**CODE EXPLANATION:**

The first step I took was to redo the polynomial regression to receive a parsimonious regression that was more in line with the grading criteria. Initially, I decided to use k-fold on the data since I felt like a manual train test split wasn’t the best idea considering how small the dataset is. Next, I decided to use a lasso regression technique because it will help me with feature selection, which is something I needed to change from my previous code. However, before jumping into the code, I knew that my final answer should ideally contain IV’s that predate the more modern IV’s in this dataset considering the principles of causality in that finding causal relationship that happen before others should ideally reduce noise. With this in mind, I was mainly looking for my lasso regression to result in a small subset of early variables like latent\_climate.

The first step I took in building this code was preprocessing of the data and essentially, I dropped all rows of countries with little to no data. Then, out of the rows that remained, I used KNN imputation to fill in missing values. KNN works by finding a samples of the data’s ‘closest neighbors’ and then imputing values based off of the averages of the neighbors. Ideally, this method of imputation is better than something like mean imputation because it should preserve the underlying patterns within the data. Although, this method is more sensitive to outliers in the data. Also I drop NA, rather than impute, on a few latent IV’s and gini (gini because that is what I am predicting) because I did know the possible underlying patterns from the calculations to create the latent vars.

Once I have a preprocessed dataset, I set up the lasso regression for the data. Within this set up I implemented a way to iterate through different values of alpha penalty and the max\_iter parameter until a best model (highest r square) was found. Within this iteration, kfold cross validation was run to create different train tests splits per iteration of each parameter. Following this regression, I return the variables with the best performance on predicting Y in the lasso and my model was able to identify ['adj\_index', 'latent\_climate', 'latent\_rugged\_ext', 'sd\_index\_ext'] as the top 4 variables with a corresponding r square of 0.57.

For the decision tree, I kept the same methods of imputing and preprocessing the data as I did in the lasso regression and then I implemented similar methods of finding the best parameters for my tree. Essentially, I first ran a kfold to implement cross validation and then I ran a loop of 3 different parameters (max\_depth, min\_sample\_split, min\_sample\_leaf, ccp\_alpha) with altering values. I initially tried LOO since I figured it would work better on a smaller dataset, but it seemed to negatively impact the results. Furthermore, the point of the loop was to find which combination of the set of parameters chosen would give the best r2 value. Also, cpp\_alpha is a pruning method I used in the loop to give a method of preventing overfitting.

**RESULTS:**

From the lasso regression, my coefficients for each variable are as follows:

latent\_rugged\_ext: 1.8935300599421956

latent\_climate: 1.9310430483654824

sd\_index\_ext: 1.7796350482923904

adj\_index: -2.5081838438870725

(typical R squared of around .6)

These variables were expected and it is clear to see that variables related to the initial formation of a particular country and its corresponding geographical area would predate other IV’s in the dataset; thus, I am making the claim that a gini score of a country can be predicted based off of factors that don’t necessarily relate to modern human interference but more so towards the inherent landscape and quality of a country’s disposition as well as the early settlement and development of a country. These results stem from the initial hypothesis that while more modern events like socialism index and battle deaths may be correlated with gini score, events that far predate these modern measures offer a stronger causal relationship with gini score.

My decision tree ended up averaging an r2 of around .5 and an MSE of 34. These results are expected because I did not anticipate a decision tree to work very well on this high dimensional and small dataset. When comparing this with my results of the lasso regression, I would be much more comfortable placing faith in the regression results for a number of reasons. Namely, my decision tree method results have high variance and I rarely get the best combination of parameters to stay the same in between runs. This is likely due to the nature of the dataset how small changes in folds can cause instability in the results. I believe lasso Is better at regularization because it more easily forces unimportant variable coefficients to be zero which should ideally work well on high dimensions like this dataset. In comparison, the inherent technique of using a decision tree makes them quite sensitive to the small amount of data in each fold and many trees could possibly contain poor predictors and suffer from this variability. Since the decision tree is a greedy algorithm, it simply makes a decision at each split while lasso iterates multiple times before it converges on a result.

Implicitly, decision trees choose which variable to split first by how well it will predict and prune those that are not as good at predicting; this aspect of decision trees makes them very reliant on the training data. However, a lasso regression uses stricter regularization for each variable and could offer a more robust model when considering noise in the data.

More specifically, my lasso regression outputs ['adj\_index', 'latent\_climate', 'latent\_rugged\_ext', 'sd\_index\_ext'] as the most important predictors while my decision tree tends to have a different list. Although it usually results in latent\_climate as the first split, the following splits tend to be a mix of adj\_index, euro2007, latent\_rugged\_ext, and latent\_earlydev. Although most of these variables are similar, I presume that variability in the training data allows certain variables like euro2007 to be interpreted as decent predictors by the decision tree.

A diagram of a tree

Description automatically generated