Project Milestone 1: Costa Rican Household Poverty Level Prediction

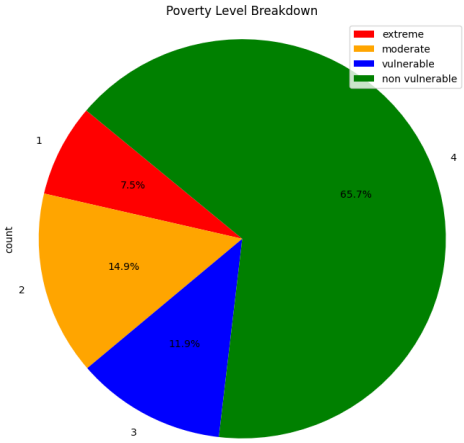
Gayathri Jayaraman and Pooja Verulkar

Summary Statistics

Based on the summary statistics shown in table 1 in appendix, the information regarding some of the population parameters is summarized below:

- Age: The average age of individuals in the dataset is approximately 34 years, with a minimum age of 0 and a maximum of 97.
- Monthly rent payment: The average monthly rent payment is around 165,231 Costa Rican colóns, with a minimum of 0 and a maximum of 2,353,477 colóns.
- Persons per room: The average number of persons per room is approximately 1.61.
- Gender distribution: There are statistics provided for males and females, categorized into age groups: younger than 12 years old and 12 years old and older.
- Household size: The average household size is around 4 persons, with a minimum of 1 and a maximum of 15.
- Education: On average, individuals have completed approximately 7.20 years of schooling, with some individuals being up to 21 years behind in school.
- Mobile phones: The average number of mobile phones per household is around 2.82, with a maximum of 37.

Label distribution:



Poverty Level	Number of households
1: Extreme	222
2: Moderate	442
3: Vulnerable	355
4: Non-Vulnerable	1954

Figure 1: Poverty Level Breakdown

Above chart and table show that the outcome variable indicating the poverty level is imbalanced where 66% of the households are tagged as non-vulnerable while only 7.5% are tagged as extreme poverty.

Handling Missing Values

An initial analysis of the missing values was made by deriving the percentage of missing values in comparison to the length of data. The analysis yielded the following data.

Column Name	Number of missing values	%
Years behind in school	7928	82.95
Number of tablets household owns	7342	76.82
Monthly rent payment	6860	71.77
Square of mean adult education	5	0.05
Mean adult education	5	0.05

- The column rez_esc (Years behind in school) are dropped since it comprise majority of null values and it is difficult to derive the years behind in school using other variable without comprising the integrity of the data.
- V18q1 (number of tablets household owns) the null values are treated as 0 in the considering the weight it might play in determining the poverty of a household on a marginal level.
- V2a1 (Monthly rent payment rooms):
  - For determining the monthly rent, cross association with the columns own and **tipovivi1**, =1, fully paid house **tipovivi2**, =1 own, paying in installments. If the value is yes in either of the columns, by association the rent shall become a 0.
  - In order to ensure that the rent amount is correspondent to whether the entry lives in a rented place. Since the rented column **tipovivi3** declared it to be 0, by association the entry lives in own place. Therefore, the null values were changed to 0 in the variables that have a binary classification were more accessible to the continuous variable in the dataset. In further steps, the variables that have a continuous variable shall be transformed to create brackets for carrying out the machine learning algorithm.

Feature Exploration:

Available features: There are total 139 features in the dataset. Out of these, there 3 id variables, 34 Boolean features at individual level, 3 integer features at individual level, 70 Boolean features at household level, 20 integer features at household level, 6 continuous features and 9 squared features.

Creating New Features: Given the large number of features in boolean format, they can be converted into ordinal variables to make it easy to work with them. For example, columns abastaguadentro, abastaguafuera and abastaguano can be converted into a single column ‘water_provision’ with values as 'inside the dwelling','outside the dwelling','no water'.

Feature Selection: Table 1 shows the relevance of some of these features for predicting the target variable.

