Household Poverty Mining in Costa Rica

Grzegorzthehero, W. Z.

Erasmus University, Rotterdam, The Netherlands

Abstract. This paper

Keywords: Data Mining · Random Forest · Poverty · Costa Rica.

1 Introduction

Poverty is prevalent among a large part of the world population. In 1990, about 28% of the developing world's population, or about 1.2 billion people, had to live for less than one dollar for a person per day. As a result of this, poverty alleviation is at the heart of the discourse on international development. As a result, the Millennium Development Goal of 2000 was the first Millennium Development Goal to reduce the share of the absolute poverty by half by 2015, namely 14%. [7] This target was attained already in 2010. Most recent estimates suggest that in 2015 the percentage of people living on less than 1,90 US dollar decreased to 10 percent. [**?**,**?**]Nevertheless, the number of people living in extreme poverty is still alarmingly high.

Despite a long history of policies that attempt to address poverty, many of these programmes have been unsuccessful. This is often a result of a lack of available information to governments or other organizations that attempt to implement such programmes. For example, governments often do not possess information about actual earnings of households. As a result, they have to estimate these figures based on other data such as owned assets, and consequently have to decide whether or not to classify the household as poor and make them eligible for an aid programme. As such, poverty alleviation programmes often fail to target the right households, due to a lack of data. [3]

One of the countries that is faced with these problems is the Republic of Costa Rica, a country where 1,1 Million people live in poverty, and the majority of them is situated in rural areas.[4] Despite the fact that Costa Rica has one of the lowest poverty rates in Central America[5] and there are several programmes that

[3] Todd et al. 2017 Is it fair to say that most social programmes dont work? https://80000hours.org/articles/effective-social-program/, access date: 19.10.2018

[4] Gibson, P. (2017) Poverty in Costa Rica https://borgenproject.org/poverty-in-costa-rica-2/,19.11.2017, access date: 11.10.2018

[5] Gibson, P. (2017) Poverty in Costa Rica https://borgenproject.org/poverty-in-costa-rica-2/, access date: 11.10.2018

aim to help the households which are below the poverty line, an improvement of the distribution of social aid is being sought through the use of data. [6]

In the case of poverty alleviation, we believe that the field of data mining offers promising opportunities for policy-makers to analyze the vast amount of data they have access to, and identify patterns. As such, they could improve their policies by incorporating this information. However, this field remains largely unexplored by policy-makers. We wish to challenge this notion by attempting to produce a model that is able to predict poverty levels based on simple household survey data.

There are numerous different approaches to data mining, each with their own strengths and weaknesses. We will investigate this problem through the use of the *random forest* method. We seek to answer the question of how to predict an individual's level of household poverty based on household survey data, using random forest. Furthermore, we will investigate which factors we can identify as important in predicting household poverty, what the accuracy is of a random forest method on this problem, and, where the random forest algorithm fails to predict correctly, we will investigate why this is the case.

We will proceed by firstly presenting the current research on the topic. Then we will go into our methodology, and provide an evaluation of the results of this methodology Lastly, we will offer our conclusions and address our research questions.

2 Literature review

2.1 Poverty predictions based on household survey information

One of the easiest but unfortunately expensive and time-consuming approaches is to predict poverty levels based on data gathered during household surveys. They very often provide reliable information about personal expenditures and assets possessed by a household. Many researchers apply the data during their studies in order to find the key determinants of poverty in developing countries and to be able to decrease its level.

Mok et al. (2007) conduct a research on the determinants of urban poverty in Malaysia using the data from Household Expenditure Survey (HES) from 2005. [10] They apply the binomial logistic regression where 0 corresponds to a household being above the Malaysias official poverty line and 1 to being below it. The study concludes that in order to eradicate the urban poverty in Malaysia the government has to put a strong emphasis on promoting education which is one of the most important determinants in Malaysia. The research supports the

[6] Costa Rican Household Poverty Level Prediction https://www.kaggle.com/c/costa-rican-household-poverty-prediction access date: 19.10.2018

statement that migrant workers tend to earn significantly less than Malaysian citizens.[7]

In another study [1] examine key determinants of poverty in Kenya using the data gathered during Demographic and Health Surveys (DHS) conducted in year 2003. In order to reduce the number of explanatory variables in the model the principal components analysis (PCA) is applied. Variables with the best explanatory power like education, religion or ethnicity are chosen and further studied using logistic regression as the model. Based on the data the model predicts whether someone is poor or otherwise.[1]

Although it should be easy to predict poverty and find its main causes, it is very often problematic due to the fact that in many underdeveloped countries there is a lack of reliable surveys on household assets and expenditure. Moreover, some poor countries cannot afford to conduct them often what results in a lack of data. Hence the researchers had to come up with alternative approaches.

