Executive Summary Report
Brandon Arias

Objective:

Use classification to identify which households have the highest need for social welfare assistance.

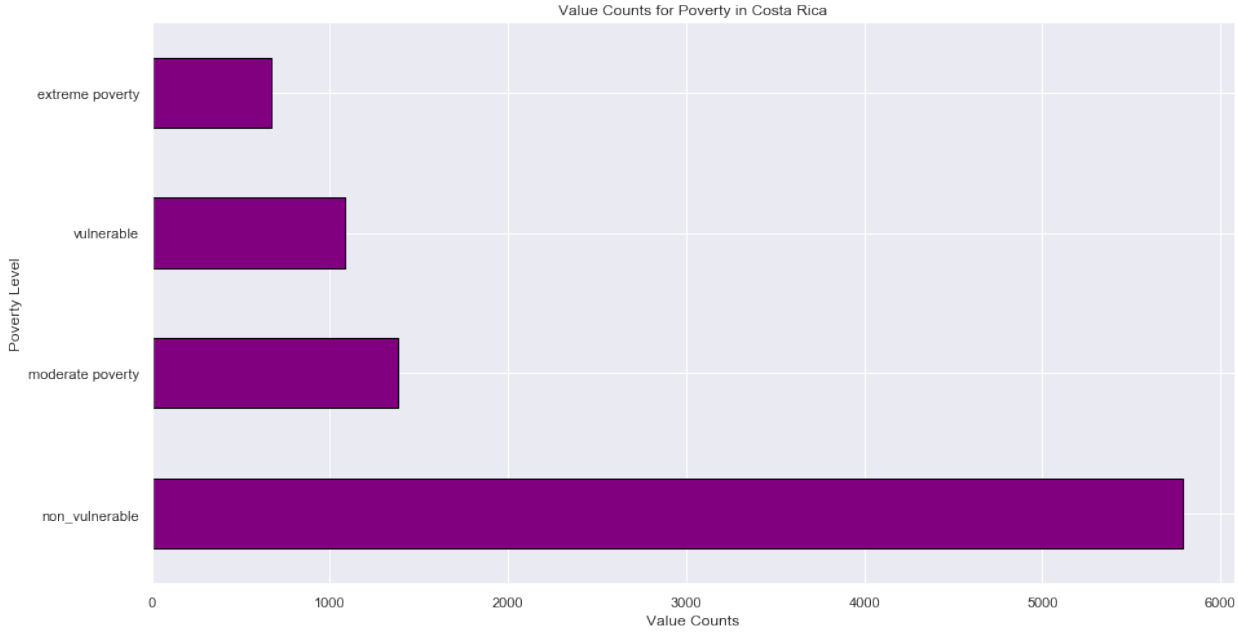
Motivation:

Social programs have a hard time directing enough aid to people that need it. It's especially tricky when a program focuses on the poorest segment of the population. In Latin America, one popular method uses an algorithm to verify income qualification, known as the Proxy Means Test (or PMT). This allows agencies to use a model that considers a family's observable household attributes like the material of their walls and ceiling, or the assets found in the home 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.

Data Analysis:

