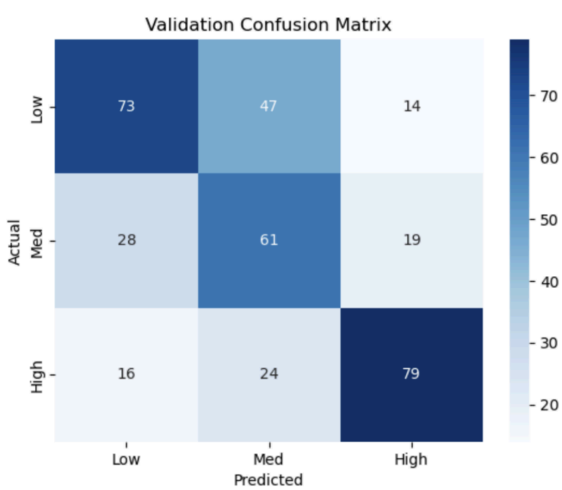
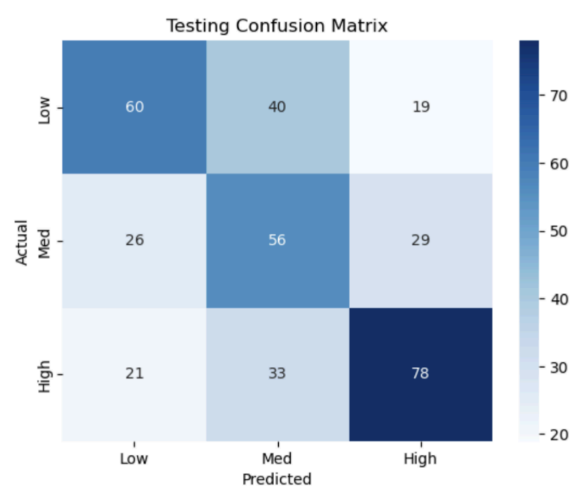
The goal of this model was to find which features are the strongest indicators for the GDP growth rate per capita and accurately predict whether a country will have a low, medium or high growth rate. To ensure our model could gain a broad understanding of economic indicators, we gathered data on macroeconomic indicators, demographic measures, and political stability indicators. The strongest model was a random forest classification with an accuracy of approximately 59% on validation data and the strongest indicators were Population Growth, Population, and Government Effectiveness.

Our data was sourced from the World Bank (WGI), Penn World Table, and Census.gov. Our initial dataset had 65 columns and nearly 9000 rows. We cleaned our data, by removing columns that contained a high percentage of nan values or were broken. To ensure that our model actually selected the most important features, columns with a correlation over 85% were removed. The target variable was split into three categories (low, medium, and high) by tertiles, to allow for more intuitive application. Lastly, the data was split into training(80%), validation(10%), and testing subsets(10%).

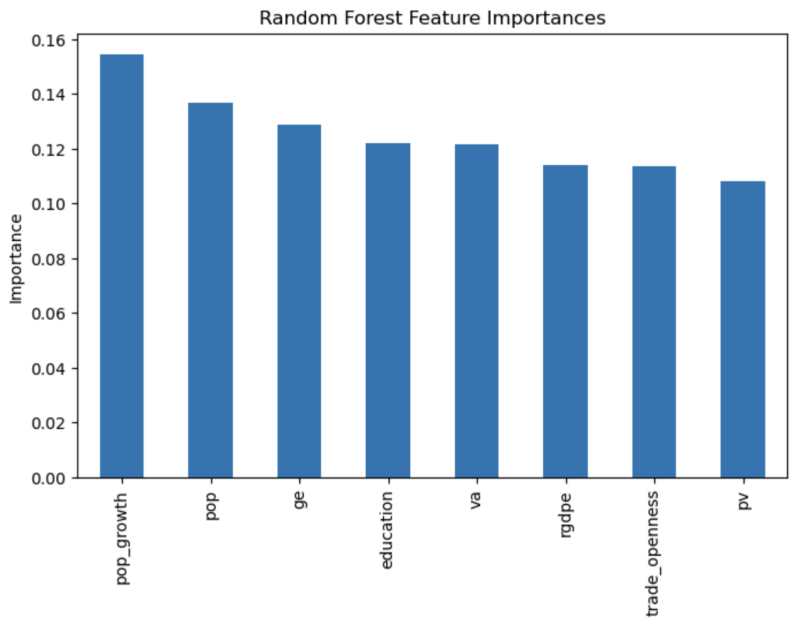
In our experiments, we chose to use the random forest algorithm because the results are both robust and easily interpretable. We also tested three other algorithms, including Random Forest, KNeighbors Regression, Lasso Regression and Linear Regression. We selected Random Forest, because it had the highest accuracy on unseen data and it was able to handle non-linear data, which is often seen in economic scenarios. We tuned our hyperparameters using grid search and the best parameters were: 'max_depth': 15, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 300.

Our model correctly predicted the growth category of countries 84%, 59% and 53% of the time on training, validation and test data respectively, which is significantly better than the expected accuracy of ⅓ if guessing. Overfitting is present, as we see reduced accuracy on unseen data than training data. However, given the variability of economic data, this is still a respectable prediction accuracy, and we can make meaningful connections to larger economic theories.





After running the model, we found that Population Growth is the most important determinant of growth in economies. This raised concerns about our model being trained to heavily predict the “per capita” aspect of our target instead of gdp growth, but when removing population and population growth, the accuracy on test data was only reduced by 2%, and our model overfit the training data more than before. Another revealing finding was the importance of Government Effectiveness (GE). This surprised us because traditional macroeconomic theoretical models, such as the Solow Growth Model, do not incorporate government effectiveness into its structure. We hypothesize that its importance in our model highlights a nuance in the actual dynamics at play.



To refine this model, we would look for more data sources and potentially experiment with ways to manipulate the data to serve better results. However, we also acknowledge that a country's growth dynamics are subject to a wide range of factors, and there will be no way to completely eliminate noise from a model.

By developing this model, we were able to predict the economic growth of countries with relative accuracy, and find important indicators to understand how economies continue to grow. Moving forward, along with finding new data sources, we would continue to experiment with ways to improve the algorithm and its effectiveness. We would look towards the use of neural networks specifically, in hopes that a more sophisticated model would be more effective at capturing the complexities of a country’s growth.