Links interesantes:

<https://www.kaggle.com/competitions/mlunlp-2024-ps-2/data>

<https://microdatos.dane.gov.co/index.php/catalog/608/datafile/F1#page=F2&tab=data-dictionary>

REPOSITORIO Template <https://github.com/ignaciomsarmiento/PS_Repo>

**Introducción**

Este conjunto de datasets está inspirado en una reciente competencia organizada por el Banco Mundial: *Pover-T Tests: Predicting Poverty*. El objetivo principal de la competencia es predecir la pobreza en Colombia, un desafío complejo, ya que medir la pobreza es costoso, lleva tiempo y resulta complicado. Tal como señala la competencia, "construir mejores modelos permitirá realizar encuestas con menos preguntas, pero más dirigidas, lo que permitirá medir rápidamente y de forma económica la efectividad de nuevas políticas e intervenciones. Cuanto más precisos sean nuestros modelos, más eficientemente podremos dirigir intervenciones y ajustar políticas, maximizando el impacto y la rentabilidad de estas estrategias".

En este contexto, el objetivo de este proyecto es predecir la pobreza a nivel de hogar, utilizando datos proporcionados por DANE y la misión del *Empalme de las Series de Empleo, Pobreza y Desigualdad – MESE*. Los datos incluyen cuatro conjuntos divididos en entrenamiento y prueba, tanto a nivel de hogar como de individuo. La variable clave para unir los distintos conjuntos de datos es el identificador *id*, que permite la vinculación entre los hogares y los individuos.

Desde una perspectiva de modelado, considero esencial construir modelos de *machine learning* que utilicen variables económicas relevantes y directamente relacionadas con el problema planteado. No es recomendable incluir todas las variables posibles, ya que esto podría resultar en una sobrecarga de datos que no aporta valor al modelo. De hecho, la inclusión de demasiadas variables podría hacer que el modelo sea más complejo sin mejorar su capacidad predictiva.

A pesar de lo mencionado anteriormente, deseo aclarar que considero que el uso de modelos más simples, con un número reducido de variables, no implica necesariamente una mejora en los resultados relacionados con la pobreza y la desnutrición. Existen factores subyacentes causales, como el nivel educativo, que influyen en la pobreza a largo plazo. El nivel educativo, por ejemplo, abre puertas a mayores oportunidades, acceso a servicios básicos y un mejor bienestar general. Estos factores no se transforman rápidamente; requieren un esfuerzo sostenido a lo largo del tiempo. Si bien las intervenciones a corto plazo, como las transferencias directas, pueden mostrar mejoras inmediatas, en el largo plazo podrían generar efectos secundarios que, en algunos casos, no resuelven las causas profundas de la pobreza. Por ello, es crucial que los modelos construidos no solo se enfoquen en la medición de la pobreza de manera inmediata, sino también en comprender los factores estructurales que requieren soluciones más complejas y sostenidas en el tiempo.