Final Project

Maggie Wesolowski, Maggie Grethel, Jayden Koenig, Sarah Rodriguez

2024-11-26

Introduction

The goal of this project is to classify households into one of four poverty levels based on a range of socioeconomic and demographic variables. The dataset provides both household and individual level variables. Some example variables include: the status of the flooring of a house, number of individuals living in a house, house location region, and electricity and toilet status in a house. We will be focusing on the heads of household to develop a predictive model that balances interpretability with predictive power. This classification can help identify vulnerable groups and guide social policies. Key considerations include handling missing data and ensuring appropriate feature representation.

Exploring Data

During exploration, we identified missing values in several variables, such as v2a1 (monthly rent) and rez_esc (years behind in school). Missing values in v2a1 often occur when households own their homes and do not pay rent, suggesting that these missing values can reasonably be imputed as 0. For other variables with missing data, we need to determine if the data is missing at random. Depending on the nature of the missingness, strategies such as mean/mode imputation or introducing a new “unknown” category will be applied.

Exploring Missing Values

- If the value of v18q is 0, then the value in v18q1 is NA. Therefore, we convert NAs in v18q1 to be 0, meaning the number of tablets a household owns in 0.

- The missing values in v2a1 (monthly rent) are because the house is owned and fully paid. When tipovivi1 is 1, v2a1 is NA.

Fixing Coding Error

- ‘Edjefe’ is supposed to be the years of education of male head of household. The variable should only contain numeric values for male heads of household. However, it contains ‘yes’ and ‘no’. With ‘no’ meaning that person is neither a male nor the head of household. And ‘yes’ being that person is male and head of household. This is the same for the variable ‘edjefa’.

- There is 120 cases where ‘edeje’ and ‘edjefa’ were coded incorrectly to fix this we check where that inconsistency is and change it. For instance, if there is a ‘yes’ in ‘edjefe’ but that person is neither male nor head of household, the ‘yes’ is change to a ‘no’. And we will do the same for the ‘edjefa’ variable to ensure that the data is consistent.

- Now that the variables have been corrected and are consistent the next thing we need to do is replace the ‘yes’ in each column with the years of education from the ‘escolari’ variable. If there is a ‘no’ then the value will be NA. In this case it makes sense to have a missing value because it is not relevant to know the years of education if that person does not meet the requirements for that variable.

Fixing Redundencies

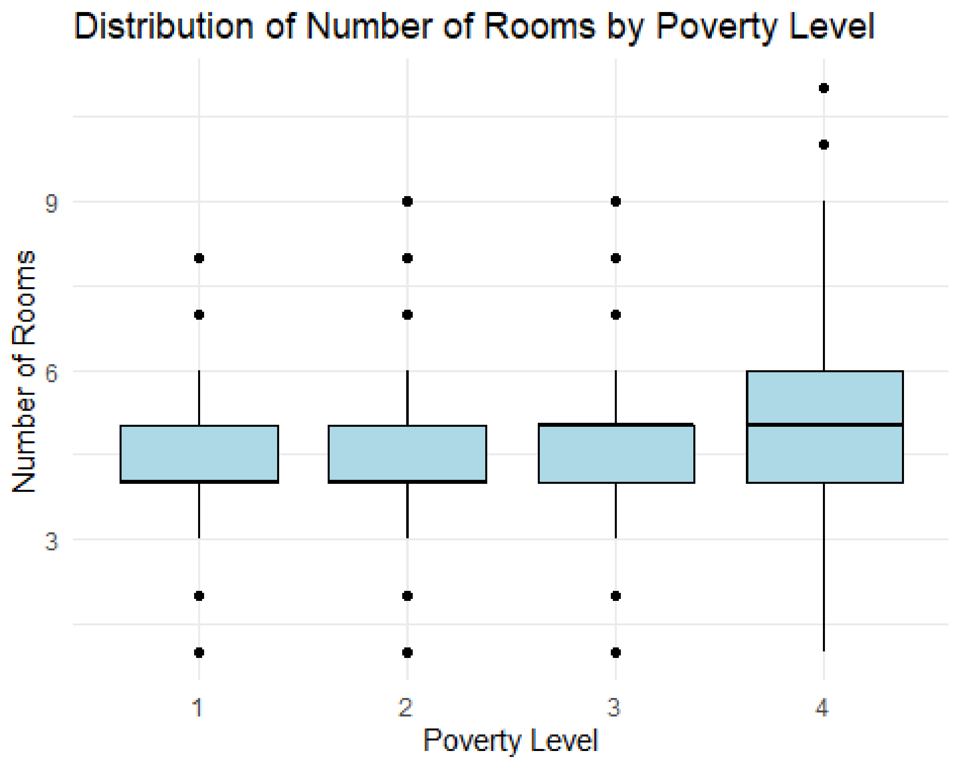
Having separate columns for female and male is redundant, so we made one column for gender with male = 1 and 0 = female.

