Universidad de Los Andes                                                                                     26 de junio de 2022

Maestría en Economía Aplicada

Big Data

**Problem Set 2 – Big Data**

**Estudiantes:** María Isabel Gómez, Paola Ortiz y Sofía Vanegas

El objetivo de este problem set es construir un modelo predictivo de la pobreza del hogar usando los datos de la Gran Encuesta Integrada de Hogares (GEIH). A continuación, se presenta el vínculo al repositorio de GITHUB:

* <https://github.com/BigData-Gomez-Ortiz-Vanegas>

**Introducción**

La medición de la pobreza ha sido uno de los retos más grandes para los hacedores de política pública. Es tremendamente relevante tener una medición certera de una problemática como la pobreza, pues permite enfocar y maximizar los programas dirigidos a reducirla. Artículos como los de [Blank & Greenberg](https://www.brookings.edu/research/improving-the-measurement-of-poverty/) (2008) y [Haughton & Khandker](https://documents1.worldbank.org/curated/en/488081468157174849/pdf/483380PUB0Pove101OFFICIAL0USE0ONLY1.pdf) (2009) discuten la medición de la pobreza. En los años recientes y gracias a los avances en tecnología, hay un creciente interés por una predicción precisa del nivel de pobreza que permita la correcta focalización de los programas. El concurso del [Banco Mundial](https://www.drivendata.org/competitions/50/worldbank-poverty-prediction/page/97/) (2018) es una muestra de la motivación y relevancia que recibe este tema.

Para este *problem set* se cuenta con los datos de la Encuesta Integrada de Hogares (GEIH) y se considera que son apropiados para predecir la pobreza de los hogares en Colombia. Los datos se pueden obtener a nivel de hogar o de integrantes del hogar (hogares y personas, respectivamente) y son un corte transversal con características relevantes que pueden relacionarse con el nivel de pobreza (e.g., ingreso total). Por eso se considera que son apropiados para el objetivo del ejercicio.

[agregar breve preview of results and main takaways]

**Datos**

Para los propósitos del presente *problem set*, se nos dio una base de entrenamiento y de prueba para hogares y personas. Es decir que contábamos con cuatro tablas de datos independientes.

Proceso de limpieza de datos

Primero, se renombraron las variables relevantes para facilitar su identificación y manejo. Después, se identificó que la variable de *ingreso total* estaba en la base de entrenamiento, pero no en la de prueba. Para sortear esta dificultad, se hizo una estimación lineal del *ingreso total* en la base de entrenamiento y se predijo dicha variable a partir de los datos de la base de prueba (la ecuación de la estimación se puede ver en el Anexo 1). En tercer lugar, se sumó el *ingreso total* a nivel de hogar en las bases de personas y se incorporaron esos datos a las bases de hogares. Por último, se consideró que tener presente el *sexo de jefe/a de hogar* podría ser relevante; así que se identificó ese dato para cada hogar y se incorporó a las bases de hogares. El resultado son dos bases de datos: hogares\_test y hogares\_train.

Estadísticas descriptivas

**Modelos y resultados**

**Conclusiones**