2.2 Poverty predictions based on satellite imagery

The following studies try to predict poverty levels based on satellite images which are the part of public domain accessible by everyone. High resolution photographs from above make it possible to reliably forecast the economic development level in a region.

In order to estimate poverty in rural regions in India [11] make use of satellite images and apply a two-step approach. Firstly, they train a multi-task fully convolutional model to predict the source of drinking water, roof material and source of lighting. Secondly, they create a new model to assess the household income level using the features observed in the first step as input. After establishing three income categories they classify a household as poor when it belongs to the category with lowest income. Finally, the results are compared to the data collected during official Census in 2011 and prove to be similar in performance.[11]

In another study Jean et al. (2016) use multistep transfer learning approach. In the beginning the convolutional neural networks are trained to predict the light intensity at night with daytime satellite imagery as input. Then the ridge regression model is trained based on features obtained from daytime satellite images along with mean village-level values from data gathered in surveys to estimate mean village-level assets and expenditures. In the end the model is able to predict the per capita outcomes on the village-level based on the features from daytime images with a high level of accuracy.[6]

3 Method

3.1 Random Forest

In general ensemble methods are techniques which produce better results than simple algorithms by creating and combining multiple models.[12] A class of ensemble methods, which is utilized in designing decision tree classifiers, is Random Forest. It was introduced by Leo Breiman in 2001 as as extension of his bagging idea. As the name suggests, this method operates by combining multiple decision trees and pooling predictions based on their individual results. This procedure gives more accurate and robust predictions than a single decision tree and overcomes the problem of overfitting[11] the dataset. Breiman (2001) explains that by the Strong Law of Large Numbers the random forests converge what solves the problem of overfitting.[2]

In a Random Forest classification algorithm, decision trees use random vectors, which are generated from a fixed probability distribution. In this manner, randomness is ensured by growing decision trees that have access only to a random part of the training set. This concept is called bootstrap aggregation, and trains M different trees on different data subsets. Replacement is also included which makes it possible for certain training points to be selected more than once. In contrast to bootstrap aggregation, random forest chooses only a subset of attributes randomly and the split on a trees node is done based on the best feature. As a result, fully grown trees are created, representing a certain outcome based on the features used. In the end, a majority voting scheme determines the outcome of the random forest, given the individual results of the decision trees.

3.2 Ordinal Forest

While the random forest classification method has been successfully applied in many cases with numerical or categorical response variables, for problems with ordinal response variables, such as the one in this research, this is not the case. The algorithm takes into account ordinality in predictor attributes, allowing only splits between adjacent categories, but it treats ordinal response variables as nominal variables. This is undesirable because it could lead to a loss of important information. Furthermore, [13] have demonstrated that the the original version of random forest by [2] has certain inherent biases in its variable selection algorithm.

In light of this, the random forest implementation developed by [4] offers interesting possibilities. It has been successfully applied to ordinal data by [5] among others. There are two main differences between the two algorithms. Fristly, whereas the algorithm by [2] uses metrics such as the Gini score to decide which attribute to split on in building the individuals trees, the version of [4] uses conditional inference tests to select the best split in a tree. This means that,

[11] modelling error which is a result of trying too closely to fit the training data which has either some level of error or noise within it

for each split, a subset (the size of the subset is a hyper-parameter) of attributes is randomly selected. Then, a statistical test is performed for the association of each attribute with the response variable. This results in a p-value. The attribute with smallest p-value is selected. In the case of an ordinal response variable, the variable is transformed to a metric scale. Thus, this method is appropriate for data with an ordinal response variable because information about the ordinal nature of the variable is taken into account when determining the splits during the construction of the decision tree. In contrast, traditional random forest algorithms use measures such as Gini or entropy that treat the classes of the response variable as unrelated to one another.