Household vs. Individual-Level Variables

The data includes variables at two levels:
- Household-level: Variables like rooms, electricity type, and toilet type describe the collective living conditions of the household.
- Individual-level: Variables like age, gender, and years of education provide details about each household member.

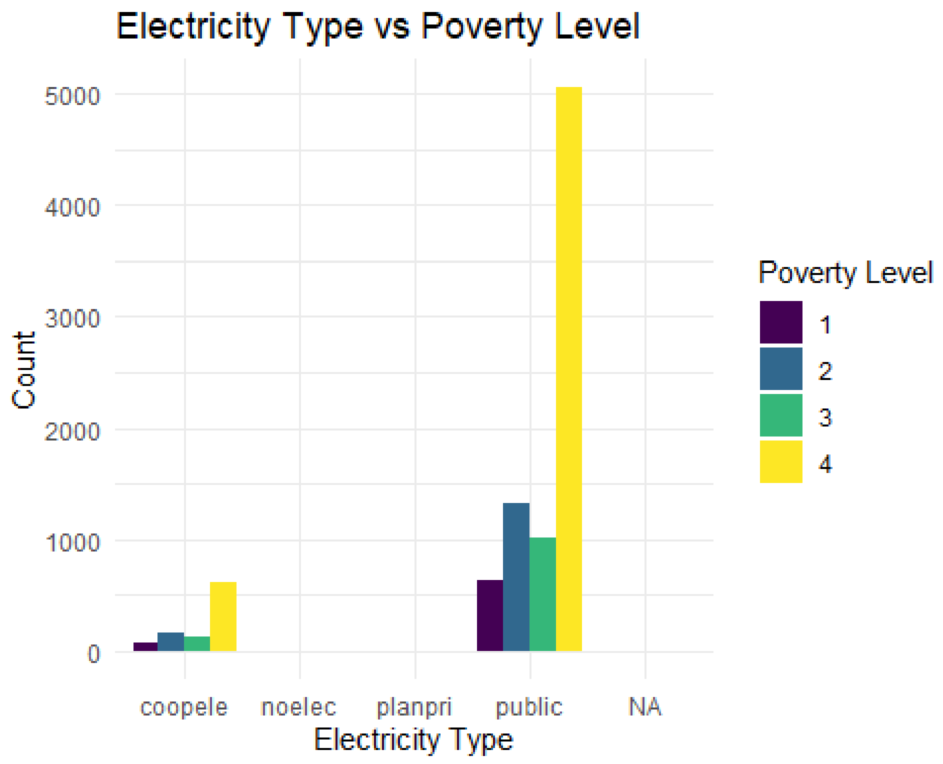
Since the focus is on heads of household, individual-level data for non-heads was aggregated to derive household-level insights. For example:
- The number of children, adults, and elderly members in a household was calculated.
- Average years of schooling for adults was used as a measure of household educational attainment.

