

# Costa Rican Household Poverty Level Prediction

Project Members: Anthony Chen (ac8480), Sudharsan Asaithambi (sa6149)

# Purpose of the ADS

Developing countries need to decide who gets access to social development programs.

Proxy Means Test (PMT) is an alternative method to measure the income of a household in developing countries where reliable income data is unavailable.



# Goal of the ADS

The Inter-American Development Bank has compiled a dataset with Costa Rican households with 100+ features that refer to the characteristics of their house, household items they own, their educational qualifications and demographics.

The goal of this ADS is to use more sophisticated means to more accurately identify poverty-stricken households.



# Dataset Characteristics

**Train** : 9557 instances (Label available)

**Test** : 23856 instances (Label not provided)

**Number of feature** : 141

Each row represents an individual. Individuals of the same household are listed in different rows and the same poverty level is assigned for the household.



# Features

- Dwelling Characteristics  
eg: House size, house wall material, ceiling
- Basic Services usage  
eg: type of sanitation, if there is electricity, water system used.
- Human capital  
eg: years of education for household members
- Household Composition  
eg: number of children in household



# Target

## Poverty Level

1 = extreme poverty

2 = moderate poverty

3 = vulnerable households

4 = non vulnerable households



# ADS Implementation

- Mr. Will Koehrsen implemented a notebook “A complete Introduction and Walkthrough”
- Exploratory Data Analysis
- Household level Dataset construction
  - Aggregate attributes of members of each household
  - Descriptive statistics: Sum, average, min, max, std



# ADS Implementation

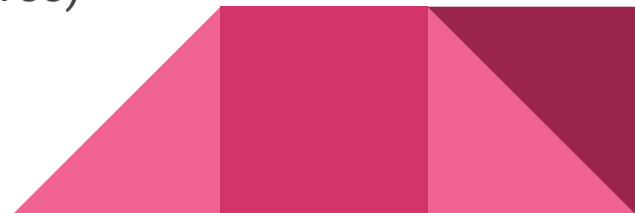
**Hyperparameter Tuning:** LSVC, GNB, MLP, LDA, Random Forest, etc.

**Final Model:** Random Forest Classifier

Number of trees : 100

Criterion : Gini

Max Features : "Auto" i.e  $\sqrt{\text{num\_features}}$





# Nutritional Label

We analyze the ADS in the dimensions below:

- Performance
- Fairness
- Transparency
- Privacy



# Performance Evaluation\*

<b>Original ADS Training set size</b>	<b>9557</b>
<b>Performance Evaluation Training set size</b>	<b>7646</b>
<b>Performance Evaluation Test set size</b>	<b>1911</b>

<b>Confusion Matrix</b>	<b>True Poor</b>	<b>True Not Poor</b>
<b>Predicted Poor</b>	67	146
<b>Predicted Not Poor</b>	23	259

<b>Accuracy</b>	<b>0.72</b>
<b>Precision</b>	<b>0.71</b>
<b>Recall</b>	<b>0.94</b>
<b>False Positive Rate</b>	<b>0.69</b>
<b>False Negative Rate</b>	<b>0.06</b>
<b>F1-Score</b>	<b>0.81</b>

\*RF was trained on 80% of the original training data  
Performance was evaluated on 20% of the original training data

# Fairness

We analyzed fairness of the ADS on the pivots below :

- Gender
- Disability
- Geographical area (Urban, Rural)
- Age
- Marital Status



# Fairness

We perform analysis on the dataset by computing

- Composition
  - Base rate of sub-populations
- Disparate Impact
  - Composition of sub-populations for poor/non-poor households.
- Model Performance
  - Variation in model performance for sub-populations



# Fairness

We observe disparate impact on Gender, Disability, Geographical area. We do not observe disparate impact with Marital Status.

For instance, we find that only 66% of the urban population are poor while 72% of the survey participants are urban.

Model performance is lower for Female, Disabled and Rural households.

For instance, households with disabled members have a model accuracy of 0.61 vs 0.75 for households of non-disabled members.



## Fairness on Protected Attribute - Disability: Composition & Disparate Impact

Disability Status	Count
Disabled	1904 (6%)
Not Disabled	31509 (94%)

Socioeconomic Status Classification/ Disability Status	Not Disabled (Privileged Group)	Disabled (Unprivileged Group)
Poor	93%	7%
Not Poor	95%	5%

## Fairness on Protected Attribute - Disability: ADS Performance Metrics

<b>Disability Status/ Metrics</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>FPR</b>	<b>FNR</b>
<b>Not Disabled (Privileged)</b>	0.75	0.76	0.94	0.68	0.06
<b>Disabled (Unprivileged)</b>	0.61	0.55	0.95	0.69	0.05