The second difference pertains to the way in which the class prediction is obtained. In the classical randoms forest algorithm by [2], predicted class probabilities are computed based on how many trees 'vote' for a class. In the version by [4], however, class probabilities are obtained by taking the average of the class probabilities predicted by the individual trees. Eventually the class of an observation is predicted in the following way:

$$\hat{Y} = r \leftrightarrow \hat{P}(Y = r) = \max_{l=1,...,k} \hat{P}(Y = l) \tag{1}$$

in this equation, $r = 1, ..., k$ and $l = 1, ..., k$ are classes. The class that is predicted in the end thus corresponds to the mode of the distribution of predicted probabilities.

In order to improve the accuracy of our model, we adjust the algorithm slightly by adding weights to the sampling method used for growing the individual trees. As is the case in the classical random forest algorithm, each tree is grown using only a portion of the data. Since our dataset contains imbalanced classes, if we use simple probability sampling we risk over-representing the largest class. This would result in our model being trained very well for the largest class, but not for the smaller classes. To prevent this, we introduce weights into the sampling process. Observations from each class are sampled with probability $\frac{1}{n}$, where n is the total number of observations in that class.

3.3 Data

We obtained all of the data used in this research from the online data mining platform *Kaggle*[13]. The dataset was posted by the Inter-American Development Bank and contains household survey data from a representative sample of the population of Costa Rica. Each observation represents an individual, but the class (level of poverty) is measured on the level of the household to which the observation belongs. The original dataset consists of 9557 observations of 143 variables.

[13] https://www.kaggle.com/edimaudo/costa-rica

Preprocessing and Feature Selection Before training the model we preprocessed the dataset, paying close attention to the description of the data by the Inter-American Development Bank. Firstly, we removed measurement errors, inconsistent values, and irrelevant features. Then we deleted observations with missing values, duplicate attributes, and attributes with a high proportion of missing values. Furthermore, we aggregated a large number of dummy variables into categorical variables. We then checked for correlation between attributes, using the correlation coefficient for continuous attributes and chi-squared tests for categorical attributes, and removed attributes that were strongly correlated with other attributes.

In order to reduce the dimensionality of our data, we select only a subset of our original features to use in the model. The *Boruta* algorithm allows us to identify which variables are important in predicting the class, and which are redundant. [8] The algorithm works as a wrapper around the random forest algorithm. It uses a test criterion to compare the original features with newly created variables (shadow variables) obtained from the shuffle of some of the original attributes. The feature importance of these shadow variables is subsequently used as a threshold for the original features, using an approximation to a z-test. [8] The algorithm works in the following way[1]:

1. Extend the information system by adding copies of all variables (the information system is always extended by at least 5 shadow attributes, even if the number of attributes in the original set is lower than 5).
2. Shuffle the added attributes to remove their correlations with the response.
3. Run a random forest classifier on the extended information system and gather the Z scores computed.
4. Find the maximum Z score among shadow attributes (MZSA), and then assign a hit to every attribute that scored better than MZSA.
5. For each attribute with undetermined importance perform a two-sided test of equality with the MZSA.
6. Deem the attributes which have importance significantly lower than MZSA as unimportant and permanently remove them from the information system.
7. Deem the attributes which have importance significantly higher than MZSA as important.
8. Remove all shadow attributes.
9. Repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

Executing the algorithm gives us the results in figure 1. Based on these results, we made two subsets, one subset including the features we thought stood out from the plot as being the most important (12 features in total), another

[1] taken directly from [8]

one containing the most important half of the features (24 in total). After some preliminary analyses, we found no significant increase in accuracy resulting from using 24 features as opposed to 12. Therefore, in order to reduce dimensionality and follow Occam's razor, we decided to use the subset with 12 features. A table containing a description of all variables can be found in the appendix, the distribution of the class variable is displayed in figure 2. The variable ranges from 1 (most poor) to 4 (least poor). The distribution clearly indicates class imbalance, which have taken into account in our methods (see section 3.2).

