CIS 545 Final Project: Worldwide Health Trends

Fall 2020 | By: Ke (Coco) Zhao & Yathushan (Yathu) Nadanapathan

Introduction & Motivation:

Dataset: Kaggle (https://www.kaggle.com/theworldbank/health-nutrition-and-population-statistics)

Within the dataset, there are 345 quantitative features, both health and statistical, for most countries/regions in the world. The data points are sampled yearly between 1960 and 2015, but there is no guarantee that every feature has a datapoint in each year for each country.

When we both initially flipped through this dataset, we were intrigued by the quantity of features within the dataset (there were over 300!). There wasn't any single feature that we wanted to explore from the beginning, but rather we wanted to explore the overall differences between countries. Our motto throughout this project was to let the data speak for itself rather than intentionally searching for something. We initially spent our time playing around with the data, whether it be selecting only certain years or ignoring certain indicators. After our exploration, we thought it would be interesting to build a machine learning classifer that attempted to predict which continent a country belonged to solely off of 29 semi-popular indicators. Below, you will find our journey through this large dataset and hopefully be amazed about the trends occurring across the globe.

1. Data Loading and Wrangling

The initial dataframe is an inconvenient form; the index is the country name and there is a specific column that is named “Indicator Name”, which lists a specific indicator. The remaining columns are the years 1960-2015, which contain the corresponding value of the indicator for each year. This dataframe is useful if we want to look up a specific indicator for a specific country, but that’s not what we want. Our goal is to explore different indicators across different years. We begin by pivoting the table such that each indicator is it’s own column and the year is a feature. We then address the overwhelming amount of null values by dropping columns that have more than 8000 null data points. Afterwards, we decided to introduce a “Continent” feature, which states the continent of the country. With this wrangling aside, we are ready to start exploring the data!

Load Data

Data Wrangling

Pivoting the table such that each row has data for each country in a specific year

Add Continents Column

Code block after this one gets rid of rows that are not countries associated with a continent.

Filter Null Values

Removes columns that have greather than 8000 null values.

Mean and Median

Create median and average rows for each country across all *55* years

Groups countries by continents then averages the countries values for each year

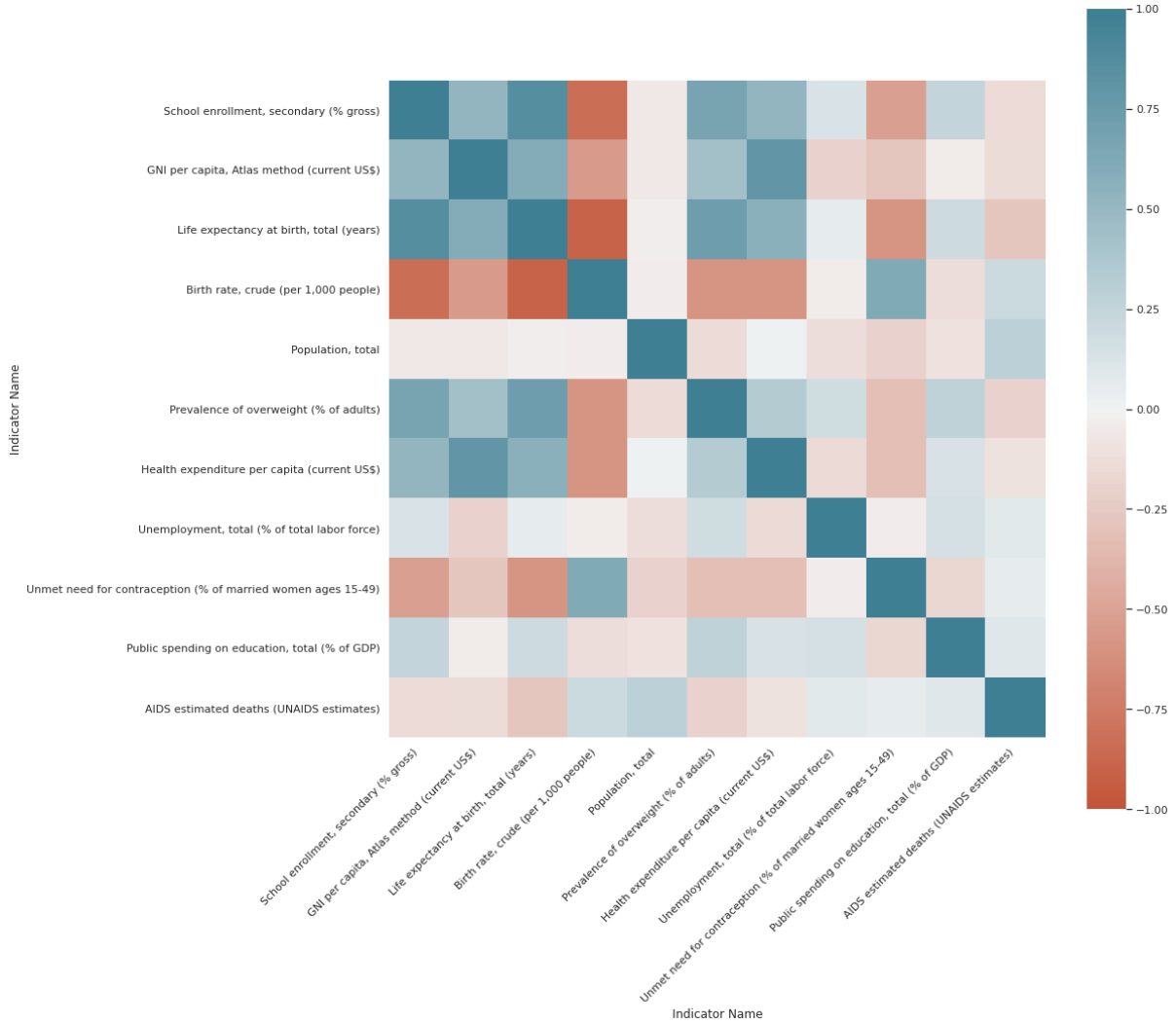
2. Visualization

Below are some visualizations that we made for certain indicators that we found interesting.

2.1 Correlations

With over 300 indicators in this dataframe, there is a lot to explore; frankly, too much! After viewing each indicator that contained a substantial amount of data points, we narrowed down our correlation matrix to these 11 indicators.

