Data Science for Public Policy

Team Project Proposal

Jasmine Jha, Maddie Dimarco, Duana Blach, David Gentili

Motivation

We aim to replicate a research project by (Muneton-Santa et al.,2023) through which they tested different machine learning models to predict Multidimensional Poverty using spatial data at the street block level in the city of Medellin, Colombia. Measuring poverty and having the capacity to target the poor is critical for effective policy design and implementation. However, the data available to measure poverty at the neighborhood level typically comes from census information, which are made every ten years and are costly and time-consuming to implement. In this sense, "a low-cost method for estimating multidimensional poverty is a crucial tool for qualifying public policy decision-making" (Muneton-Santa et al.,2023).

Approach and data sources

- We will use the Multidimensional Poverty Index (MPI) calculated at the street block level from the National Department of Statistics of Colombia, available in: *https://geoportal.dane.gov.co/visipm/*. The MPI in this data set will be used as the outcome of interest in our analysis.
- As explanatory variables we will attempt to use the distance from each block to public infrastructure in the city such as police stations, hospitals, schools, universities, churches, airports, banks, and bus stops. This information can be downloaded from Open Street Map using the library(osmdata). The information about this open source portal can be found here:

    - https://www.openstreetmap.org/about
    - https://www.openstreetmap.org/#map=12/10.8738/-74.9280

- The process will be to train different machine learning models on predicting MPI based on the distance of each street block to different public infrastructure items. We will test the best model and will plot the results for the actual and predicted MPI locations.

Technical Challenges

- Calculating the distances of each block to public infrastructure items is going to be a challenge. Figuring out the code and the logic behind getting the distances from each block to each nearest "item" (e.g. school, hospital, etc.).
- So far we have been able to download the geographical data for individual public infrastructure items (e.g. all bus stops in the city of Medellin), but we need to see how can we use all data sets to perform the analysis and plot the maps.