3.4 Data Analysis

We have conducted all of our data analyses in R. All of the functionalities of the random forest implementation by [4] are available in an R package called 'party'. The Boruta algorithm is available in a package called 'Boruta'.

Fig. 1. Plot of the results from running the Boruta algorithm

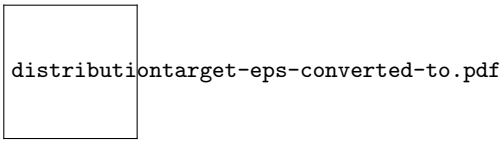

Fig. 2. Plot of the results from running the Boruta algorithm

4 Evaluation

4.1 Training and Testing our Model

In order to evaluate the performance of our model, we split our dataset into a training set used for training the model, and a test set used for evaluating the model's performance. The training set contains 75% of the total data [2], the remaining 25% of the data makes up the test set[3]. We divide observations by

[2] $n = 7112$
[3] $n = 2371$

stratified sampling, because of the class imbalance in our data set. The stratified sampling method ensures that the class proportions in the test and training set are the as in the original data set. Furthermore, we assign a relatively large proportion of the data to the test set, in order to ensure that it can yield statistically significant results, even for the smallest class. We present several performance metrics

4.2 Performance Metrics

In order to be able to assess the ability of the model to predict the level of poverty for previously unseen observations in the data set, we will present performance metrics of the model when tested on the test data set. Because of the structure of the data being imbalanced, assessing the performance of the model is not straightforward. Therefore, besides simple accuracy we include four different metrics that are more sensitive to class imbalance.

Accuracy Firstly, the model's accuracy is shown. Accuracy is the fraction of the data that the model accurately predicted (see equation 1). Due to the class imbalance and the presence of multiple classes in the data set, accuracy alone does not give us a complete picture of the model's performance. Therefore, we use several additional performance metrics.

$$\text{accuracy}, a = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \tag{2}$$

Precision, Recall, and F_1 Measure Precision and recall are widely used metrics for evaluating the performance of a model on a dataset with imbalanced classes. [12] The formal definitions of the measures are given in equations 2 and 3, where TP refers to true positive (, FP to false positive, and FN to false negative

$$\text{precision}, p = \frac{TP}{TP + FP} \tag{3}$$

The precision metric indicates the records that are correctly classified as a class, as a proportion of the total number of records (correctly or falsely) classified as that class.

$$\text{recall}, r = \frac{TP}{TP + FN} \tag{4}$$

The recall metric, on the other hand, indicates the records that are correctly classified as a class, as a proportion of the total number of records belonging to that class.

In general, models that have high precision have lower recall and vice versa. Ideally, however, a model would perform well both in terms of precision and recall. We can assess this using another metric, the F_1 measure, which is essentially the harmonic mean of precision and recall (see equation 4), and thus summarizes the model's precision and recall for a given class. [12]

$$F_1 = \frac{2rp}{r+p} \tag{5}$$

Kappa Metric The Kappa (κ) metric is insightful in situations with class imbalance as well. This metric (see equation 5) compares a model's observed accuracy (p_o) to its expected accuracy (p_e). In other words, it compares the performance of the model to the performance that any model would be expected to achieve based on the distribution of classes in the data. The formula for expected accuracy is given in equation 6. In this formula, k represents a class, N the total number of records, and n_k the total number of records with class k. Values for κ can range from 0 to 1, with 0 meaning no agreement between the model's predictions and the actual values, and 1 meaning complete agreement. [3] There is no standard interpretation of κ. It suggested in [9] to interpret 0.0-0.20 as slight agreement, 0.21-0.40 as fair agreement, 0.41-0.60 as moderate agreement, 0.61-0.80 as substantial agreement, and 0.81-1.0 as almost perfect agreement. [9]

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6}$$

$$p_e = \frac{1}{N^2} \sum_k n_{k \text{ predicted}} \times n_{k \text{ actual}} \tag{7}$$