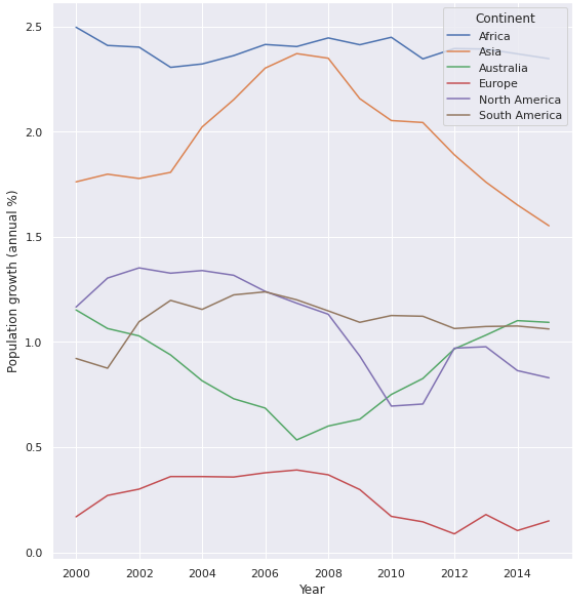
Here are some key takeaways: birth rate has a highly negative correlation with second school enrollment and life expectancy. This suggests as birth rate decreases, both higher education enrollment and life expectancy increases. Another interesting, related observation is that as higher education enrollment increases so does life expectancy.

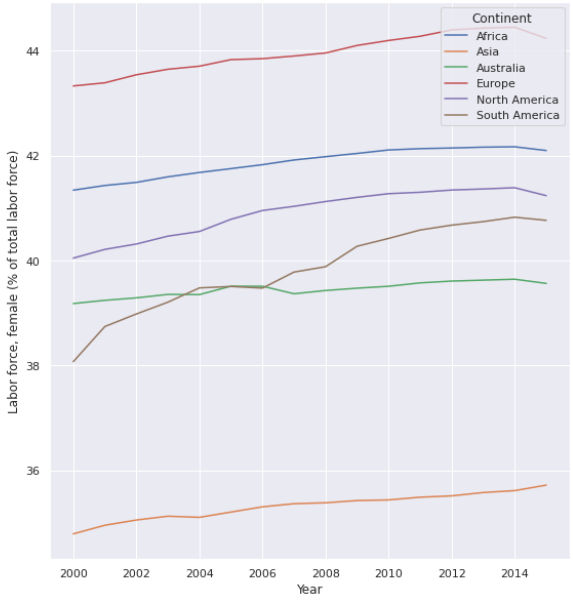


2.2 Further Exploration of Several Interesting Indicators

Population Growth and Female Labor Force Percentage Over Time

An interesting take away from the population growth chart on the left is the impact of the 2008 global recession on population growth. Leading up to 2008, there was a period of increasing population growth, but, suddenly in 2008, the population growth slowed down significantly. This potentially suggests that individuals were reluctant in producing offspring when the global economy became unsteady. Oddly, individuals in Australia/Oceania (the continent), did not follow this trend; instead they had a significant population boom in 2008. The chart on the right shows that women across the globe have been increasing their presence in the labor force. Sadly, there are stark differences between the continents, but the steady growth provides great optimism.

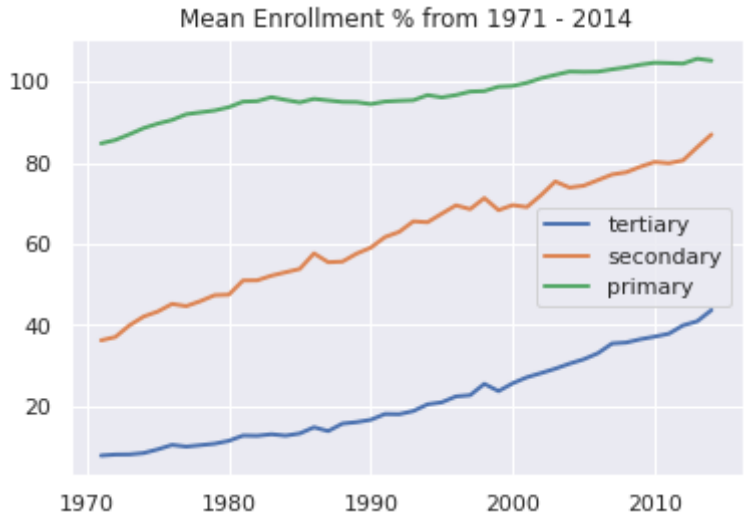


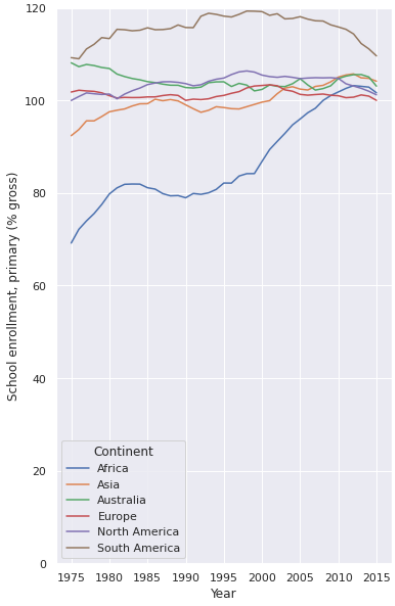


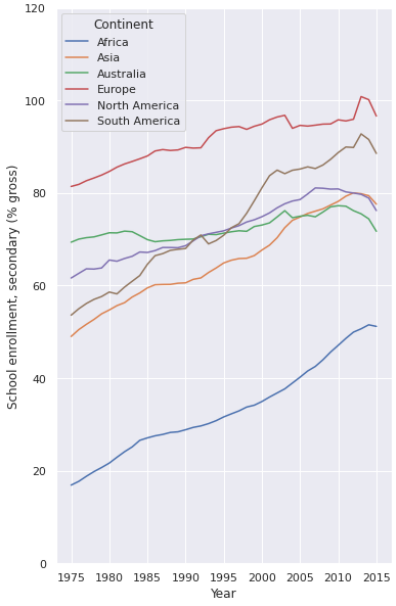
Gross Enrollment Ratio (GER)