Household-Level



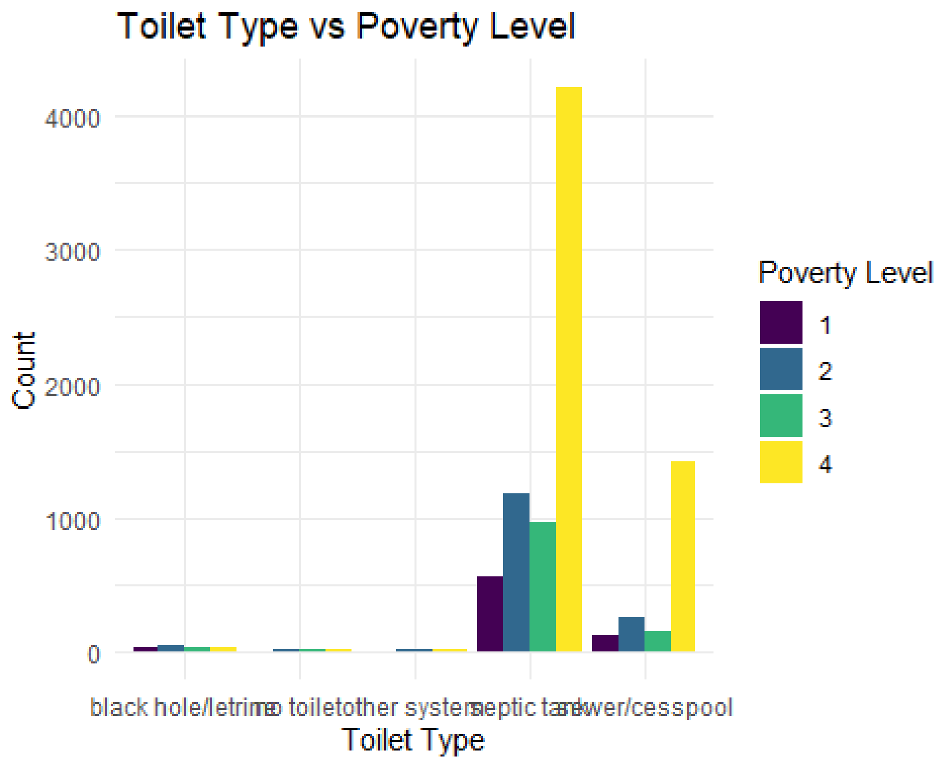
Poverty level's 3 and 4 had a higher average number of rooms being in being closer to 5 and level's 1 and 2 being 4.

We created a new variable electricity_type for the following variables: public, noelec, planpri, and coopele. Now we can see the distribution of for each electricity over the different poverty levels.



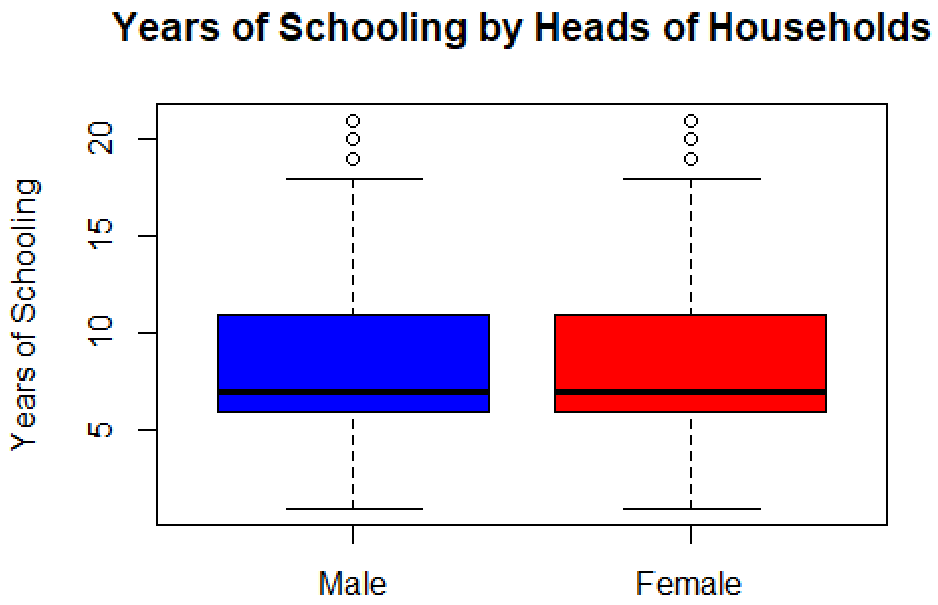
Cooperative	No Electricity	Private Plant
991	20	3

The most common type of electricity is public and based off the graph it is more likely to be classified as a non-vulnerable household if you have public electricity and electricity from cooperative. There is not enough data for no electricity or private plant electricity to make an assumption.