4.3 Evaluating the Model's Performance

Table 1. Confusion matrix of the model on the test set

		Predicted Class			
		1	2	3	4
Actual Class	1	**128**	29	8	27
	2	5	**283**	14	101
	3	1	40	**167**	101
	4	10	47	39	**1371**

Applying our model to the test set and comparing the predicted results to the actual values, yields the confusion matrix shown in table 1. The performance metrics calculated from these results are displayed in table 2. At first glance, we can say that the overall performance of our model is fairly good, obtaining an accuracy of 0.82. However, at the same time we do observe interclass differences in the model's performance. Comparing F_1-scores shows that the model's performance is best for class 4, the largest class. For the smallest class 1, however, the performance in terms of F_1-score is the second best. For this class, we must note that the model's recall is much better than its precision. Concretely, this

Table 2. Performance metrics of the model on the test set

	Class			
	1	2	3	4
Accuracy (95% Conf. Interval)	0.82 (0.8,0.84)			
Precision	0.67	0.7	0.54	0.93
Recall	0.89	0.71	0.73	0.86
F_1	0.76	0.71	0.62	0.89
κ	0.67			
N	192	403	309	1467

means that there are more persons incorrectly classified as not having class 1 (false negatives), than persons incorrectly classified as having class 1 (false positives). Furthermore, the κ statistic of the model is 0.67, which we can interpret as indicating substantial agreement between the model and the actual values. [9]

Table 2 also shows that the model has the worst performance for class 3. We might seek the reason for this in the data. The class 3 represents 'vulnerable households' that are in-between being in the middle-class and being in moderate poverty. It is imaginable that such households exhibit characteristics of both class 2 and 4, making it difficult to distinguish them as class 3. Another reason could be that there is in facts an attribute which correlates highly with class 3, but it is missing from our training data. Indeed the confusion matrix shows that the model mainly mistakes class 3 records for being class 4. Besides this issue, table 1 also shows that despite a fairly high F_1-score for class 2, the model incorrectly classifies 101 class 2 record as class 4. While proportionally this error is not particularly large, the fact that the two categories are non-adjacent is concerning.

4.4 Feature Importance

To evaluate the importance of each feature in the overall model, we will also present feature importance metrics. Following Janitza et al. (2016), who present novel ways to compute feature importance for random forest models with ordinal response variables. [5] The importance of a feature is usually derived based on the difference in accuracy when a random permutation of that feature is used instead. For nominal response variables, most measures of feature importance are based on the error rate (similar to the measure of accuracy we presented in equation 1). This, however, ignores the ordinal nature of the response variable and is therefore undesirable for our model. Janitza et al. (2016) find that a measure based on the ranked probability score (RPS) gives the best results for data with ordinal response variables. Equation **??** shows the form of the RPS for the classification of $i = 1, ..., n$ observations with true classes Y_1 into k number of response classes. RPS sums for each observation i the squared difference between the predicted probability $\hat{\pi}$ of observation i to be in class $r = 1, ..., k$, and the

indicator function I taking value 1 if $Y_i \leq r$ = true. It then sums these sums of squared distances for all observations i and divides this number by n.

$$\text{RPS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{r=1}^{k} (\hat{\pi}_i(r) - I(Y_i \leq r))^2 \tag{8}$$

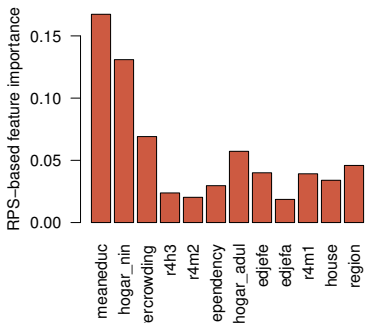

Fig. 3. Importance of features based on the Ranked Probability Score (RPS, see equation **??**), for variable descriptions see appendix.

Figure 3 shows that in our model, the most important features for predicting an observation's level of household poverty are the average years of education for adults (18+) in the household, the number of children (0-19) in the household, and the number of persons per room in the household.