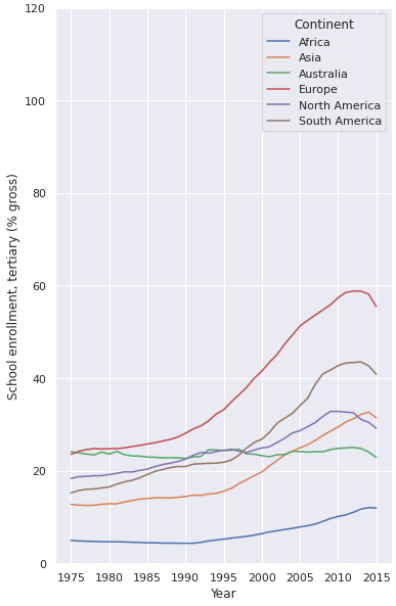
- School enrollment, tertiary (% gross)
- School enrollment, secondary (% gross)
- School enrollment, primary (% gross)

Gross Education Enrollment (GER) is a statistical measure used to show the ratio of the number of students who live in that country to those who qualify for a particular grade level (primary, secondary, tertiary). It can be over 100% as it includes students who may be older or younger than the official age group.


Mean Enrollment % from 1971 - 2014
tertiary
secondary
primary
100
80
60
40
20
1970
1980
1990
2000
2010


School enrollment, primary (% gross)
Year
Continent
Africa
Asia
Australia
Europe
North America
South America


School enrollment, secondary (% gross)
Year
Continent
Africa
Asia
Australia
Europe
North America
South America


School enrollment, tertiary (% gross)
Year
Continent
Africa
Asia
Australia
Europe
North America
South America

Worldwide

There's an overall increasing trend for enrollment rates for all grade levels. From 1970 to 2015, the mean secondary and tertiary enrollment rate almost doubled. However, until 2015, the higher education enrollment rate is still below 50%. We still have a lot of room for improvement.

Continental

We further analyze the education enrollment rate based on the continental. There are two reasons for that. First, we want to see how the education level differs around the world. Secondly, we want to classify continents later, so it's important to take a look at the general trends first.

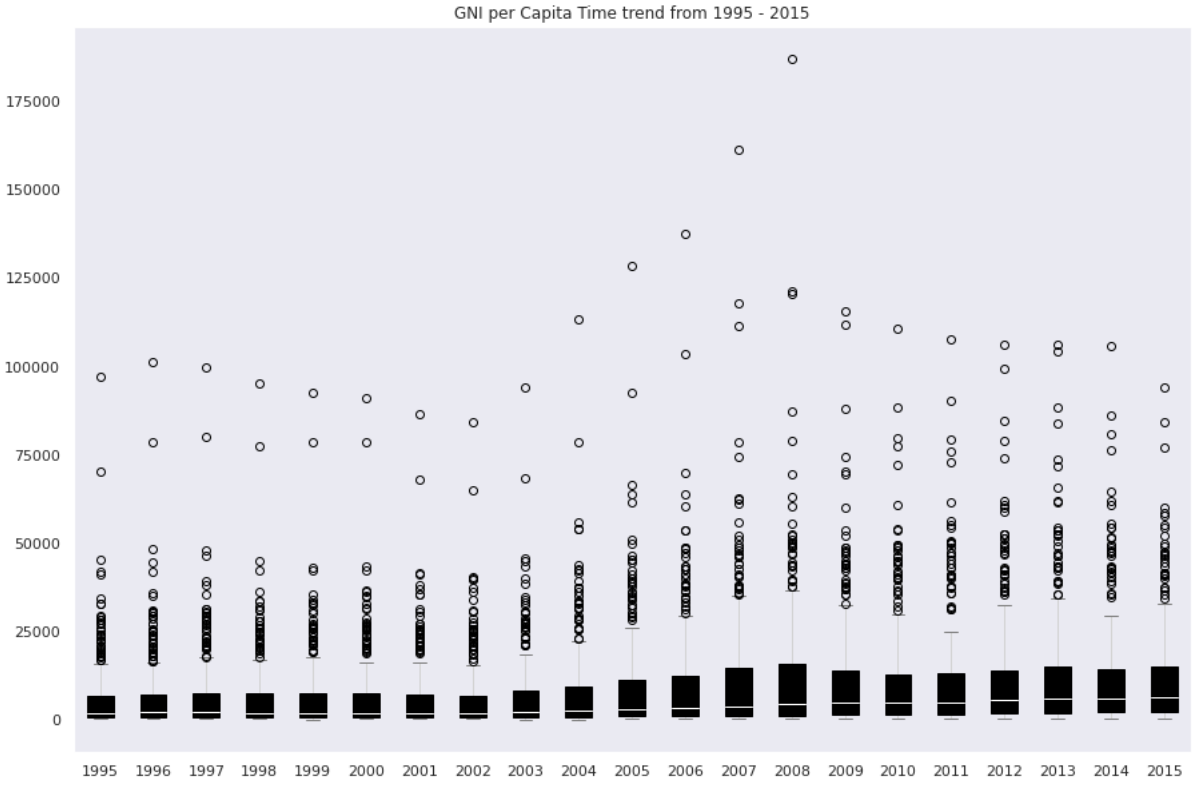
From the figures below, we find that Africa shows a steady improvement from 1975 to 2015. Although its GER for secondary and higher-level education still significantly lower than other continents, its mean GER for primary education has come to about 100%. Meanwhile, Europe continuously leads GER for the last 40 years, and its secondary GER has also almost reached 100%. It's interesting to find an inverted U-shape curve for South America's GER primary education with a peak of about 120%. It means that lots of students who were older than the official age group went to primary school in the 90s.

GNI per capita

The GNI per capita is the dollar value of a country's final income in a year, divided by its population. It reflect the average before tax income of a country's citizens and suggest the economic strength and the general standard of living of a country.

Worldwide GNI per capita from 1995 - 2015.

