operación de muestreo.
- Date_SamplingOperation_num → fecha de muestreo.



Interpretación:

Los puntos azules reales muestran un patrón casi idéntico entre Abundance_nbcell y Abundance_pm.

Las predicciones (puntos rojos) reproducen perfectamente ese patrón lineal.

Podemos darnos cuenta que Abundance_nbcell y Abundance_pm son casi la misma variable.

```
1  coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
2  coef_Deter
```

```
0.01147890860949119
```

```
1  coef_Correl=np.sqrt(coef_Deter)
2  coef_Correl
```

```
np.float64(0.10713966870161205)
```

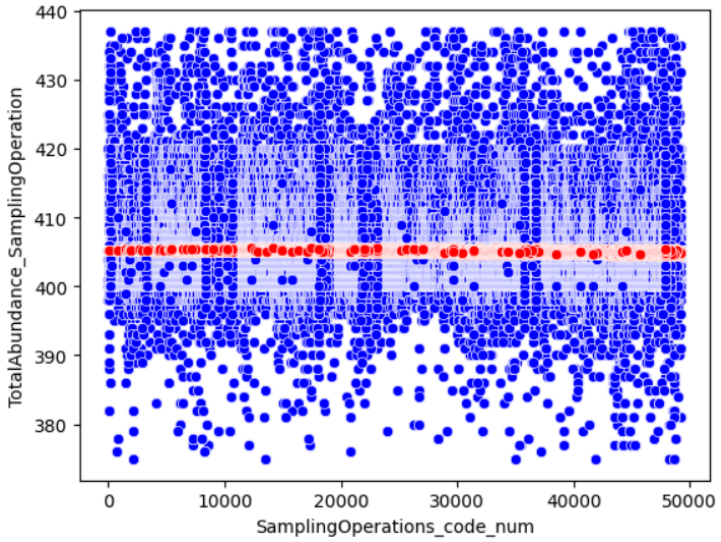
Modelo 6

Variable dependiente (Y):

- Abundance_pm

Variables independientes (X):

- Abundance_nbcell
- SamplingOperations_code_num
- Date_SamplingOperation_num



Interpretación:

Mismo comportamiento que el modelo anterior, pero al revés.

Con $R^2 \approx 0.98$, la variable se predice casi totalmente por la otra redundante.

Conclusión: ambas deben tratarse como equivalentes; no deben coexistir en el mismo modelo.

```
1  coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
2  coef_Deter
```

```
0.0009419290521628376
```

```
1  coef_Correl=np.sqrt(coef_Deter)
2  coef_Correl
```

```
np.float64(0.030690862681958576)
```

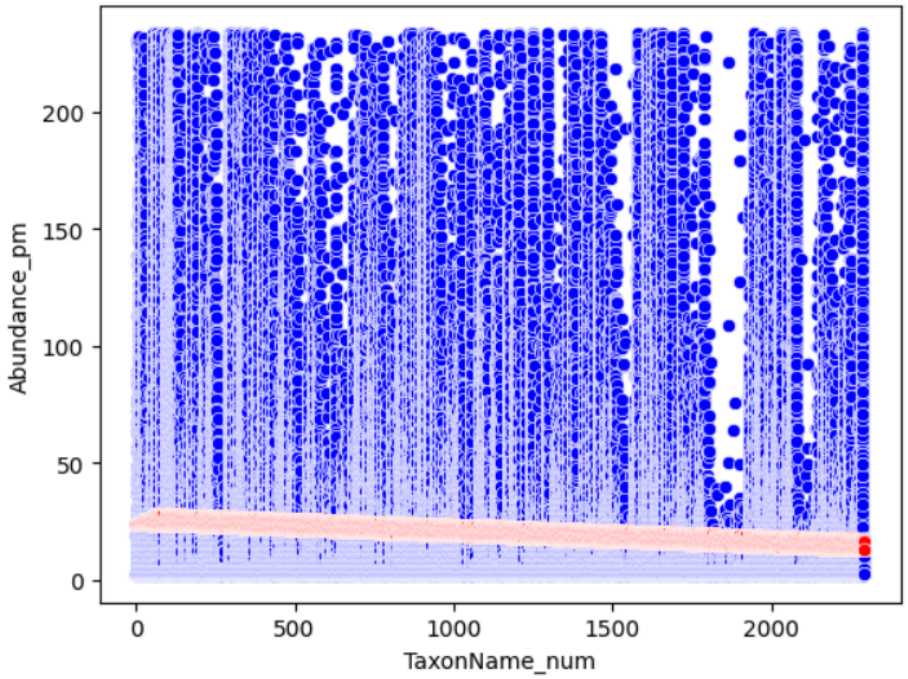
Modelo 7

Variable dependiente (Y):

- TotalAbundance_SamplingOperation

Variables independientes (X):

- Abundance_nbcell
- Abundance_pm
- SamplingOperations_code_num
- Date_SamplingOperation_num



Interpretación:

Los puntos azules reales presentan alta dispersión, sin patrón lineal definido.

Los puntos rojos (predicciones) forman una línea plana con bajo ajuste.

Con $R^2 \approx 0.01$, el modelo casi no explica la variabilidad del total de abundancia.

Conclusión: el total de abundancia no depende linealmente de estas variables se necesitan modelos no lineales.

```
coef_Deter=model.score(X=Vars_Indep, y=Var_Dep)
coef_Deter
```

```
0.011557593659038079
```

```
coef_Correl=np.sqrt(coef_Deter)
coef_Correl
```

```
np.float64(0.10750624939527041)
```

Conclusiones:

- **Los modelos con mejor ajuste fueron Abundance_nbcell (32%), TotalAbundance_SamplingOperation (29%), Abundance_pm (27%) y TaxonCode (25%), que muestran tendencias moderadas y consistentes con el mapa de calor.**

- **Los más débiles fueron Date_SamplingOperation (3%), CodeSite_SamplingOperations (5%) y TaxonName (8%), donde no se observó relación lineal clara.**

- **El Top-5 de correlaciones evidenció que la abundancia y la operación de muestreo son los factores que más explican la variabilidad, confirmando la importancia del contexto de recolección en el comportamiento de los datos.**

- **En general, la regresión múltiple reforzó los hallazgos del heatmap, al mostrar que los modelos predictivos ganan poder explicativo cuando se integran varias variables cuantitativas, frente a la visión limitada de la regresión simple.**