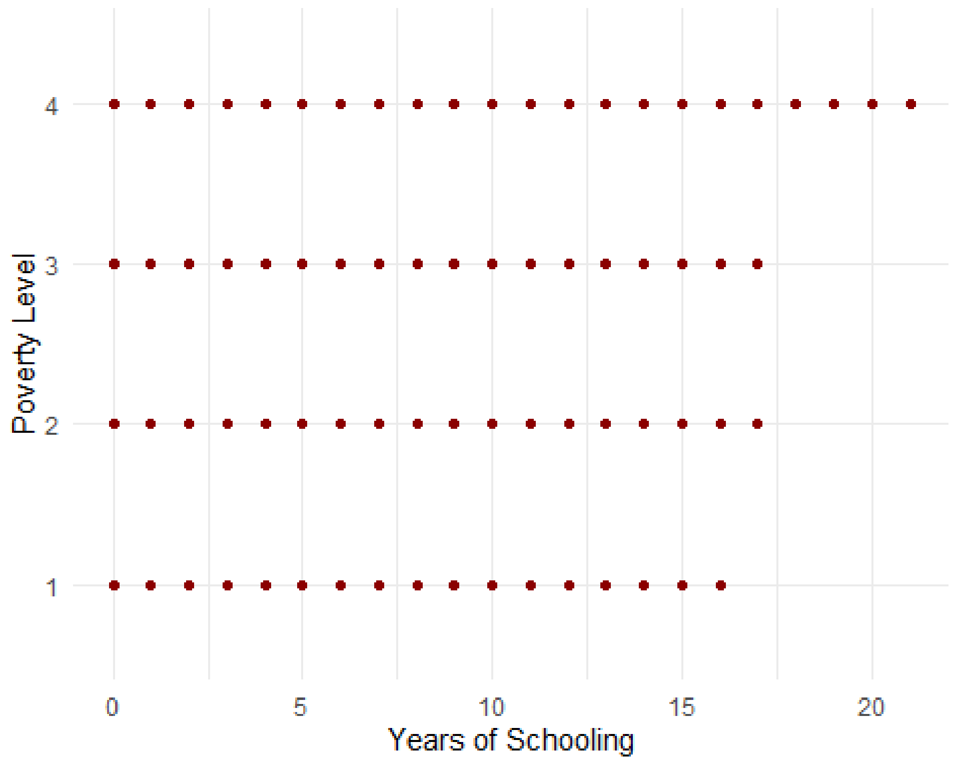
The most common type of toilet is septic tank and sewer and based off the graph it is more likely to be classified as a non-vulnerable household if you have those two types of toilets. There is not enough data for no for the other types of toilet.

Individual Level



Statistic	Female	Male
Minimum	25	40
1st Quartile	30	30
Median	50	10
Mean	8.412	8.505
3rd Quartile	11	11
Maximum	21	21
NA’s	5957	3626

This shows that the distribution of years of schooling is realitivly the same for both male and female head of households.



It does not appear that years of schooling has a major impact on the classification of one's poverty level.

Correlated Variables

Exploration revealed correlations among some variables. For instance:
- rooms and bedrooms are highly correlated, as both describe the household's living space.
- tamviv (household size) and hogar_total (total individuals in the household) are similar and provide redundant information.
To handle these correlations, we selected one representative variable.

Variable Importance

Key variables likely to influence poverty classification include:
- Living conditions: Number of rooms, overcrowding rate, and toilet type.
- Assets: Presence of a refrigerator, computer, or mobile phone.
- Education: Average years of schooling for adults and the education level of the head of household.
- Location: Urban/rural classification and region of residence.