Figure

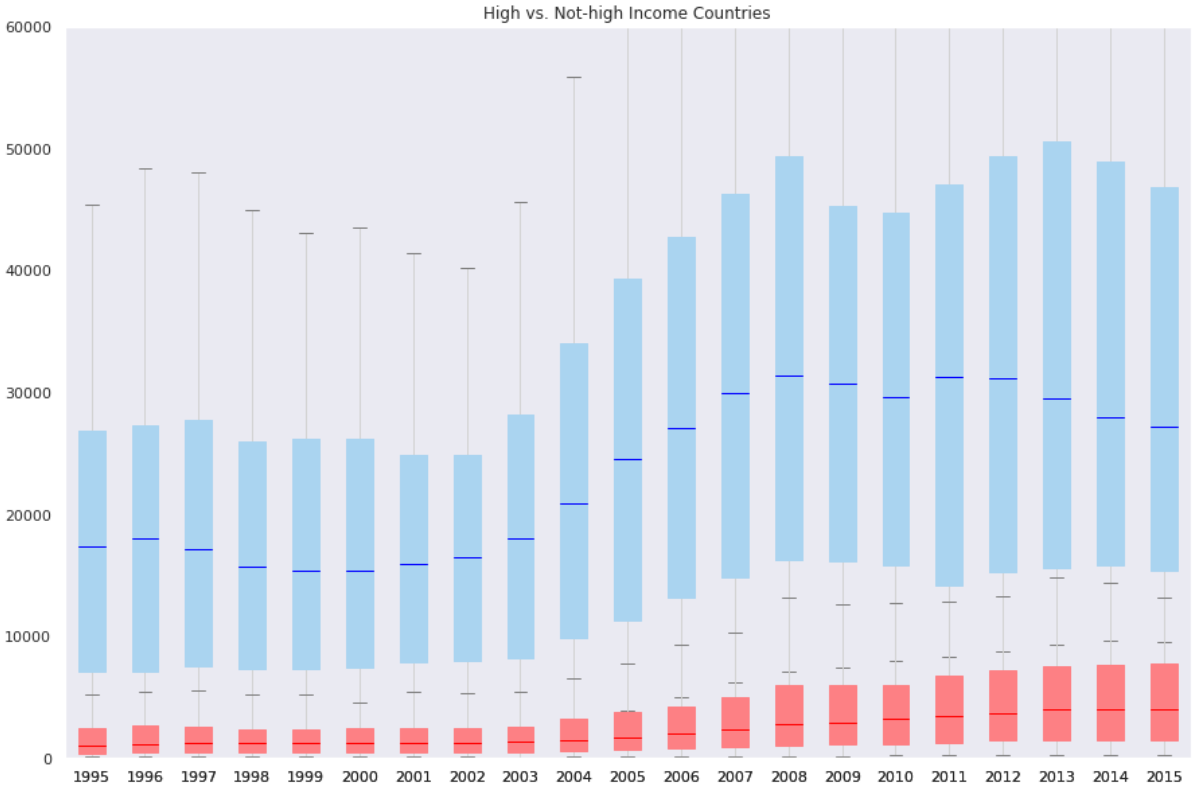


The above figure shows the economical world trend from 1995 to 2015. Overall, it's excited to find a slightly increasing trend from 1995 to 2015. However, there are also lots of outliers with significantly higher GNI per capita than average, suggesting the vast differences in living standards among countries.

To better understand the world trend of GNI per capita in the past 20 years, we divided countries into two groups (High-income economies and Non-high-income economies). By definition from World Bank (https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups), high-income economies are countries with GNI per capita higher than $12,476 (in 2015).

The figure below suggests that the non-high-income-economies countries (red boxes) show an overall increasing trend from 1995 to 2015 while high-income-economies countries (such as the US and many European countries) reach the peak in 2009 and then have a decreasing trend afterward. The 2008 economic crisis can probably explain the trend for high-income countries.