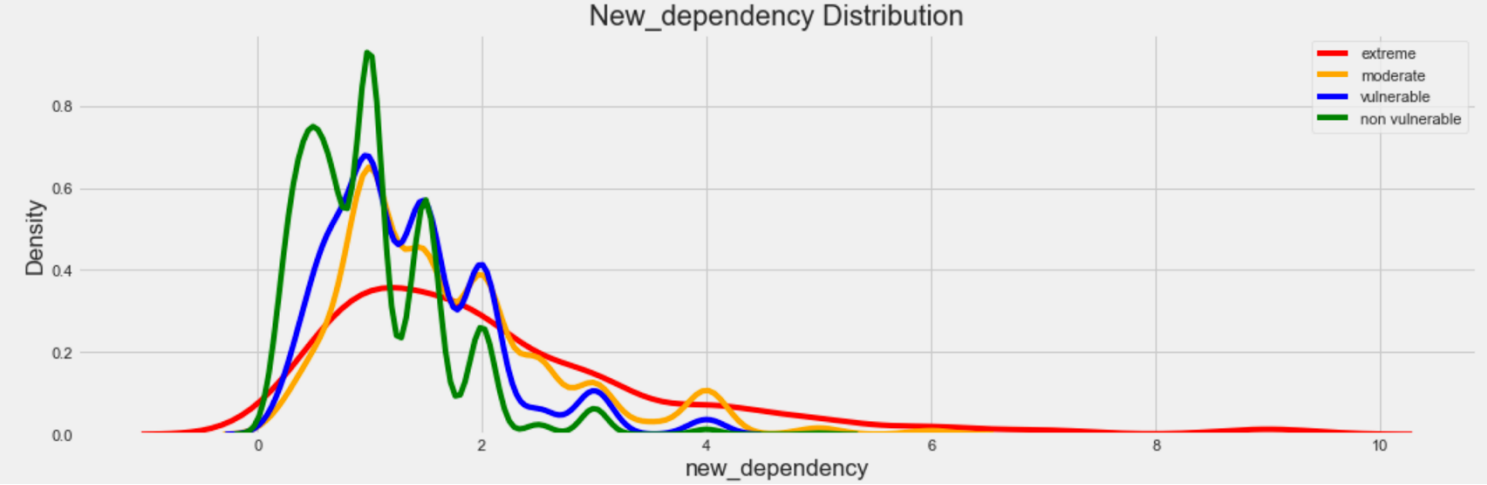
From the data provided, we can note of that most of the observations fall in the non vulnerable sections. Upon exploring the percentage, we can note the following distribution:

```
Non-vulnerable cases make up 64.84 % of all cases.
Moderate poverty cases make up 15.51 % of all cases.
Vulnerable poverty cases make up 12.17 % of all cases.
Extreme poverty cases make up 7.48 % of all cases.
```



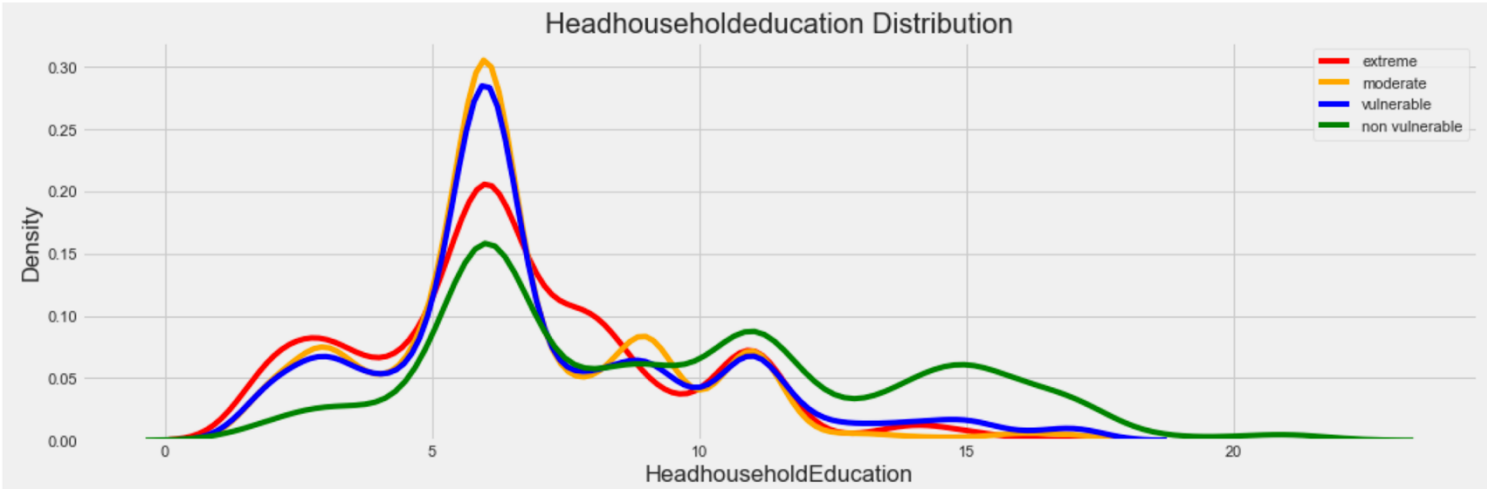
From these distribution, our models need to be performing at an accuracy better than 64.84% to be impactful since a naïve solution would be labeling all observations as non vulnerable.

