# Final Project - Poverty

STAT 443/543

## Overview

Many social programs have a hard time making sure the right people are given enough aid. It's especially tricky when a program focuses on the poorest segment of the population. The world's poorest typically can't provide the necessary income and expense records to prove that they qualify. Instead, your team will consider a family's observable household attributes like the material of their walls and ceiling, or the assets found in the home, to classify poverty levels.

You'll be graded on two primary elements. The first element is having correct technical content and thorough analysis that addresses the question prompts. You must have working R code and a knittable .Rmd file. The second element is having an appropriate and well-written narrative throughout the report. Your narrative should explain what you are doing and why you are doing it. Your narrative should respond to the question prompts at each state below. The narrative should be written so that someone with a similar statistical background, but no knowledge of this particular project, could read along and reconstruct your analysis. You should choose graphs and tables that support your conclusions. Grammar, spelling, and clarity count for the narrative.

Unlike most homework problems, there isn't a single right answer in this project. There are many choices to make in building a predictive model and I'll be looking to see how well you've explained your choices as well as how you've analyzed the impact of those choices.

## Format

Your final report (max 15 knitted pages) should include (at minimum) the following sections:

- Introduction
- Data Exploration
- The Statistical Model (including model selection)
- Results Summary

## Points and Due Date

- 100 Points Total
  - Analysis: 40 points
  - Narrative: 35 points
  - Predictions: 10 points
  - Peer and Project Review: 15 points
- Report, Review, and Predictions Due: December 17, 4:45pm

# Task and Questions

## Primary Task

Your task for this project is to classify individuals into one of four poverty levels. Your goal in this report is to balance interpretability with predictive power.

For this project, we are only concerned with predictions for "heads of household" (`parentesco` variable below). Data for other individuals (other than "heads of household") are included in this dataset as well, and may or may not be useful for prediction.

## Potential Questions

The following questions are meant to provide a *starting point* for your analysis. As such, these questions are neither explicitly required nor exhaustive.

- Are there any issues with any of the variables and how they are coded?
- Is there any missing data? If so, where is it? How should it be handled?
- Are the variables coded at the "household" level or the "individual" level?
- What useful information could be gathered from non-head-of-household individuals? How could this be mapped to the head-of-household data?
- Are there any pairs or sets of correlated variables? How should these be handled (e.g., selection, grouping, feature engineering)?
- Which variables are most important for prediction?
- How do you measure accuracy for your predictions?

# Data Description

The datasets can be found on Canvas. A description of the datasets can be found below. The main dataset is called "poverty.csv" and the testing data (blinded) is called "poverty-test-blinded.csv".

Each row of the dataset represents a single individual. There are 142 predictor variables and 1 response variable ("Target") in this dataset:

- v2a1, Monthly rent payment
- hacdor, =1 Overcrowding by bedrooms
- rooms, number of all rooms in the house
- hacapo, =1 Overcrowding by rooms
- v14a, =1 has bathroom in the household
- refrig, =1 if the household has refrigerator
- v18q, owns a tablet
- v18q1, number of tablets household owns
- r4h1, Males younger than 12 years of age
- r4h2, Males 12 years of age and older
- r4h3, Total males in the household
- r4m1, Females younger than 12 years of age
- r4m2, Females 12 years of age and older
- r4m3, Total females in the household
- r4t1, persons younger than 12 years of age
- r4t2, persons 12 years of age and older
- r4t3, Total persons in the household
- tamhog, size of the household
- tamviv, number of persons living in the household
- escolari, years of schooling
- rez_esc, Years behind in school
- hhsize, household size