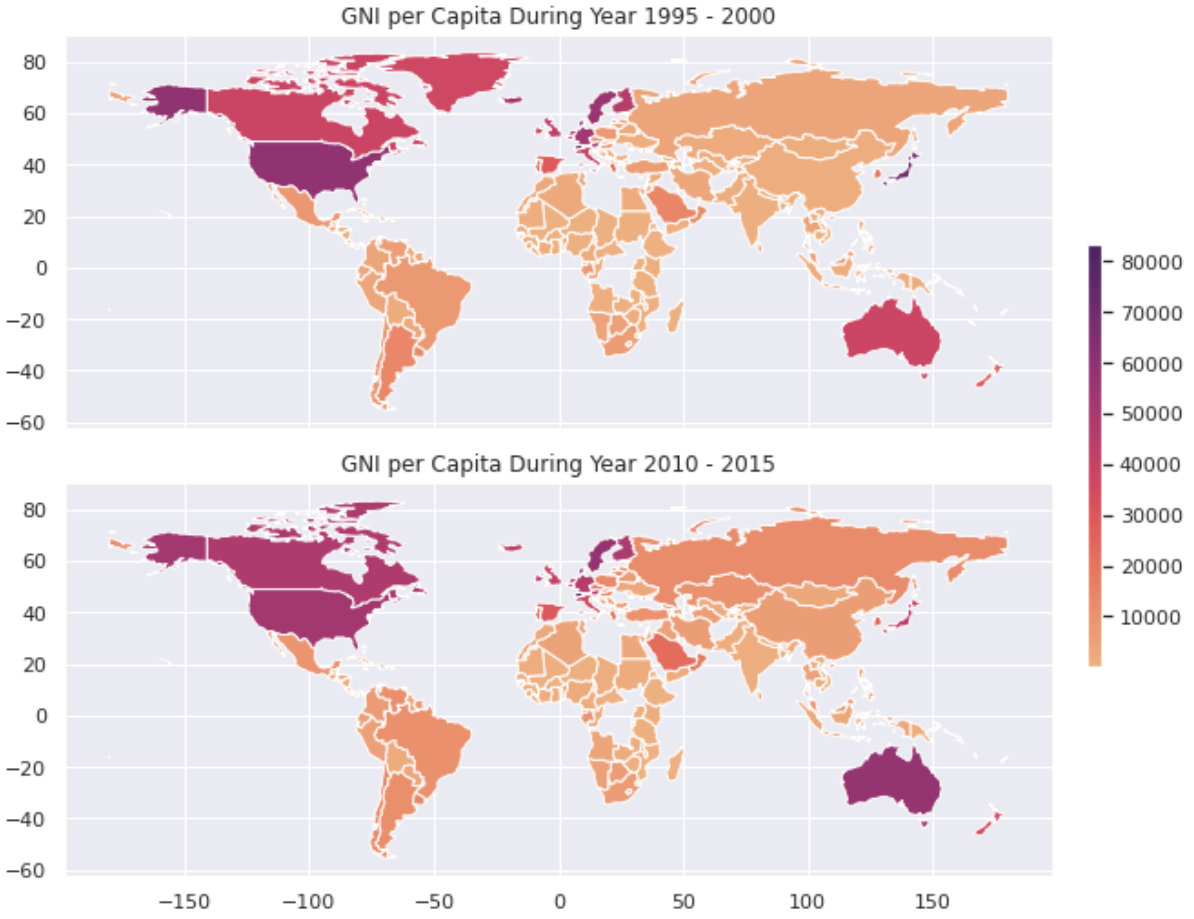
Figure

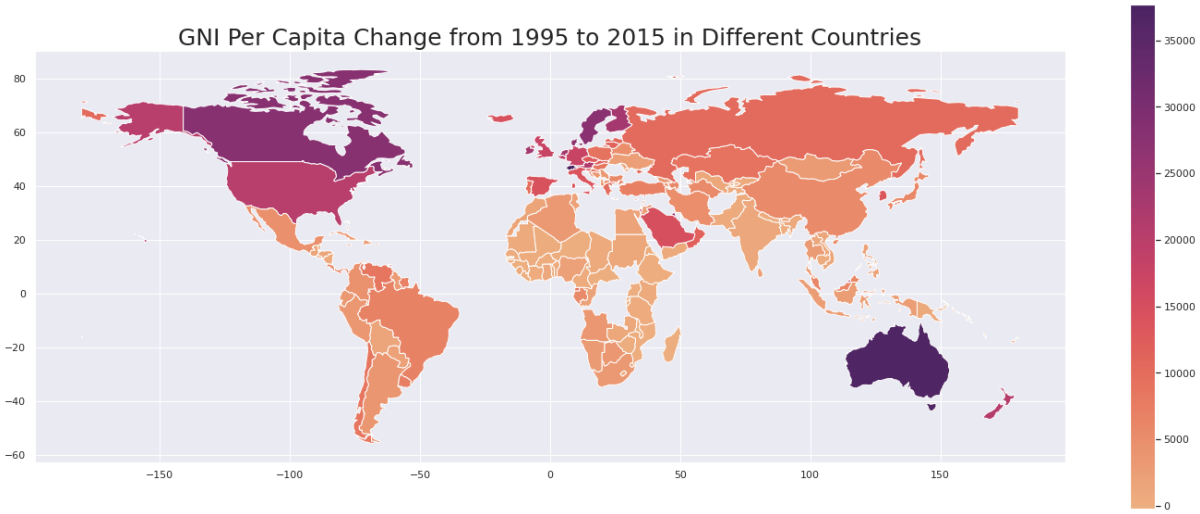


Maps

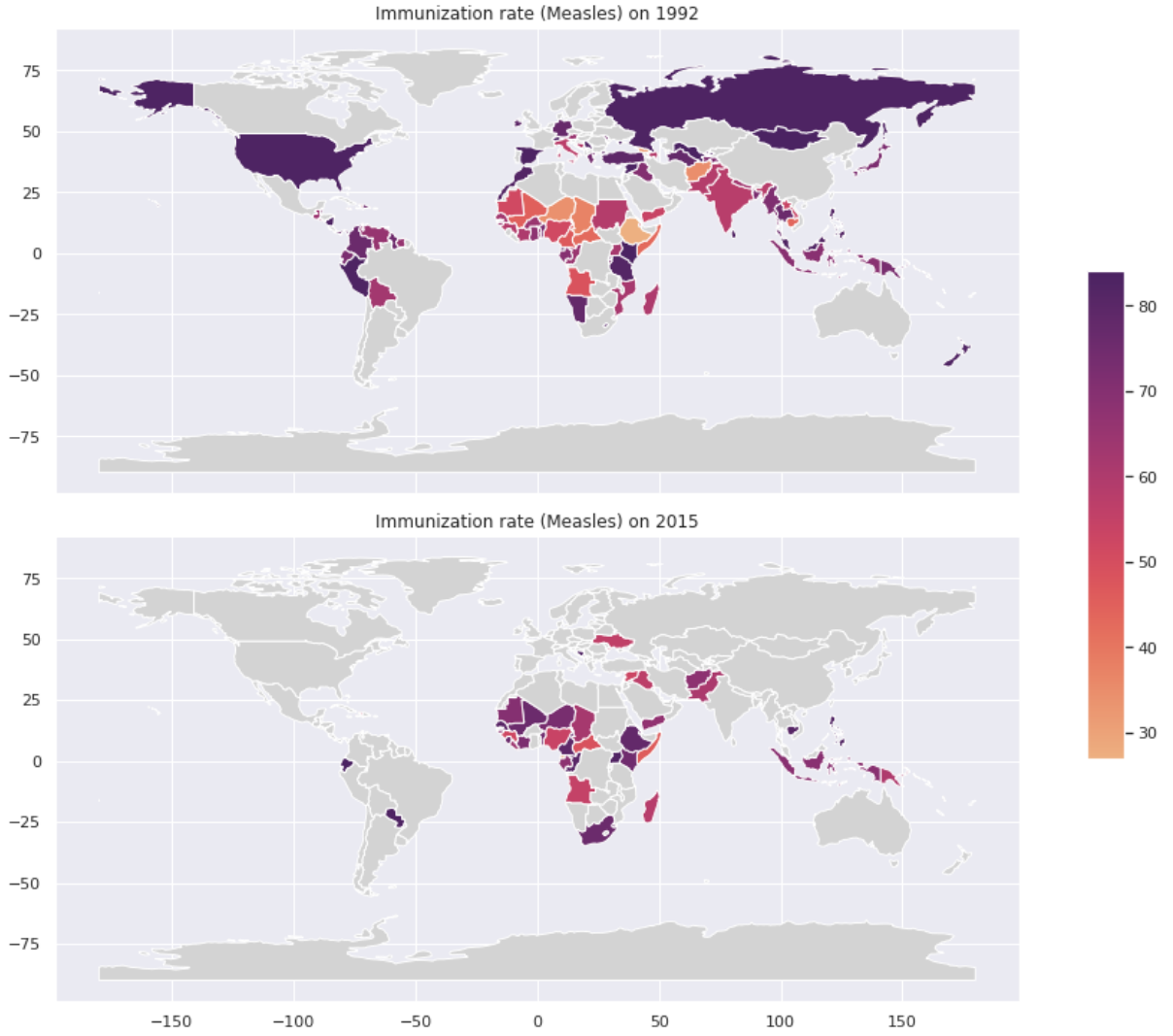
The first two maps below show the GNI per capita in the two different year ranges, and the last map in this section shows the growth (in $) from 1995 to 2015. The deeper the color, the stronger the economic strength of a certain country. Countries that don't have enough data in the dataset is blank on the map.