Executive Summary Report
Brandon Arias



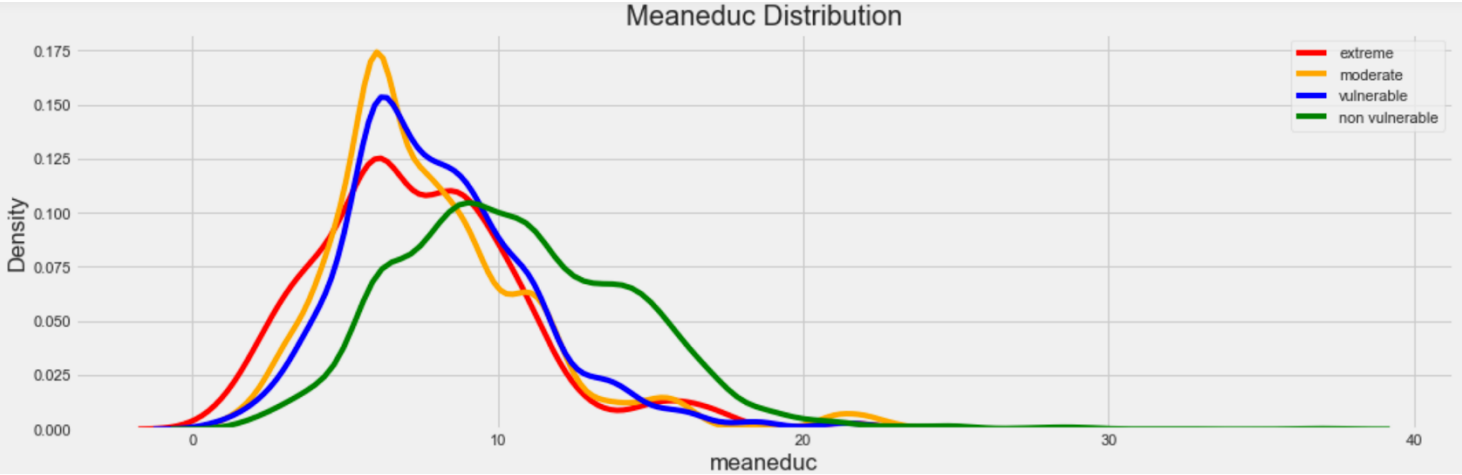
When exploring the dependency ratios, defined as ratio of children and elderly over adults, we can note that as dependency ratios get higher, families will fall outside of a non-vulnerable. In fact, those with higher dependencies are more likely to fall in a more serve category.

If this data is representative of the population, we can observe that it is not so common for family distributions to be anywhere bigger than a ratio of four. In fact, we can capture most of the distribution from zero to two.



