# Primera entrega de proyecto

**POR:**

Alejandro Herrera Rivera

Juan José Isaza

**MATERIA:**

Introducción a la inteligencia artificial

**PROFESOR:**

Raul Ramos Pollan


UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2021

# 1. Planteamiento del problema

El Banco Interamericano de Desarrollo enfrenta un desafío en su labor de ayudar a las familias más vulnerables cumpliendo con sus objetivos de combatir la pobreza extrema. Para este caso su desafío radica en la clasificación de las familias para sus programas sociales, especialmente aquellas que se encuentran dentro de los estratos más empobrecidos de la población.
En América Latina se utiliza un enfoque conocido como la Prueba de Medios por Poder (PTM) para determinar quiénes son aptos para recibir ayudas del gobierno. El algoritmo usa diversos atributos de los hogares como la calidad y cualidades de la vivienda, la presencia de ciertos bienes, los ingresos promedios, entre otros factores.
El objetivo de este proyecto será utilizar un conjunto de datos con información sobre las características de los hogares en Costa Rica, desarrollando un modelo que clasifique correctamente a las familias que necesiten ayuda social.

# 2. Dataset

El dataset a utilizar proviene de una competencia de kaggle en la cual se proporcionan datos de los hogares en Costa Rica. El dataset está compuesto por un conjunto de archivos .xlsx y .csv que proporcionan la información requerida tales como la descripción de las variables, los datos de entrenamiento y los datos de prueba.

El archivo que contiene los datos de entrenamiento es nombrado como *train.csv* y contiene la siguiente información:

- **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 PacÃƒÂfico central
- **lugar4** =1 region Brunca
- **lugar5** =1 region Huetar AtlÃƒÂ¡ntica
- **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

El archivo llamado test.csv contiene las mismas variables a excepción de la variable objetivo ‘Target’.

## 3. Métricas

La métrica de evaluación principal del modelo es el F1 score. Esta es una métrica muy utilizada en problemas en los que el conjunto de datos a analizar está desbalanceado. Esta métrica combina la precisión y el recall, para obtener un valor mucho más objetivo.

F1 = 2 * ((recall * precision)/(recall + precision))

La métrica de negocio está basada en que es muy grave para la alcaldías, el banco u organizaciones predecir incorrectamente un hogar de extrema pobreza, es decir asignarle una calificación mayor a estas familias, perjudicando la obtención de los beneficios para su clase verdadera, además se espera que las familias pertenecientes realmente a la clase no desfavorecidas no sean catalogadas de un menor nivel socioeconómico, previniendo un gasto innecesario, para equilibrar la situación anterior se utilizará el accuracy, el cual equilibra estas dos posturas, permitiendo evaluar la exactitud del modelo.

## 4. Desempeño

Por la razón anterior se espera que el Recall (sensibilidad) sea mayor al 85%, es decir que el modelo sea capaz de predecir al menos el 85% de las familias que en realidad se encuentran en extrema pobreza, además para la clase no vulnerable se espera al menos una precisión del 85%, es decir que el 85% de las familias catalogadas en esta clase realmente pertenezcan a esta clase, evitando que familias de esta pertenecientes a esta clase sean catalogadas de clases menores.

## 5. Bibliografía

- Fabián Sánchez, Gary Soto, Julia Elliott, Luis Tejerina, Phil Culliton. (2018). Costa Rican Household Poverty Level Prediction. Kaggle. https://kaggle.com/competitions/costa-rican-household-poverty-prediction