Overall, North American and European countries' economic level is significantly stronger than countries on the other continent. It's surprising to find that Australia and Canada have the most significant economic growth in the past 25 years. We initially thought China would lead the economic growth for the last 25 years, but apparently, the population has greatly influenced the final result.





Immunization



Models

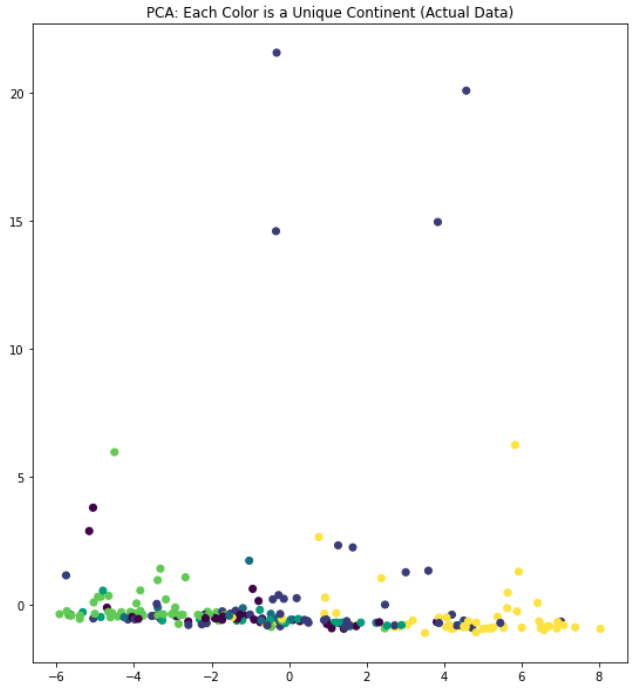
Objective: The goal of our modeling is to classify continents based on the 29 indicators we are interested in. This includes indicators (school enrollment, GNI per capita, population, life expectancy, etc) we visualized above, but also other indicators like life expectancy and mortality rates that we didn't visualize.

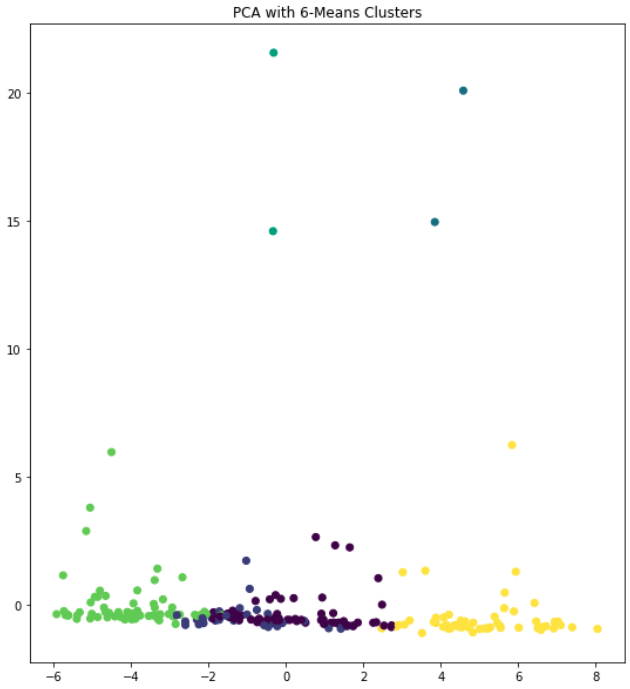
We first tried unsupervised learning (PCA and k-means), hoping to see whether the continents clustered together. Although we didn't get really good accuracy from it (acc = 50%), we saw that European and African countries had a significant border between them whereas Asian countries didn't cluster as nicely. You can find the code and theory behind its development in the Colab notebook 'Extra Visuals & Models'.

In addition, we tried 4 supervised learning models (two neural feedback models, random forest classifier, and SVM) and three of them are presented below. We use ¾ of the dataset for training and use the rest ¼ of them for testing. For each model, we show both accuracy and confusion.

PCA plot

```
X-Axis is PCA1 & Y-Axis is PCA2
Yellow: Africa
Light Green: Europe
Greenish Blue: Australia
Blue: South America
Purple: Asia
Accuracy Score:  0.5155555555555555
```





