



**MACHINE LEARNING:- PROJECT**

DESCRIPTION

Identify the level of income qualification needed for the families in Latin America.

**Problem\_Statement\_Scenario:**  
Many social programs have a hard time ensuring that the right people are given enough aid. It’s tricky when a program focuses on the poorest segment of the population. This segment of the 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.

Following actions should be performed:

* Identify the output variable.
* Understand the type of data.
* Check if there are any biases in your dataset.
* Check whether all members of the house have the same poverty level.
* Check if there is a house without a family head.
* Set poverty level of the members and the head of the house within a family.
* Count how many null values are existing in columns.
* Remove null value rows of the target variable.
* Predict the accuracy using random forest classifier.
* Check the accuracy using random forest with cross validation.

The following would be the target variable values:

• 1 = extreme poverty

• 2 = moderate poverty

• 3 = vulnerable households

• 4 = non vulnerable households

First, we applied a common target label to each member of a household. As described in the initial background, we concluded that the analysis of only the head of a household is done in order to predict the target label for that particular household.

Consequently, each member of a household should have the same target label as that of the head as they all live in the same household.

The second step to clean the data, involved filling up missing values. To achieve this, we narrowed down the features like *number of house owned*, *monthly rent payment* and *years in school* as they were the ones with missing values. We moved on to the *monthly rent payment* feature and here we put 0 as a value in a missing value row, in case the respective household owns the property. For the other households, we can leave the missing values to be imputed later, but for now we added a flag (Boolean) column indicating that these households have missing values. Lastly, we moved on to the feature *years behind in school* and for this we first identified that this variable is defined for the age group of 7-19 year olds only. Hence, in case of missing values, we put 0 as a value for those data points which do not fall into the 7-19 age group and otherwise we flagged it.

***Feature Engineering***

*1. Removing Squared Variables:* There are some squared variables included in the dataset to help with feature engineering in case of nonlinear data modelling. But since we are using more complex models, these squared variables are redundant and are removed. Including these variables in the features may hurt the result as they are highly correlated with their non-squared version.

*2. Remove highly correlated household and individual variables:* We saw the correlations between different variables to see which ones are highly correlated. If there are any that are too highly correlated, then we might want to remove one of the pair of highly correlated variables. We considered pairs with correlation more than 0.95 as fit for removing.

Extreme poverty is smallest count in train data set. Hence data is baised.

The correlation matrix defines it very well.

3. *Creating Ordinal Variable:* We tried to construct some ordinal variables by compressing several related Boolean variables. We mapped the description based on the data description given to us in the problem. For example, in case of electricity: -

• 0: No electricity

• 1: Electricity from cooperative

• 2: Electricity from CNFL, ICA, ESPH/JASEC

*4. Feature Aggregation:* In order to incorporate the individual data into the household data, we need to aggregate it for each household. The simplest way to do this is to group by the *family*

MODEL

As a first attempt at looking into our model, we can visualize the distribution of predicted labels on the test data. We would expect these to show the same distribution as on the training data. Since we are concerned with household predictions, we'll look at only the predictions for each house

and compare with that in the training data. This is necessary because the raw counts differ in

the training and testing data. Furthermore, the predicted distribution looks close to the training distribution although there are some differences Now, after our confusion matrix from RF we derived for the poverty level, we can conclude that our model really does not do that well for classes other than Non Vulnerable. It only correctly identifies very little % of the Vulnerable households, classifying more of them as moderate or non-vulnerable. Overall, these results show that imbalanced classification problems with relatively few observations are very difficult.

There are some methods we can take to try and counter this such as oversampling or training multiple models on different sections of the data, but at the end of the day, the most effective method may be to gather more data and perform RFCLASSIFIER with CROSS VALIDATION on it as it will provide with confusion matrix that has removed oversampling and data is sampled in variations and tested with many k folds. The model is efficient and thus concluded our objective that From the figure in the end of source code, meaneduc, dependency, overcrowding has significant influence on the model.