Ashley and Kerigan

Team Project Proposal - Draft

The data we plan to use for this project is the Human Development Report from 2015, taken from Kaggle (and found here: https://www.kaggle.com/undp/human-development). The data is partly an overview of different dimensions of human development, including life expectancy, levels of education, gross national income, and more. There are also other values of the data that are calculated summaries for each country's achievements, including the gender development index, coefficient of inequality, and the overall ranking of human development compared to the other countries. This data is important as a representation of which countries are more developed than others and how. It (in addition to our planned work) may help the UN make decisions on things such as funding.

Through exploration of the data, we hope to determine a function that will allow us to determine the correlation between the Human Development Index (HDI) and the Multivariate Poverty Index (MPI). This is our question. By using a regression model we will be able to produce a function that inputs the HDI and outputs a peredicted MPI.

To begin building this function we need to be able to compare the UN's MPI data to the UN's HDI data. The HDI data is spread out throughout several CSV files so it will be necessary for us to begin by concatenating the HDI and Gender Development Index (GDI) information. We can then compare features that affect the GDI and HDI and compare it to the data in the MPI CSV. Since the HDI is directly determined by the same data as the GDI, it makes sense to keep them in the same file. However, the MPI only focuses on the Global South so we will need to edit the HDI data by creating a dictionary of countries in the MPI CSV file.Then we run a simple if-not for loop that removes all of the countries from the HDI files that are not in the MPI file . Following this, we will need to reformat the CSV files by importing ".." as a NaN function. Since we are missing some information on countries like Afghanistan. Namely, Afghanistan does not give information from the"Population Below $1.25 per Day" variable. So, to fix this we can create smaller regression models in order to fill in the necessary information. We will need to do this for the following features: Population Below $1.25 per Day, Population Below The national poverty line, Life Expectancy at Birth (Female), Life Expectancy at Birth (Male),Estimated Gross National Income per Capita (Female), and Estimated Gross National Income per Capita (Female),Gender Development Index (GDI), Human Development Index (Female), and Human Development Index (Male) at least. Furthermore, we will not be able to keep data from Dominica, Andorra, and some other countries as their is not enough data to compare them to the other counties reliably.

In conclusion, by finding the relationship, using a regression model, between the HDI and the MPI, there may be more ability to make decisions about where aid and funding are necessary. Our next steps primarily include feature engineering with countries with higher or lower amounts of data. Then, we will begin testing regression models to see where data can be filled in best.