Preparing the Models

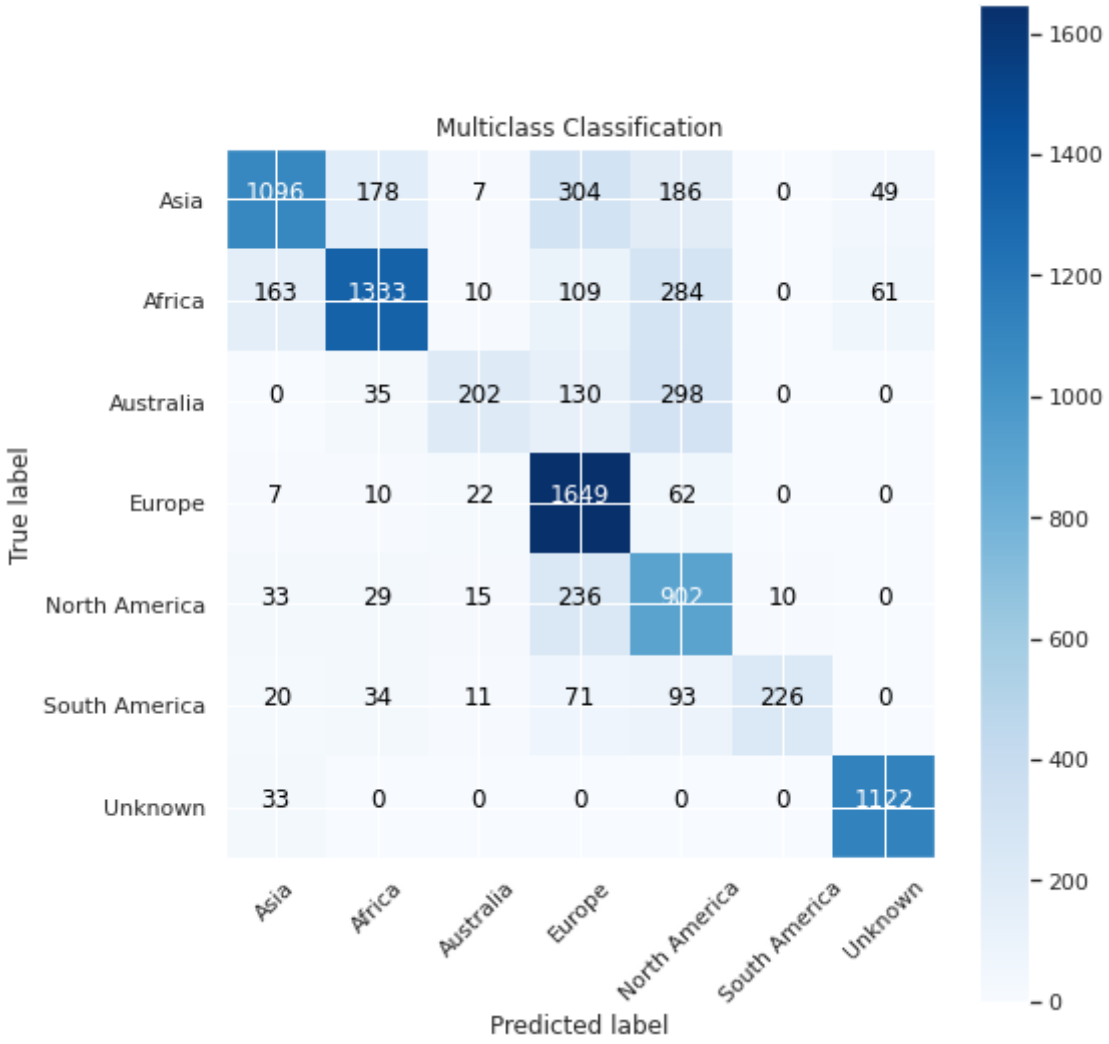
Neural Network

MulticlassClassification

Modeling, Training, and Evaluation

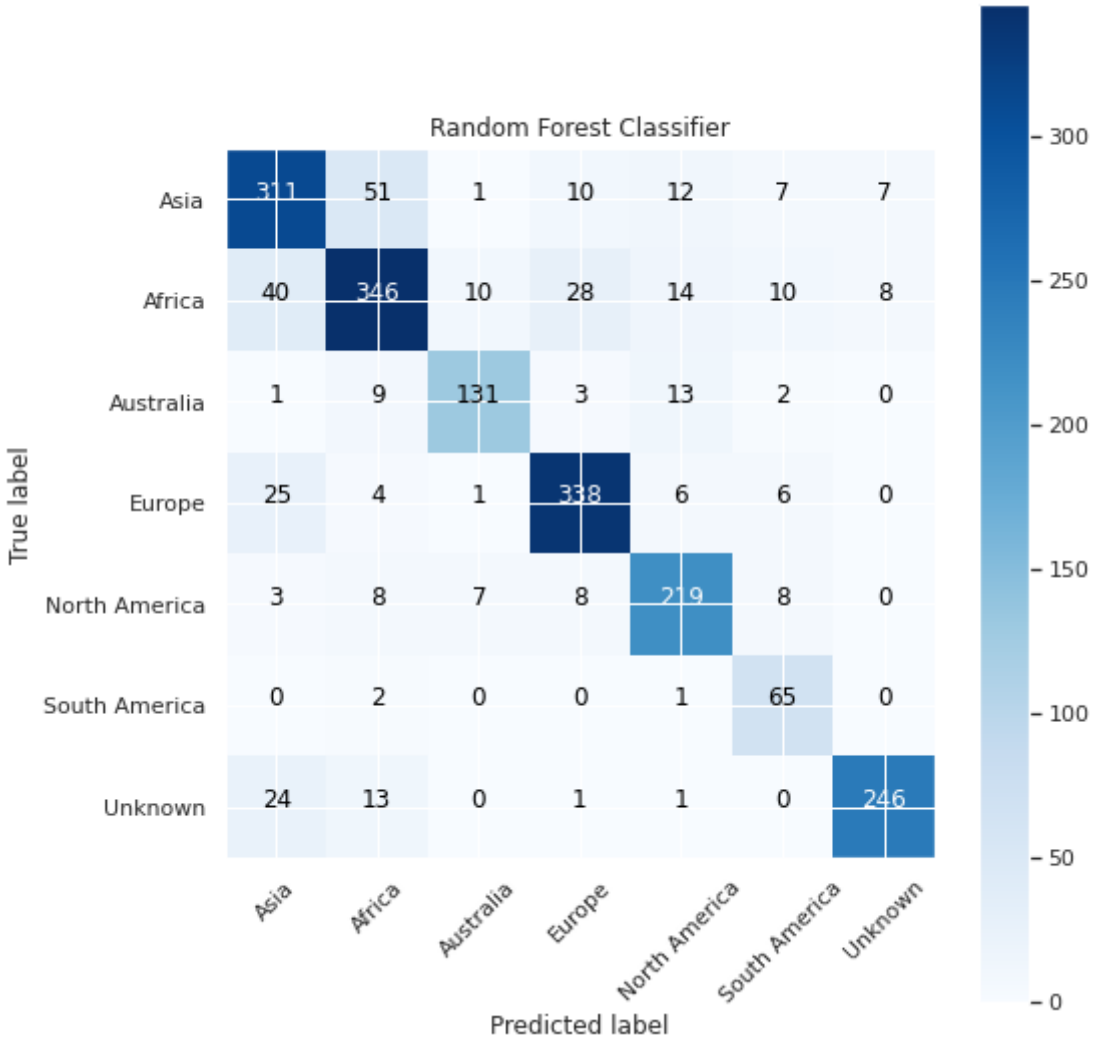
Running the Model

```
MulticlassClassification: Final Accuracy: 72.3145, Final F1 Score: 71.6
753
```