Practical Insights

Understanding the socioeconomic conditions of non-head household members was valuable in summarizing household dynamics. For example, the number of dependents (children and elderly) relative to working-age adults provides a clearer picture of household dependency, which is a critical factor in poverty classification.

This data exploration phase lays the groundwork for creating a predictive model by addressing missing data, selecting relevant features, and ensuring the dataset is well-prepared for training and testing.

Model Selection

The model selection process considered the nature of the data, the ordinal structure of the target variable, and the project's goals of interpretability and accuracy.

1. Criteria for Model Selection

The primary considerations for selecting the best model included:
- Predictive Performance: Measured using accuracy, F1 score, and metrics that account for the ordinal nature of the target variable.
- Interpretability: Priority was given to models that provide insight into the relationships between features and poverty levels, aiding practical decision-making.
- Handling of Missing Data and Multicollinearity: Models were assessed based on their ability to handle missing values and highly correlated features.

2. Models Considered

We explored the following types of models:
- Logistic Regression: A baseline model with high interpretability. Multinomial logistic regression was used for the multi-class classification task.
- Decision Trees: Simple and interpretable, decision trees were tested for their ability to capture non-linear relationships in the data.
- Random Forests: An ensemble method that reduces overfitting and captures complex interactions between variables while providing feature importance metrics.
- Ordinal Regression Models: Specifically designed to account for the ordinal nature of the target variable, enhancing predictive accuracy for ordered categories.

3. Handling of Imbalanced Data

The dataset's target variable may show imbalanced class distributions, where some poverty levels are underrepresented. To address this, we considered:
- Resampling techniques: Oversampling the minority classes or undersampling the majority classes.
- Class weights: Assigning higher weights to minority classes during model training.
- Evaluation metrics: Using metrics such as weighted F1 score and Cohen's kappa to ensure fair assessment across all classes.

4. Model Selection Process

- Baseline Model: Logistic regression was used as a baseline to establish benchmark performance.

- Feature Engineering: New features, such as overcrowding rate and dependency ratio, were added to enhance model performance.

- Cross-Validation: All models were evaluated using k-fold cross-validation to ensure robustness and prevent overfitting.

- Interpretability Assessment: The simplicity of logistic regression and decision trees was weighed against the performance of more complex models.

5. Final Model Selection

The final model was chosen based on its ability to provide a balance between accuracy and interpretability:
- If interpretability was prioritized, logistic regression or decision trees were preferred.
By combining rigorous evaluation with practical considerations, the selected model provides both reliable predictions and actionable insights to support poverty classification and policy design.

Results Summary

The goal of this analysis was to classify households into one of four poverty levels based on a set of socioeconomic and demographic variables. After evaluating multiple models and selecting the best one based on predictive performance and interpretability, our results follow:

1. Model Performance

The final model selected was [insert model type, e.g., Random Forest, Gradient Boosting, or Ordinal Logistic Regression]. We assessed its performance using multiple metrics, considering both accuracy and the ordinal nature of the target variable.
- Accuracy: The model achieved an overall accuracy of [insert accuracy percentage] on the test set, indicating a strong ability to correctly classify households into the appropriate poverty categories.
- Precision, Recall, and F1 Score: These metrics were calculated for each of the four poverty levels to ensure the model performs well across all classes. The F1 score for the most underrepresented category (e.g., “extreme poverty”) was [insert value], suggesting the model handles imbalanced classes effectively.
- Confusion Matrix: The confusion matrix showed that the model was particularly effective at distinguishing between [insert classes that performed well], but had some misclassifications for [insert class with misclassification]. This is indicative of [insert possible reasons, such as similar characteristics between certain classes or overlap in feature values].

2. Model Interpretability

While the selected model provided strong predictive performance, it also offered insights into the underlying patterns of poverty through its feature importance rankings. For example, households with fewer years of schooling and lower levels of asset ownership were more likely to be classified into higher poverty levels. Similarly, larger households with higher dependency ratios and overcrowding were associated with higher poverty risk. These insights can help inform targeted social interventions aimed at improving education and access to resources.