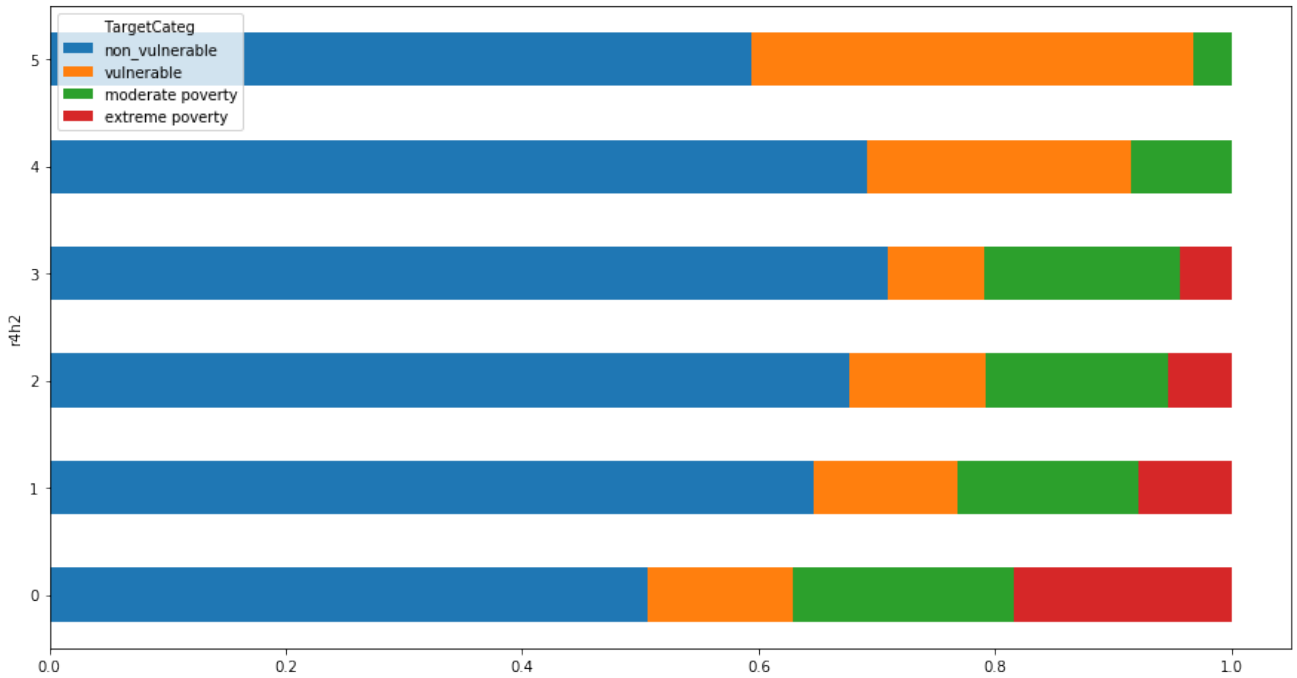
For this visualization, I considered the education of the head of the household. By looking at the distribution of the head of the household and breaking them down by level of vulnerability, we note that grade six becomes where a big chunk of the distribution lies. At this peak, more people tend to fall outside of the non-vulnerable categorization. If this sample is representative of the population, this could be telling the population's common education levels and where they fall in vulnerability. It is additionally worth nothing at grade level 9, non-vulnerable families take the lead while there are less families with that fall outside of that category.

Executive Summary Report
Brandon Arias



When looking at the mean education of families, families that have higher education tend fall in less severe or non-vulnerable areas. This can potentially be explained by that with more education, families can be employed in more areas where they get paid more. As such, they manage to live a lifestyle where they are more likely to be safe.

