* **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. 1, % 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. 2, % of households with cement floor is higher for non-vulnerable houses hence this could be a relevant feature |
| Roof material | As shown in fig. 3, % of households for each type of roof materian is almost the same across poverty levels hence this may not be a relevant feature |

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.

|  |  |
| --- | --- |
| **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 |

A graph of poverty level

Description automatically generated A graph of a poor level

Description automatically generated with medium confidence A graph of poverty level

Description automatically generated

Figure 1 Figure 2 Figure 3

A graph of water poverty

Description automatically generated A graph of poverty level

Description automatically generated A graph of poverty level

Description automatically generated

Figure 1 Figure 2 Figure 3

A graph of a poor quality

Description automatically generated A graph of poor quality

Description automatically generated A graph of a poor level

Description automatically generated with medium confidence

Figure 1 Figure 2 Figure 3

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.