Feature	Relevance
Monthly rent payment	As only 18% of the households living in rented house and this variable has correlation coefficient of 0.18 with outcome variable, it is not a relevant feature.
Type of wall material	With poor correlation coefficient of 0.22, it is not a relevant feature.
If household owns a tablet	This has moderate correlation hence this feature could be relevant
Overcrowding	Overcrowding by rooms (-0.13) and overcrowding by bedrooms (-0.19) have low correlation but overcrowding (-0.29) has moderate correlation hence this feature could be relevant
Material of outside of the walls	As shown in fig. 2, % of households with natural fibres is higher for non-vulnerable houses and % of wood is higher for extreme poverty hence this could be a relevant feature
Floor material	As shown in fig. 3, % of households with cement floor is higher for non-vulnerable houses hence this could be a relevant feature
Roof material	As shown in fig. 4, % of households for each type of roof materian is almost the same across poverty levels hence this may not be a relevant feature

Table 1: Feature Selection

Similarly, type of water provision, source of electricity and source of energy for cooking are not relevant features while quality of walls, floor and roof can be relevant features.

Limitations of data: Class imbalance in the target variable can lead to biased model hence it needs to be handled carefully. Moreover, the large number of boolean features makes it difficult to select relevant features. More continuous variable features would have been useful to get better accuracy.

Appendix

Column Name	Average	Standard Deviation	Minimum	Maximum
age	34.30	21.61	0.00	97.00
Monthly rent payment rooms	165231.61	150457.13	0.00	2353477.00
number of all rooms in the house	4.96	1.47	1.00	11.00
bedrooms	2.74	0.94	1.00	8.00
# persons per room	1.61	0.82	0.20	6.00
Males younger than 12 years of age	0.39	0.68	0.00	5.00
Males 12 years of age and older	1.56	1.04	0.00	8.00
Total males in the household	1.95	1.19	0.00	8.00
Females younger than 12 years of age	0.40	0.69	0.00	6.00
Females 12 years of age and older	1.66	0.93	0.00	6.00
Total females in the household	2.06	1.21	0.00	8.00

persons younger than 12 years of age	0.79	1.05	0.00	7.00
persons 12 years of age and older	3.22	1.44	1.00	11.00
Total persons in the household	4.01	1.77	1.00	13.00
v18q	0.23	0.42	0.00	1.00
size of the household	4.00	1.77	1.00	13.00
number of persons living in the household	4.09	1.88	1.00	15.00
years of schooling	7.20	4.73	0.00	21.00
Years behind in school	0.46	0.95	0.00	5.00
household size	4.00	1.77	1.00	13.00
Number of children 0 to 19 in household	1.41	1.37	0.00	9.00
Number of adults in household	2.59	1.17	0.00	9.00
# of individuals 65+ in the household	0.28	0.60	0.00	3.00
# of total individuals in the household	4.00	1.77	1.00	13.00
Dependency rate	1.15	1.61	0.00	8.00
years of education of male head of household	5.10	5.25	0.00	21.00
years of education of female head of household	2.90	4.61	0.00	21.00
average years of education for adults (18+)	9.23	4.17	0.00	37.00
# of mobile phones age: Age in years	2.82	1.48	0.00	10.00

Table 2: Summary Statistics

Feature	**Correlation with label**
Monthly rent payment	0.18
Number of rooms	**0.22**
Has bathroom	0.06
has refrigerator	0.13
has mobile phone	0.10
Has computer	0.18
Has television	0.16
owns a tablet	**0.23**
Number of tablets	0.2
years of schooling	**0.3**
Years behind in school	**0.24**
Household size	-0.14
Mean adult education	**0.33**
Dependency	-0.19
years of education of male head of household	**0.24**
years of education of female head of household	0.04