5 Conclusion

We started this research based on the question of how to predict an individual's level of household poverty based on household characteristics. To this end, we used survey data from Costa Rican households. We have found that a random forest model with ordinal regression trees, developed by [4], is able to predict household poverty levels in Costa Rica with moderate to good accuracy [**?**]. Our model performs best for predicting observations in the class with lowest level of poverty, and in the class with the highest level of poverty. However, the model's performance is worst for predicting the class of so-called vulnerable households. Furthermore, we have identified the average years of education for adults (18+) in the household, the number of children (0-19) in the household, and the number of persons per room in the household as the three features with the largest effect on household poverty in our model. While it is plausible that education has a direct effect on poverty, we should be careful with drawing such conclusions. It

could be that number of children in the household, for example, does not have a direct effect on poverty, but instead 'translates' the effect of a different cause. In our research we have found the extension of random forest by Hothorn et al. (2006) to perform well on a classification problems with an ordinal class variable and class imbalance. A suggestion for future research may be to further explore the working of this algorithm on different datasets with ordinal class variables.

References

1. Achia, T.N.O., Wangombe, A., Khadioli, N.: A Logistic Regression Model to Identify Key Determinants of Poverty Using Demographic and Health Survey Data. European Journal of Social Sciences **13**(1), 38–46 (2010)
2. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (oct 2001). https://doi.org/10.1023/A:1010933404324, https://doi.org/10.1023/A:1010933404324
3. García, S., Fernández, A., Luengo, J., Herrera, F.: A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Computing **13**(10), 959 (dec 2008). https://doi.org/10.1007/s00500-008-0392-y, https://doi.org/10.1007/s00500-008-0392-y
4. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics **15**(3), 651–674 (2006). https://doi.org/10.1198/106186006X133933, https://doi.org/10.1198/106186006X133933
5. Janitza, S., Tutz, G., Boulesteix, A.L.: Random forest for ordinal responses: Prediction and variable selection. Computational Statistics and Data Analysis (2016). https://doi.org/10.1016/j.csda.2015.10.005
6. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science **353**(6301), 790–794 (2016). https://doi.org/10.1126/science.aaf7894, http://science.sciencemag.org/content/353/6301/790
7. Klasen, S.: Armutsreduzierung im Zeitalter der Globalisierung. Tech. rep., Ibero-America Institute for Economic Research (2006)
8. Kursa, M.B., Rudnicki, W.R.: Feature Selection with the Boruta Package. Journal of Statistical Software **36**(11) (2010)
9. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics (1977). https://doi.org/10.2307/2529310
10. Mok, T.Y., C, G., A, S.: The Determinants of Urban Household Poverty in Malaysia. Journal of Social Sciences **3** (2007)
11. Pandey, S.M., Agarwal, T., Krishnan, N.C.: Multi-Task Deep Learning for Predicting Poverty from Satellite Images. Tech. rep., The Thirtieth AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-18), Ropar, India (2018)
12. Pang-Ning, T., Steinbach, M., Kumar, V.: Introduction to Data Mining (2006). https://doi.org/10.1016/0022-4405(81)90007-8
13. Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics **8**(1), 25 (jan 2007). https://doi.org/10.1186/1471-2105-8-25, https://doi.org/10.1186/1471-2105-8-25

A Feature Description

Feature Name	Type	Range	Description
Target	Ordinal	1-4	Feature indicating groups of income levels. 1 = extreme poverty 2 = moderate poverty 3 = vulnerable households 4 = non vulnerable households
meaneduc	Discrete	0-37, median = 9	Average years of education for adults (18+)
hogar_nin	Discrete	0-9, median = 1	Number of children 0 to 19 in household
overcrowding	Discrete	2-38, median = 17	Number of square meters in the house per person
r4h3	Discrete	0-8, median = 2	Total males in the household
r4m2	Discrete	0-6, median = 2	Females 12 years of age and older
dependency	Continuous	1-31, median = 21	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)
hogar_adul	Discrete	1-9, median = 2	Number of adults in the household
edjefe	Discrete	1-22, median = 18	Years of education of male head of household
edjefa	Discrete	1-22, median = 21	Years of education of female head of household
r4m1	Discrete	0-6, median = 0	Females younger than 12 years of age
house	Categorical	5 categories	Type of house ownership, categories are: own, own but paying in instalments, precarious, rented, other
region	Categorical	6 categories	Region where the household lives