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 data frame: *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 dos modelos por cada par de variables, dando un total de seis modelos:
   - Par 1: Abundance_pm ↔ TaxonName
     - Modelo racional: f(x) = ax2 + b / cx2
     - Modelo logarítmico: f(x) = a log(x) + b
   - Par 2: Abundance_pm ↔ TaxonCode
     - Modelo racional: f(x) = ax2 + b / cx2
     - Modelo logarítmico: f(x) = a log(x) + b
   - Par 3: SamplingOperations_code ↔ CodeSite_SamplingOperations
     - Modelo cuadrático: f(x) = ax2 + bx + c
     - Modelo logarítmico: f(x) =a log(x) + b
4. **Se aplicó la función curve_fit()** para ajustar los parámetros. Se calculó el coeficiente de determinación R2 y su valor absoluto para evaluar el grado de ajuste del modelo. El modelo logarítmico del primer par presentó un R2=0.0906 y un coeficiente de correlación R=0.301, lo que refleja una relación débil entre las variables.

Hallazgos

- Los pares con correlación perfecta (1.00) corresponden 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 indica que ambos códigos podrían estar organizados bajo una misma estructura o esquema de identificación.
- Los resultados de los seis modelos fueron los siguientes:
    - Par 1 (Abundance_pm–TaxonName):

      Ambos modelos (racional y logarítmico) presentaron bajo poder explicativo y alta dispersión de los datos.
    - Par 2 (Abundance_pm–TaxonCode):

      Se repitió el patrón de ajuste débil, con valores de R2 bajos y sin tendencia clara.
    - Par 3 (SamplingOperations_code–CodeSite_SamplingOperations):

      Ninguno de los modelos (cuadrático ni logarítmico) logró representar adecuadamente la relación; ambos mostraron dispersión y ajuste limitado.
- En general, los resultados sugieren que las variables elegidas no mantienen una relación funcional clara, lo que evidencia la complejidad del fenómeno y la posible necesidad de incluir más factores explicativos.

Conclusiones

En total se estimaron seis modelos de regresión, dos por cada par de variables analizadas, con el propósito de evaluar la existencia de relaciones significativas entre ellas. Los resultados demostraron que la mayoría de las correlaciones eran débiles o producto de redundancia entre variables equivalentes, como en los casos de TaxonName - TaxonCode y Abundance_nbcell - Abundance_pm, donde los valores de correlación perfecta (1.00) reflejan información duplicada más que una dependencia real. La única relación moderada observada fue entre SamplingOperations_code y CodeSite_SamplingOperations, lo que sugiere una posible correspondencia estructural entre los códigos de operación y los sitios de muestreo, aunque su

fuerza predictiva sigue siendo limitada. En todos los modelos (racional, logarítmico y cuadrático) el nivel de ajuste fue bajo, confirmando que las variables no presentan una relación funcional clara y que probablemente intervienen otros factores no considerados. Estos hallazgos evidencian la necesidad de depurar las variables redundantes, incorporar atributos contextuales adicionales y emplear modelos más complejos o multivariados para obtener resultados más precisos y representativos del fenómeno estudiado.