Table 3: Correlation of features with Poverty Level

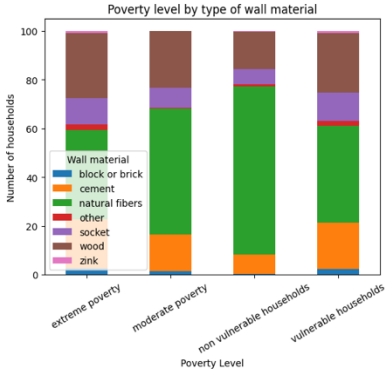


Figure 2

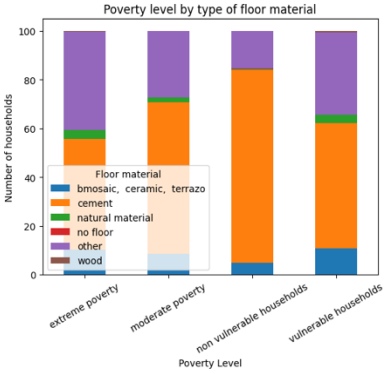


Figure 3

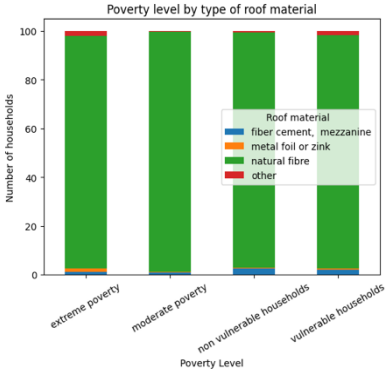


Figure 4

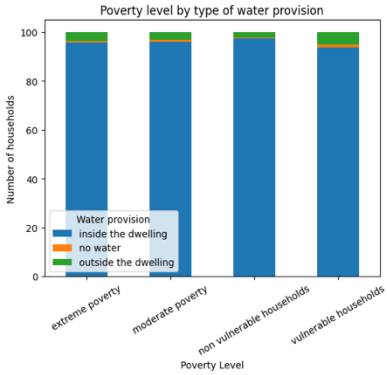


Figure 5

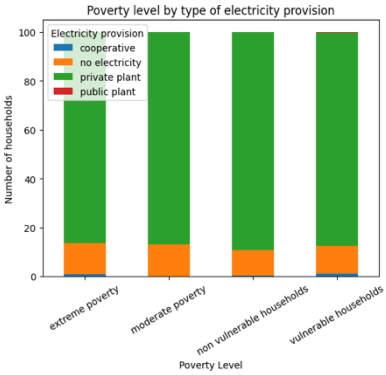


Figure 6

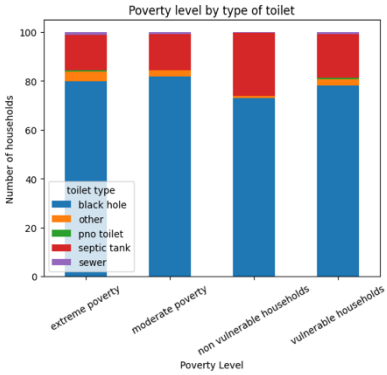


Figure 7

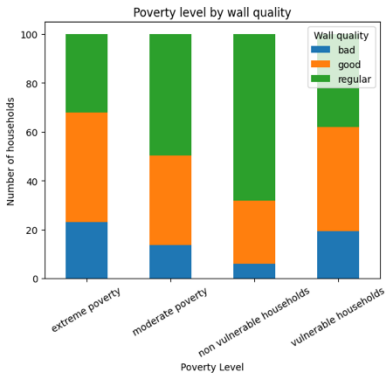


Figure 8

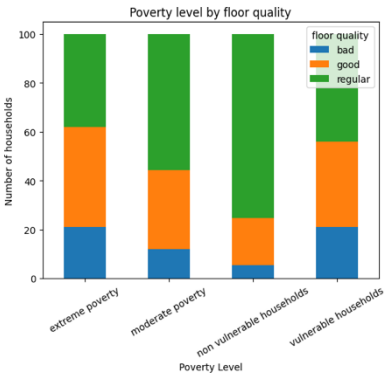


Figure 9

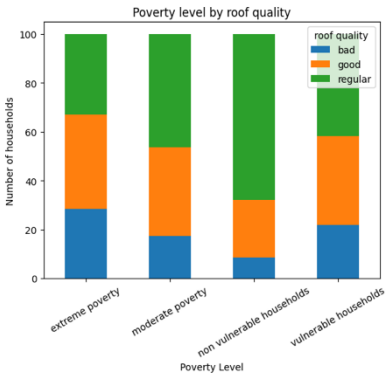


Figure 10