**Project2: Income Qualification**

# Import os and Warnings

Graphical user interface, text, application

Description automatically generated

# Problem Statement Scenario:

Many social programs have a hard time making sure the right people are given enough aid. It’s tricky when a program focuses on the poorest segment of the population. This segment of population can’t provide the necessary income and expense records to prove that they qualify.

In Latin America, a popular method called Proxy Means Test (PMT) uses an algorithm to verify income qualification. With PMT, agencies use a model that considers a family’s observable household attributes like the material of their walls and ceiling or the assets found in their homes to classify them and predict their level of need. While this is an improvement, accuracy remains a problem as the region’s population grows and poverty declines.

The Inter-American Development Bank (IDB) believes that new methods beyond traditional econometrics, based on a dataset of Costa Rican household characteristics, might help improve PMT’s performance.

A picture containing graphical user interface

Description automatically generated

## Let us explore our dataset before moving furtherGraphical user interface, text, application Description automatically generated

## Let us identify our target variable

Graphical user interface

Description automatically generated with medium confidence

## Lets Understand the type of data.

Graphical user interface, text, application

Description automatically generated

**We have mixed data types. Specified as below:**

* float64 : 8 variables
* int64 : 130 vriables
* object :5 variables

Graphical user interface, text

Description automatically generated

Below is Data dictionary for above object variables

* ID = Unique ID
* idhogar, Household level identifier
* 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

Graphical user interface

Description automatically generated with medium confidence

A picture containing graphical user interface

Description automatically generated

## **Lets Convert object variables into numerical data.**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Now all data is in numerical form

## **Lets identify variable with 0 varinace**

Graphical user interface, text, application

Description automatically generated

elimbasu5 : 1 if rubbish disposal mainly by throwing in river, creek or sea.

Interpretation: From above it is shown that all values of elimbasu5 is same so there is no variability in dataset therefor we will drop this variable

Check if there are any biases in your dataset.

Text

Description automatically generated

Text, letter

Description automatically generated

Therefore, variables ('r4t3','hogar\_total') have relationship between them. For good result we can use any one of them.

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Therefore, variables ('tipovivi3','v2a1') have relationship between them. For good result we can use any one of them.

Graphical user interface, text, application

Description automatically generated

A picture containing graphical user interface

Description automatically generated

Therefore,variables ('v18q','v18q1') have relationship between them. For good result we can use any one of them.

**Conclusion : Therefore, there is bias in our dataset.**

A picture containing text

Description automatically generated

## Check if there is a house without a family head.

"parentesco1" =1 if household head

Table

Description automatically generated with medium confidence

A picture containing keyboard, computer, electronics, lined

Description automatically generated

Interpretation : **Above cross tab shows 0 male head and 0 female head which implies that there are 435 families with no family head.**

## Count how many null values are existing in columns.

Graphical user interface, text, application, email

Description automatically generated

Interpretation: **There are no null values in Target variable. Now lets proceed further and identify and fillna of other variable.**

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Interpretation and action : ***'v2a1', 'v18q1', 'rez\_esc' have more than 50% null values, because for v18q1, there are families with their own house so they won't pay rent in that case it should be 0 and similar is for v18q1 there can be families with 0 tablets.***

***Istead we can drop a column tipovivi3,v18q***

* tipovivi3, =1 rented
* v18q, owns a tablet

**as v2a1 alone can show both** \*\*as v18q1 alone can show that if respondent owns a tablet or not

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Interpretation : **Now there is no null value in our datset.**

Graphical user interface, text, application

Description automatically generated

## Set the poverty level of the members and the head of the house same in a family.

Now for people below poverty level can be people paying less rent and don't own a house. and it also depends on whether a house is in urban area or rural area.

Graphical user interface, text, application, email

Description automatically generated

* For rural area level if people paying rent less than 8000 is under poverty level.
* For Urban area level if people paying rent less than 140000 is under poverty level.

Graphical user interface, text, application, email

Description automatically generated

Interpretation :

* ***There are total 1242 people above poverty level independent of area whether rural or Urban***
* ***Remaining 1111 people level depends on their area***

**Rural :**

Above poverty level= 445

**Urban :**

Above poverty level =1103

Below poverty level=1081

Graphical user interface, text, application, email

Description automatically generated

## Applying Standard Scalling to dataset

Text

Description automatically generated

## Now we will proceed to model fitting



Lets identify best parameters for our model using GridSearchCv

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

## Lets apply cleaning on test data and then find prediction for that.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Interpretation : **Above is our prediction for test data.**

## Conclusion :

**Using RandomForest Classifier we can predict test\_data with accuracy of 90%.**