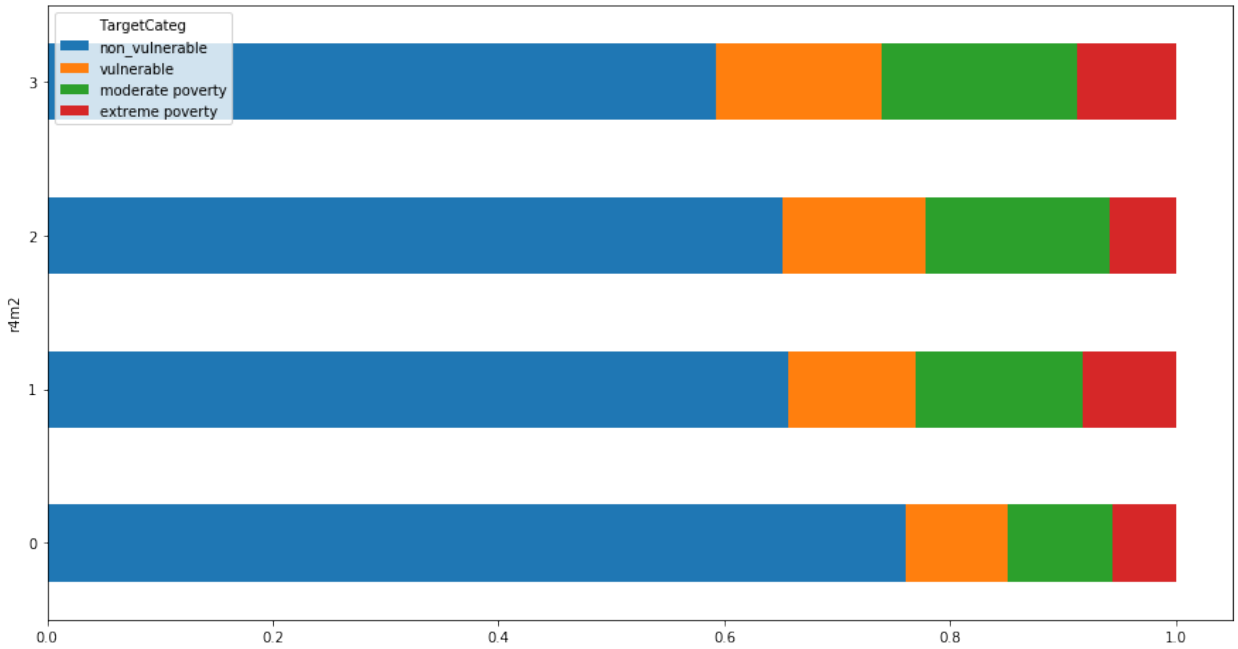
Number of Males > 12 Years Old



From this adjusted we can see a clear trend that the more males that are older than 12 years old, decreases the chance of up until 3, then we start seeing the chances of falling outside a non-vulnerable group increasse. In fact, we can note that the severity of poverty also decreases as the number of males that older than 12 years old increases.

Executive Summary Report
Brandon Arias

Number of Females > 12 Years Old

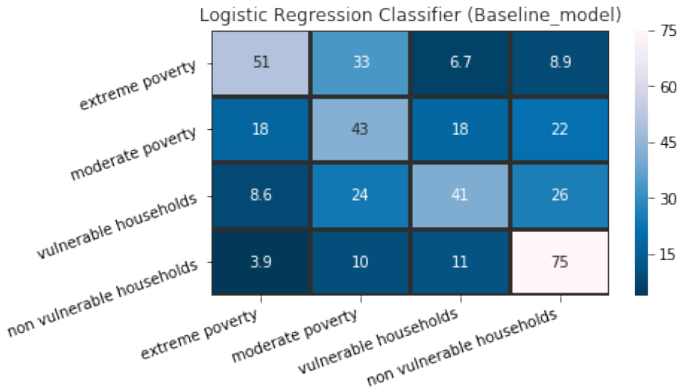


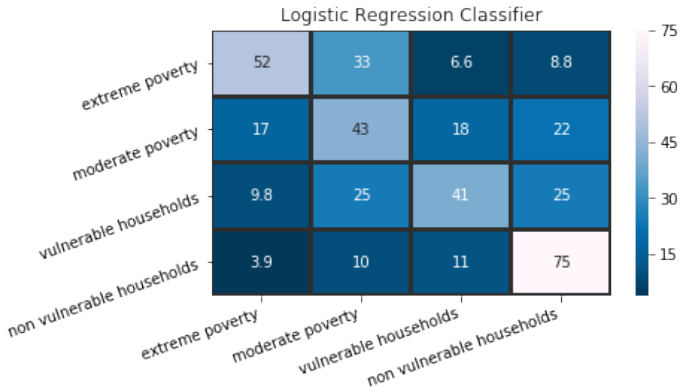
From this figure, we can see a clear trend that the more females older than 12 years old, the higher the rate of falling outside of a non-vulnerable group. In fact, we can note that the severity of poverty also increases as the number of female that older than 12 years old increases. This is the same idea for females younger than 12 years old except less severe.

