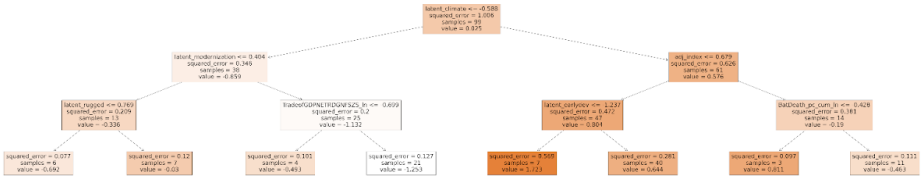
Inequality Project Decision Tree Model Essay

Fixed Linear Regression Model from OLS:

gini_disp_ext = 39.5571 + 2.2910*latent_climate - 1.6418*latent_earlydev + 0.5685*sd_index_ext + 1.3483*latent_rugged

Decision Tree Model:



In the process of selecting the right data for the linear regression model, I used Excel to find all the countries that were missing data for the important independent variables, and deleted them. To further filter out which countries to delete, I looked at a summary of the data for all the important independent variables between all the filled data and all the data in general and found that the means and standard deviation only changed significantly for the latent_climate and latent_earlydev variables. All the countries that didn't have values for either of these two variables were not important. For the countries that had missing Gini Coefficients(Marshall Islands, North Korea, Liechtenstein, and Morocco), I looked up their Gini coefficients and plugged those values in.

In class, it was discussed that latent_climate and latent_earlydev were the most important factors in determining a country's Gini coefficient. These variables set a precedent for a country's social, economic, and political development. I fixed my linear regression model to only include variables that could not easily be changed. Using this logic, all modern variables were irrelevant and I only used variables that explain characteristics inherent to a country. That explains why only latent_climate, latent_earlydev, sd_index_ext, and latent_rugged were included in my linear regression model. With an r-squared of .551, the equation is a solid estimate for gini coefficient.

Selecting data for the decision tree was much easier than linear regression, because the DecisionTreeRegressor class from scikit filters out which classes to do for you. I decided to remove all the countries from the data set that was missing data because it would have been hard to impute data for all 32 independent variables for all important countries. This resulted in only 111 data points, which was a big limitation of this decision tree.

The decision tree was created with a max depth of 3, which was determined after testing higher max_depths. Other than that, I did not do anything special with In theory, because we know that latent_climate and latent_earlydev are the most important factors, they should take up the nodes at the top. However, from the decision tree, it looks like adj_index is a better predictor variable because it is able to split the subset of data into two subsets with higher purity. Another potential issue of this decision tree is that the split from latent_rugged is not very helpful because half of the observations are in each subset, so the tree could keep splitting from there. However, if I added another layer to the tree, it may become too complex and overfit. It is also interesting that battle deaths was a node because it seems like number of battle deaths is irrelevant to determining Gini coefficient. Although my r-squared was a high .82, I don't think this model is good because once the random state parameter changes, the r-squared drops significantly.

As discussed above, the decision tree has a lot of limitations that the linear regression model does not have. In addition to that, I would trust the linear regression model more in determining Gini coefficient because the variables included in the equation make sense as to why it affects Gini coefficient. However, some variables(like Battle Deaths) in the decision tree seem to be irrelevant but are still included in the equation.