- paredblolad, =1 if predominant material on the outside wall is block or brick
- paredzocalo, “=1 if predominant material on the outside wall is socket (wood, zinc or absbesto”
- paredpreb, =1 if predominant material on the outside wall is prefabricated or cement
- pareddes, =1 if predominant material on the outside wall is waste material
- paredmad, =1 if predominant material on the outside wall is wood
- paredzinc, =1 if predominant material on the outside wall is zink
- paredfibras, =1 if predominant material on the outside wall is natural fibers
- paredother, =1 if predominant material on the outside wall is other
- pisomoscer, =1 if predominant material on the floor is mosaic, ceramic, terrazo”
- pisocemento, =1 if predominant material on the floor is cement
- pisoother, =1 if predominant material on the floor is other
- pisonatur, =1 if predominant material on the floor is natural material
- pisonotiene, =1 if no floor at the household
- pisomadera, =1 if predominant material on the floor is wood
- techozinc, =1 if predominant material on the roof is metal foil or zink
- techoentrepiso, “=1 if predominant material on the roof is fiber cement, mezzanine”
- techocane, =1 if predominant material on the roof is natural fibers
- techootro, =1 if predominant material on the roof is other
- cielorazo, =1 if the house has ceiling
- abastaguadentro, =1 if water provision inside the dwelling
- abastaguafuera, =1 if water provision outside the dwelling
- abastaguano, =1 if no water provision
- public, =1 electricity from CNFL, ICE, ESPH/JASEC
- planpri, =1 electricity from private plant
- noelec, =1 no electricity in the dwelling
- coopele, =1 electricity from cooperative
- sanitario1, =1 no toilet in the dwelling
- sanitario2, =1 toilet connected to sewer or cesspool
- sanitario3, =1 toilet connected to septic tank
- sanitario5, =1 toilet connected to black hole or letrine
- sanitario6, =1 toilet connected to other system
- energcocinar1, =1 no main source of energy used for cooking (no kitchen)
- energcocinar2, =1 main source of energy used for cooking electricity
- energcocinar3, =1 main source of energy used for cooking gas
- energcocinar4, =1 main source of energy used for cooking wood charcoal
- elimbasu1, =1 if rubbish disposal mainly by tanker truck
- elimbasu2, =1 if rubbish disposal mainly by botan hollow or buried
- elimbasu3, =1 if rubbish disposal mainly by burning
- elimbasu4, =1 if rubbish disposal mainly by throwing in an unoccupied space
- elimbasu5, “=1 if rubbish disposal mainly by throwing in river, creek or sea”
- elimbasu6, =1 if rubbish disposal mainly other
- epared1, =1 if walls are bad
- epared2, =1 if walls are regular
- epared3, =1 if walls are good
- etecho1, =1 if roof are bad
- etecho2, =1 if roof are regular
- etecho3, =1 if roof are good
- eviv1, =1 if floor are bad
- eviv2, =1 if floor are regular
- eviv3, =1 if floor are good
- dis, =1 if disable person
- male, =1 if male
- female, =1 if female
- estadocivil1, =1 if less than 10 years old

- estadocivil2, =1 if free or coupled uunion
- estadocivil3, =1 if married
- estadocivil4, =1 if divorced
- estadocivil5, =1 if separated
- estadocivil6, =1 if widow/er
- estadocivil7, =1 if single
- parentesco1, =1 if household head
- parentesco2, =1 if spouse/partner
- parentesco3, =1 if son/doughter
- parentesco4, =1 if stepson/doughter
- parentesco5, =1 if son/doughter in law
- parentesco6, =1 if grandson/doughter
- parentesco7, =1 if mother/father
- parentesco8, =1 if father/mother in law
- parentesco9, =1 if brother/sister
- parentesco10, =1 if brother/sister in law
- parentesco11, =1 if other family member
- parentesco12, =1 if other non family member
- idhogar, Household level identifier
- hogar_nin, Number of children 0 to 19 in household
- hogar_adul, Number of adults in household
- hogar_mayor, # of individuals 65+ in the household
- hogar_total, # of total individuals in the household
- dependency, Dependency rate, calculated = (number of members of the household younger than 19 or older than 64)/(number of member of household between 19 and 64)
- edjefe, years of education of male head of household, based on the interaction of escolari (years of education), head of household and gender, yes=1 and no=0
- edjefa, years of education of female head of household, based on the interaction of escolari (years of education), head of household and gender, yes=1 and no=0
- meaneduc,average years of education for adults (18+)
- instlevel1, =1 no level of education
- instlevel2, =1 incomplete primary
- instlevel3, =1 complete primary
- instlevel4, =1 incomplete academic secondary level
- instlevel5, =1 complete academic secondary level
- instlevel6, =1 incomplete technical secondary level
- instlevel7, =1 complete technical secondary level
- instlevel8, =1 undergraduate and higher education
- instlevel9, =1 postgraduate higher education
- bedrooms, number of bedrooms
- overcrowding, # persons per room
- tipovivi1, =1 own and fully paid house
- tipovivi2, "=1 own, paying in installments"
- tipovivi3, =1 rented
- tipovivi4, =1 precarious
- tipovivi5, "=1 other(assigned, borrowed)"
- computer, =1 if the household has notebook or desktop computer
- television, =1 if the household has TV
- mobilephone, =1 if mobile phone
- qmobilephone, # of mobile phones
- lugar1, =1 region Central
- lugar2, =1 region Chorotega
- lugar3, =1 region Pacfico central
- lugar4, =1 region Brunca

- lugar5, =1 region Huetar Atlantica
- lugar6, =1 region Huetar Norte
- area1, =1 zona urbana
- area2, =2 zona rural
- age, Age in years
- SQBescolari, escolari squared
- SQBage, age squared
- SQBhogar_total, hogar_total squared
- SQBedjefe, edjefe squared
- SQBhogar_nin, hogar_nin squared
- SQBovercrowding, overcrowding squared
- SQBdependency, dependency squared
- SQBmeaned, square of the mean years of education of adults (>=18) in the household
- agesq, Age squared
- Target - the target is an ordinal variable indicating groups of income levels.
  - 1 = extreme poverty
  - 2 = moderate poverty
  - 3 = vulnerable households
  - 4 = non vulnerable households