# Transparency & Interpretability

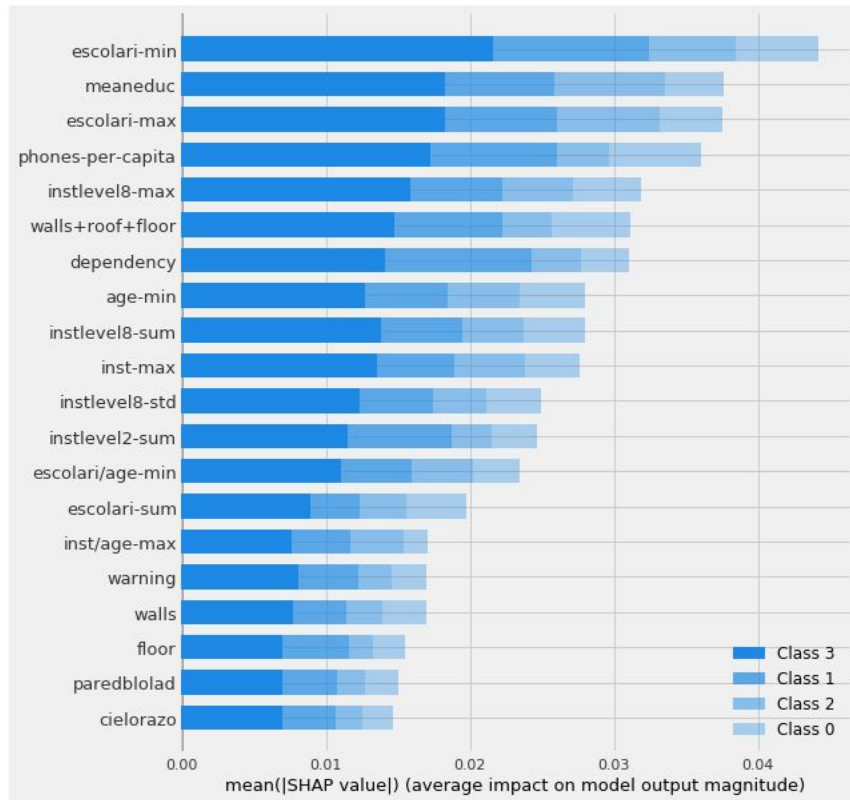
Given that the ADS deals with sensitive information about its survey participants, it should be more transparent for individuals.

As shown in the SHAP plot:

- Education level
- owning a phone
- having quality living space
- dependency rate
- minimum age of a resident in a household

provide the most weight in classifying whether a person is non-vulnerable.

The minimum years of schooling of the entire household provides a strong indication whether the family needs further social aid.

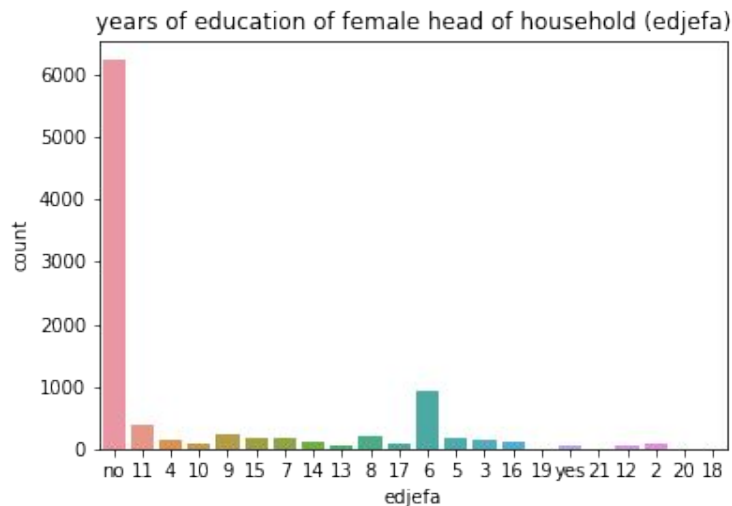





# Privacy

A privacy concern inherent in the data, is the potential for linkage attacks to re-identify individuals.

We have shown that for the particular feature 'edjefa' there are only 2 individuals with the head of household having 20 years of education. Many other features can be used to further obtain details on where they live, and their socioeconomic status.



# Summary

- Privacy concerns in this ADS
  - Non-robustness in terms of fairness for certain protected attributes: gender & disability.
  - Optimizing on the false negative rate will allow more people to receive aid.
  - Optimizing on the false positive rate will allow the IDB to conserve more of its budget for extremely poor individuals.
  - We would not be comfortable deploying this ADS in the public sector without a confidentiality preprocessing method such as Data Synthesizer
  - Recommended to have a cost-efficient approach to manage the inflow and outflow of data, along with careful outreach to poor demographics.
  - Awareness that some participants may be new immigrants.
- 

# References

<https://www.kaggle.com/code/willkoehrsen/a-complete-introduction-and-walkthrough>

<https://publications.iadb.org/publications/english/document/Social-Assistance-Poverty-and-Equity-in-the-Dominican-Republic.pdf>

<https://olc.worldbank.org/sites/default/files/1.pdf>

<https://jpia.princeton.edu/news/accuracy-proxy-means-tests-immigrant-populations-case-study-colombia>