Random Forest Classifier

```
Test accuracy: 0.828
```

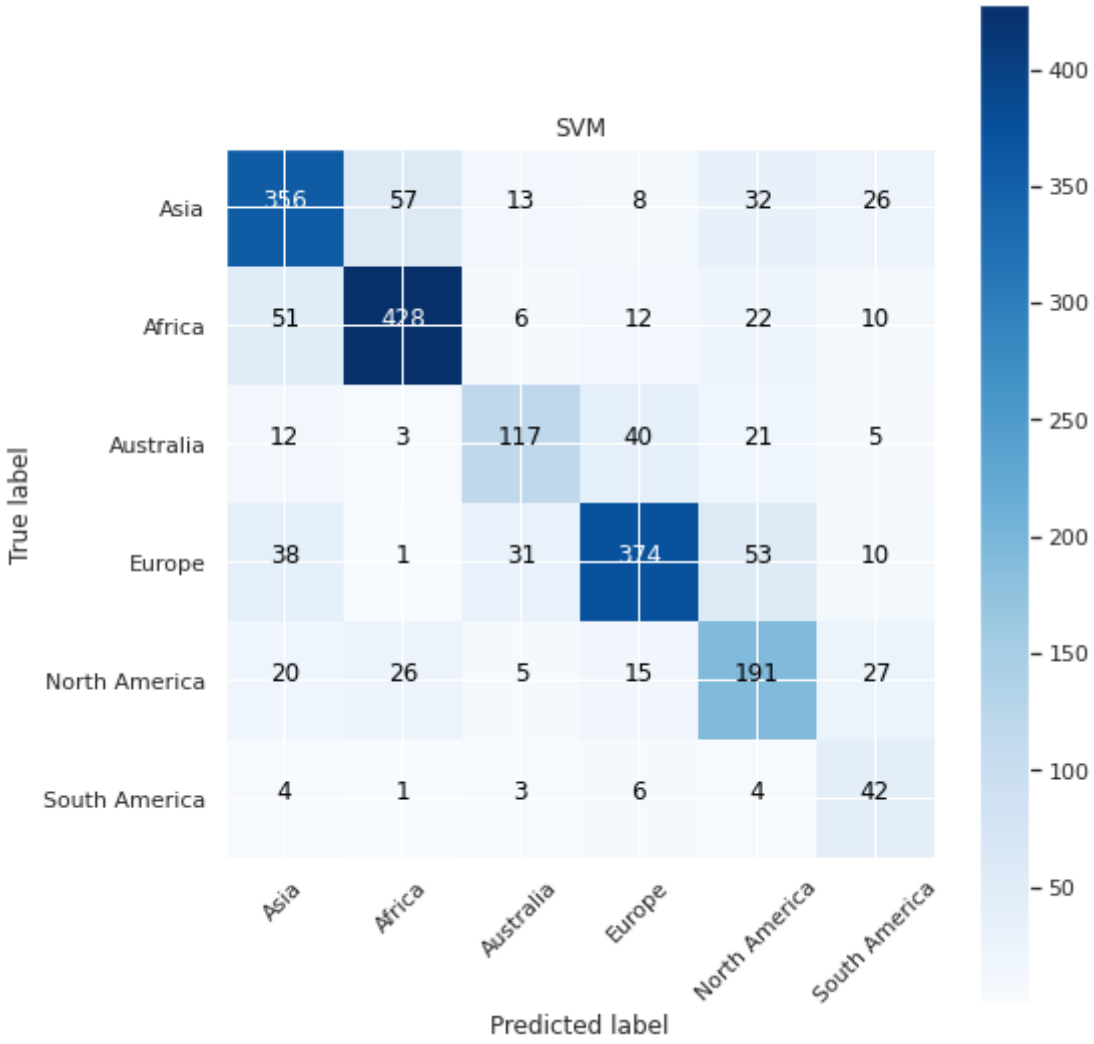


SVM Model

Preparing the model

Running the model

```
Test accuracy: 0.7285024154589372
```



After we run all the models, the **random forest classifier** performs best, with an accuracy of 84% (seems promising!). Both the neural network (multiclass classification) and SVM have accuracy rates of around 75%. From the confusion matrix, we also find that Asia and Africa have the highest similarities, making it difficult to accurately predict (as a side note: this might be a consequence of including most Middle Eastern countries in Asia rather than separately categorizing them; moreover, Southeast Asian countries have unique properties that cause them to be more similar to Africa rather than central Asia).

Challenges and Obstacles

Our first, and probably the most difficult challenge was to determine what to do with null values in the data frame. At first, we naively tried to drop any row that contained a null value, but consequently, we were left with around 40% of what we began with; this clearly wasn't the option. With the help of our project TA, Arielle, we were able to pivot the table to make it user-friendly. From there, we decided to filter out any columns that had more than 8000 null entries; this number was decided by looking at the distribution of null entries across the columns. However, this didn't entirely solve the problem since there were still null entries within our data frame. Moreover, these null entries prevented us from doing any machine learning. After some Googling, we found that pd.interpolate() with a linear method was a decent way to fill in null values since it used linear regression to fill missing values.

Our next challenge was deciding what to build our machine learning model upon. We had an abundant amount of features, but we didn't know which one would be interesting in predicting. We debated predicting population or GNI per capita (a regression-based classifier), but decided that they wouldn't be interesting. We thought it would be cool to see if we would be able to predict what continent a country belonged to. However, then we had to decide which features out of the 345 we would include. We narrowed it down to 29 by simply choosing well-known indicators along with some that we found particularly interesting.

Future Directions

The data we used only contains information between 1960 and 2015. One interesting thing to try would be to take data from 2019 for each country and see how our model classifies them. A question that we have is would more countries be classified as European since their statistics will have assumingly improved over the past 4 years.

Another thing that one could try is selecting different features other than the 29 we choose to train new models. Moreover, we could combine our 3 classifiers into a voting-like classifier where we let each of our classifiers cast a vote for its classification and then take whichever vote occurs the most; in the case of a 3-way tie, we could favor the tree classifier.

Another interesting thing to try would be to build some classifier that knows the number of countries in each continent; this additional knowledge would serve as a constraint for the classifier, making it less likely to over classify one particular continent (Africa has a lot of false negatives). In addition, after we have a better understanding of these features, we could weigh them differently and get a better prediction.