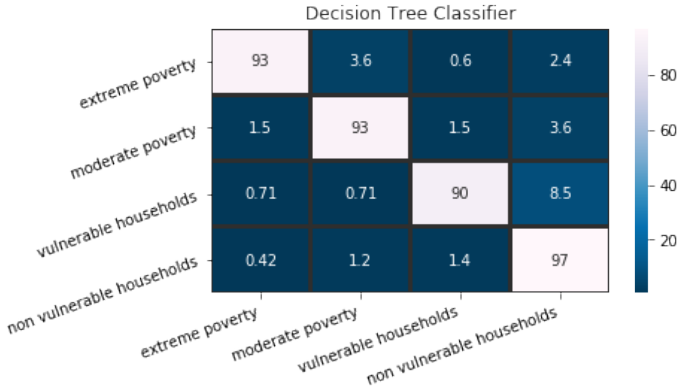
In explaining both figures, we can potentially note the gender explaining for these figures: Men are expected to be working and providing income for the homes meanwhile women are expected to be caretakers in the home

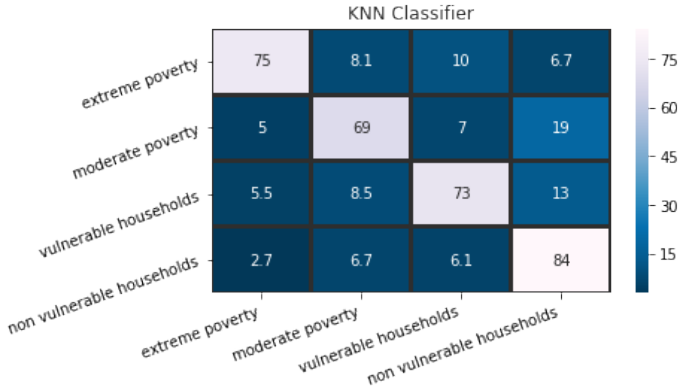
Executive Summary Report
Brandon Arias

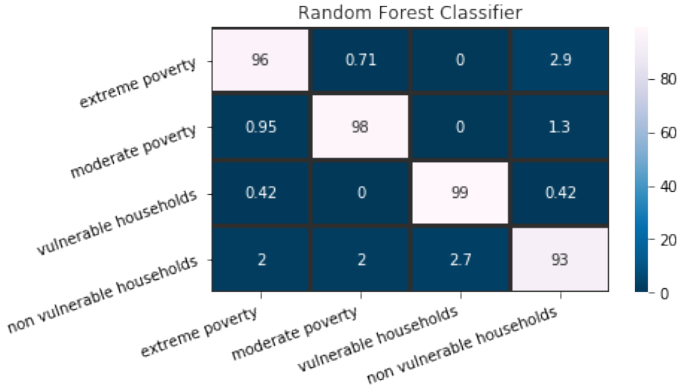
Modeling:

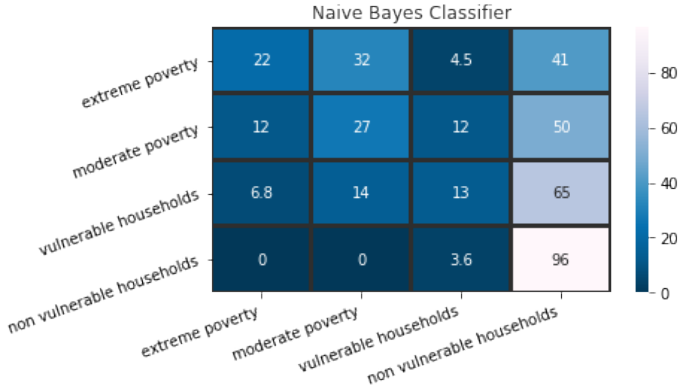


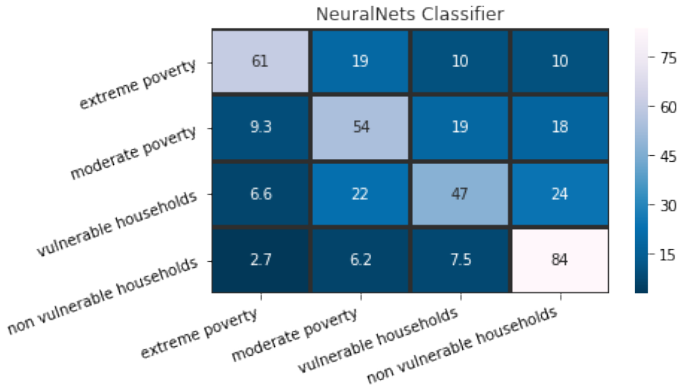








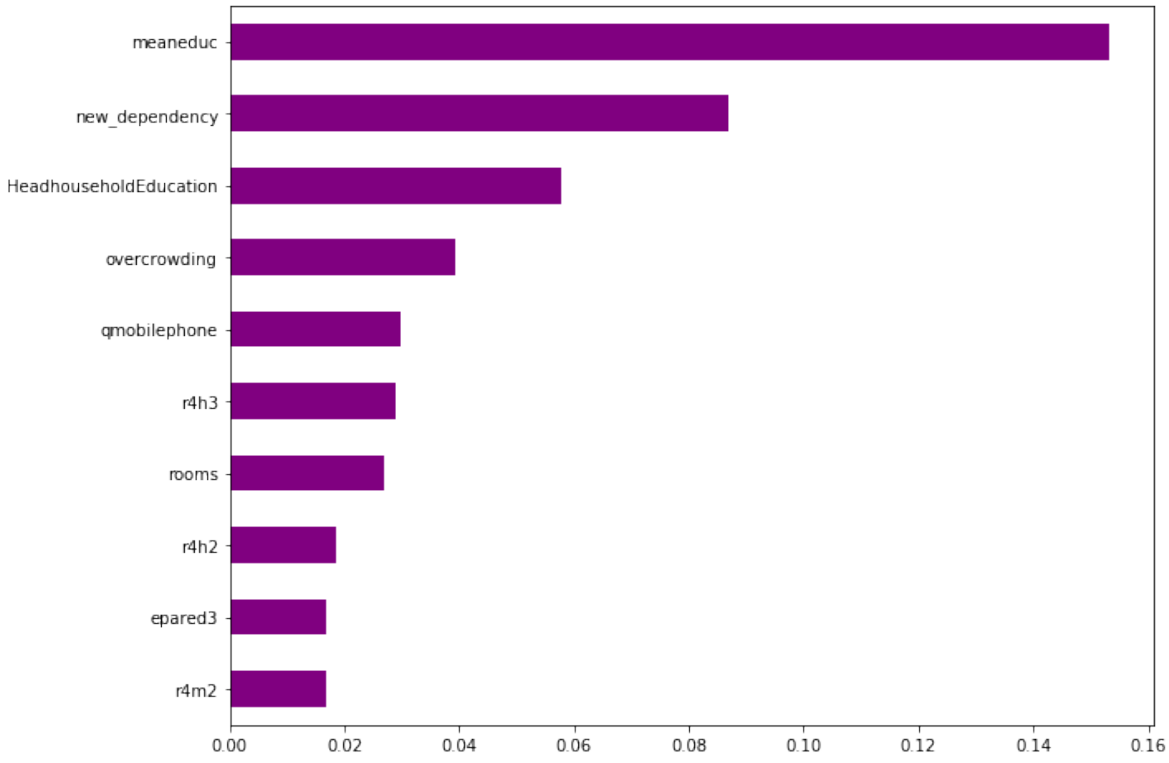




Executive Summary Report
Brandon Arias

	Model	Accuracy_score	Recall_score	Precision	f1_score
0	Logistic Regression(Baseline_model)	0.692204	0.692204	0.692204	0.642325
1	Logistic Regression	0.693100	0.693100	0.693100	0.643906
2	Decision Tree	0.952957	0.952957	0.952957	0.952931
3	KNN Classifier	0.806900	0.806900	0.806900	0.798023
4	Random Forest Classifier	0.947133	0.947133	0.947133	0.945981
5	Naive Bayes	0.160394	0.160394	0.160394	0.096167
6	Neural Networks	0.749104	0.749104	0.749104	0.734512

From the various models, we can note that Random Forest and Decision Trees perform the best. For evaluating the best model, I chose to look at an f1_score since it would be the balance between accuracy and recall. In addition, I would recommend using Random Forest as the best model since it is in theory more robust than Decision Tree in that Random Forest is an ensemble of decision trees.



When looking at the feature variables that are important in making the correct classifications, I found that mean education, dependency, education level of the head of the household, and overcrowding are the best ones.

Executive Summary Report
Brandon Arias

Conclusions:

Next Steps:

Limitations: