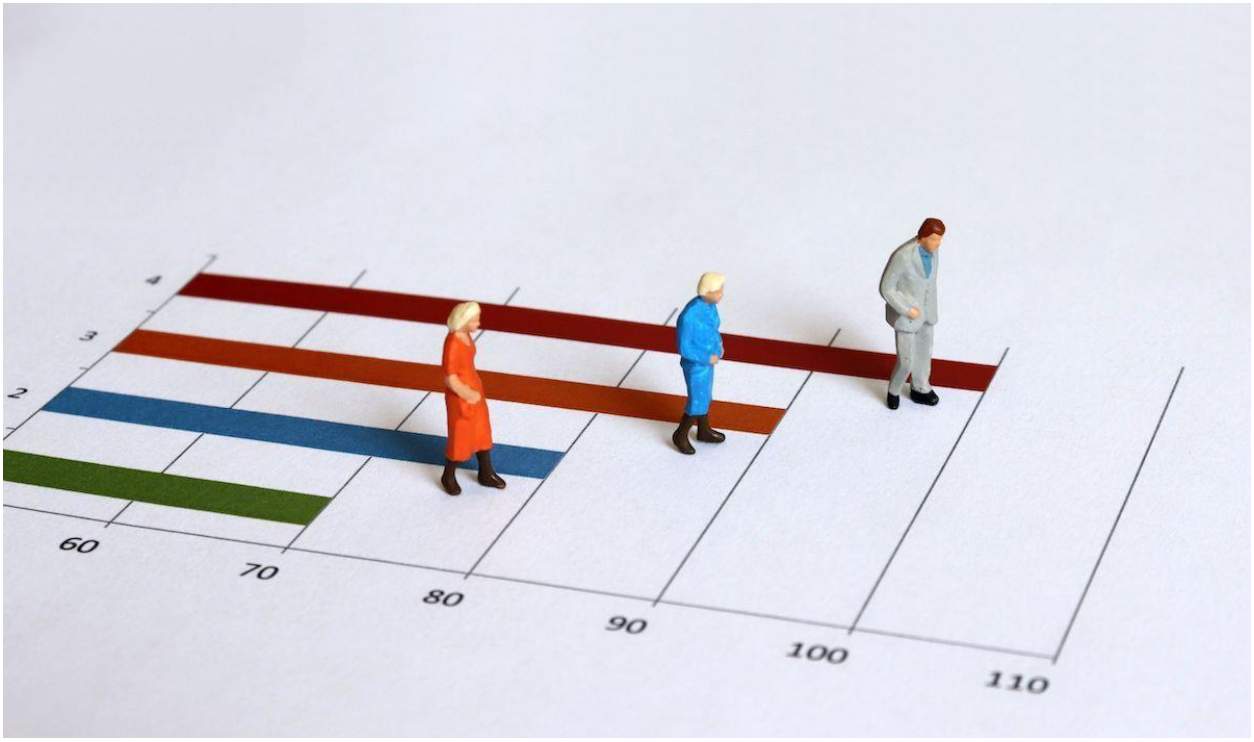
Predicting Life Expectancy

Capstone 3 Springboard Data Science Career Track

Final Project Report

Soumyadip Kundu



1. Acknowledgements

- Mentor – Kenneth Gil Pasqual
- Dataset – World Health Organization
- Cover image – Google Images

2. Background

The most important statistic for measuring population health is life expectancy. Life expectancy captures mortality across the whole life course, making it more comprehensive than the restricted measure of newborn and child mortality, which concentrates only on mortality at a young age. It provides information on the typical

death age for a population. A big topic at the World Health Organization for years has been life expectancy in various nations. Only a small number of parameters, including demographic demographics, income distribution, and death rates, have previously been taken into account. The National Centre for Health Statistics and the Center for Disease Control (CDC) have been closely examining various global characteristics and comparing them to the global statistic in order to better understand life expectancy in the United States. The life expectancy statistics have changed significantly as a result of the increased amount of growth in the health sector over the past 20 years, including the implementation of systematic immunization programs in many nations. This new dataset, which contains data spanning 15 years, now includes a plethora of different predictive variables.

3. Project Statement

Assessing the relationship between the various predictive factors that affect an individual's life expectancy in 193 countries. The impact of education and immunization coverage will also be evaluated. This problem will also look at the various predictive factors that will increase an individual's life expectancy in a developing country versus a developed country by a certain number of years. In short, we would investigate the predictive factors that health organizations must consider in order to improve life expectancy.

4. Data

The World Health Organization's Global Health Repository monitors all countries' health status and the factors that may influence this rate. This dataset contains data from all 193 countries over a 15-year period. The dataset consisted of 22 columns and 2938 rows, with 20 predictive features. The data was obtained from the Kaggle website. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

5. Executive Summary

To evaluate the factors, 22 features were taken directly from the dataset or derived from it. This was a regression problem, and the following regression models were used:

- Linear Regression
- Lasso Regression
- Ridge Regression

- Random Forest Regressor
- XG Boost Regressor

The dataset was split into a 80-20 train-test split with shuffle before the models were evaluated. Cross validation using GridSearchCV and RandomizedSearchCV were done and hyperparameters were tuned on both the Random Forest Regressor Model and the XG Boost Regressor models, since they were the best performing models. After tuning, both the Random Forest Regressor model (0.968) and the XG Boost Regressor model (0.962) were the best models. Both the RMSE scores and the R2 scores were considered for this project. Both the models predicted HIV/AIDS to be the best predictor for life expectancy.

6. Data Cleaning and Preprocessing

The dataset contained 2938 rows and 22 columns. There were two categorical variables: country name and status(developing or developed). The rest of the features were all numerical. Out of the 22 columns, population, Hepatitis B and the GDP column had the most missing values(~25%). Data cleaning and preprocessing are critical steps in gaining an understanding of the data. Preprocessing is required to feed the data into the algorithms so that they can function properly. Missing values can be a real problem because almost all algorithms demand the complete set of data yet fail to execute properly if certain points are missing. Data cleansing is therefore crucial. There are several ways to impute the missing values. In this project, we used the Simple Imputer function to fill the missing values with the median values. The dataframe was separated into numerical and categorical data. The numerical data was normalized using Min-Max Scaler so that the values were between 0 and 1. The categorical data was encoded using Label encoder on the status column. It was determined that the column year had no correlation with life expectancy. Hence it was dropped. The data was finally split into a 80-20 train-test split with shuffling.

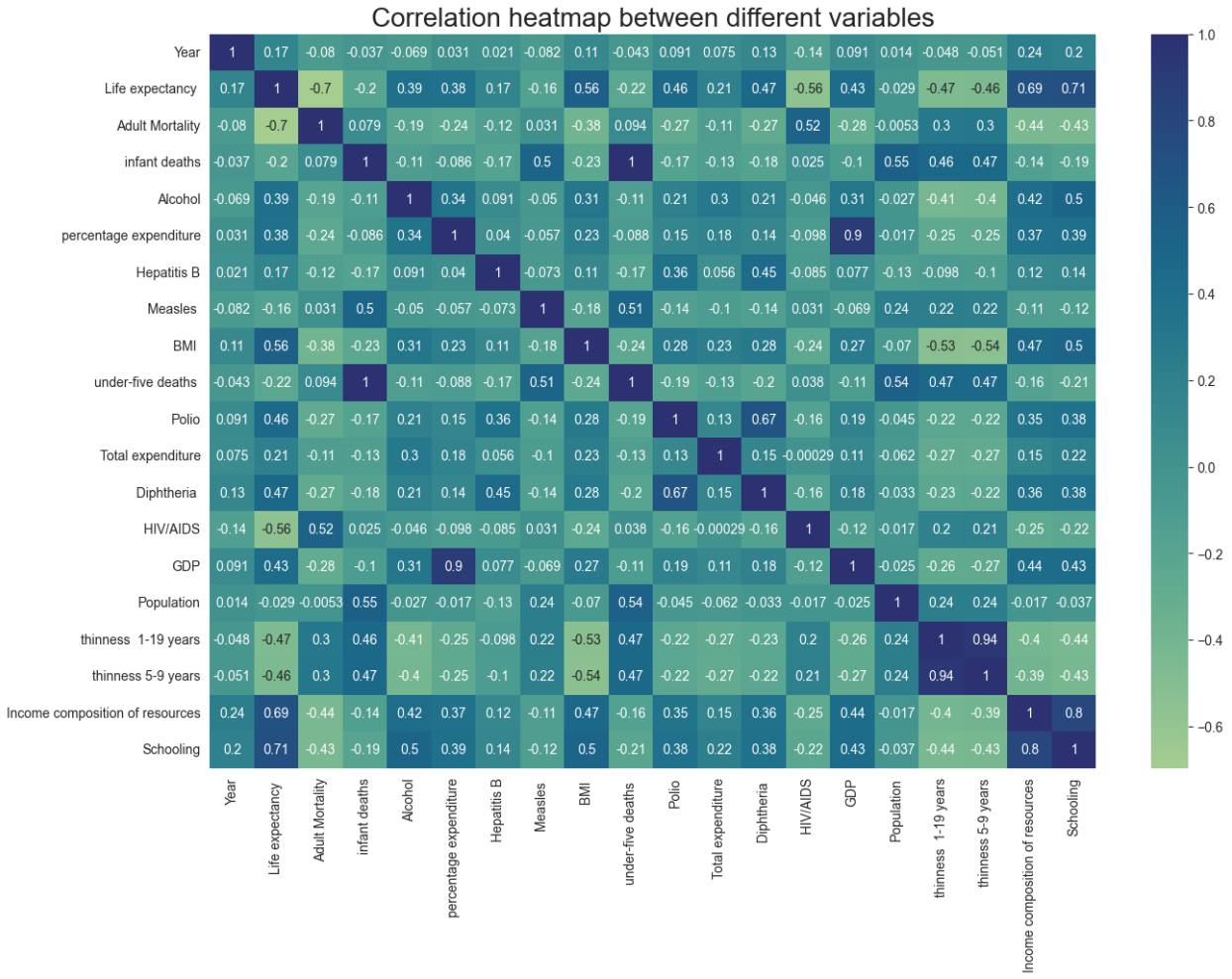
7. EDA

Prior to modeling, EDA was conducted on a number of variables that were significant risk factors for cervical cancer. The first relationship between the features was visualized using Python's Matplotlib, Seaborn, Plotly, and using Chloropleth maps.

Correlation heatmap

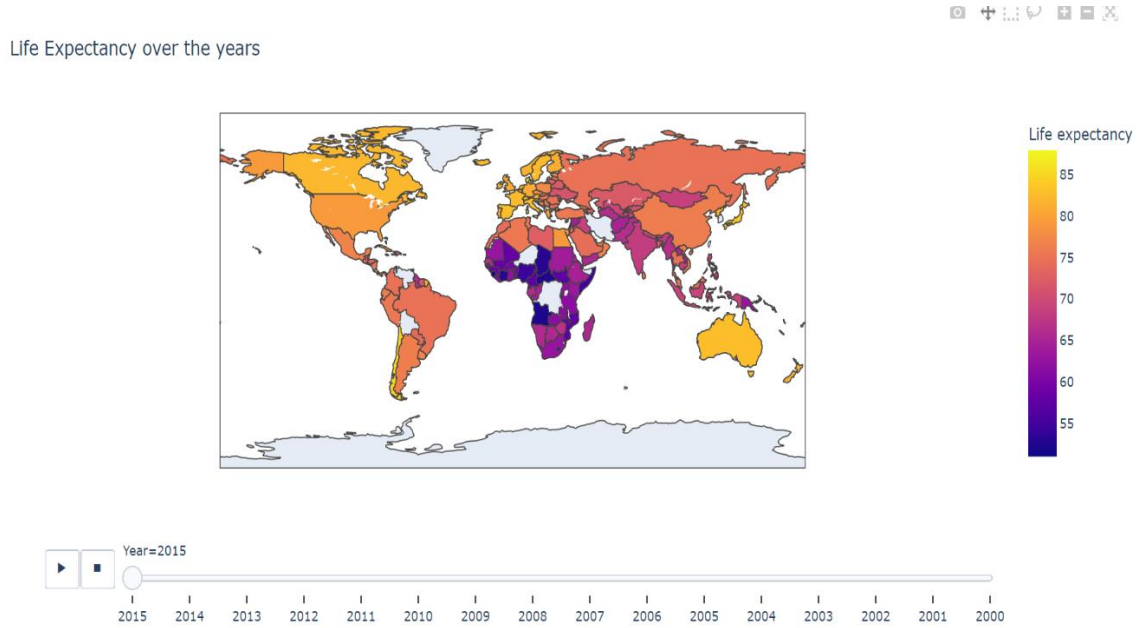
I generated a preliminary heatmap with different variables that were associated with life expectancy. From the correlation heatmap, the following features have strong to moderate correlations with life expectancy:

- Positive correlations: Schooling (0.71), Income composition of resources (0.69),GDP(0.43)
- Negative correlations: Adult Mortality (-0.70), HIV/AIDS(-0.56)

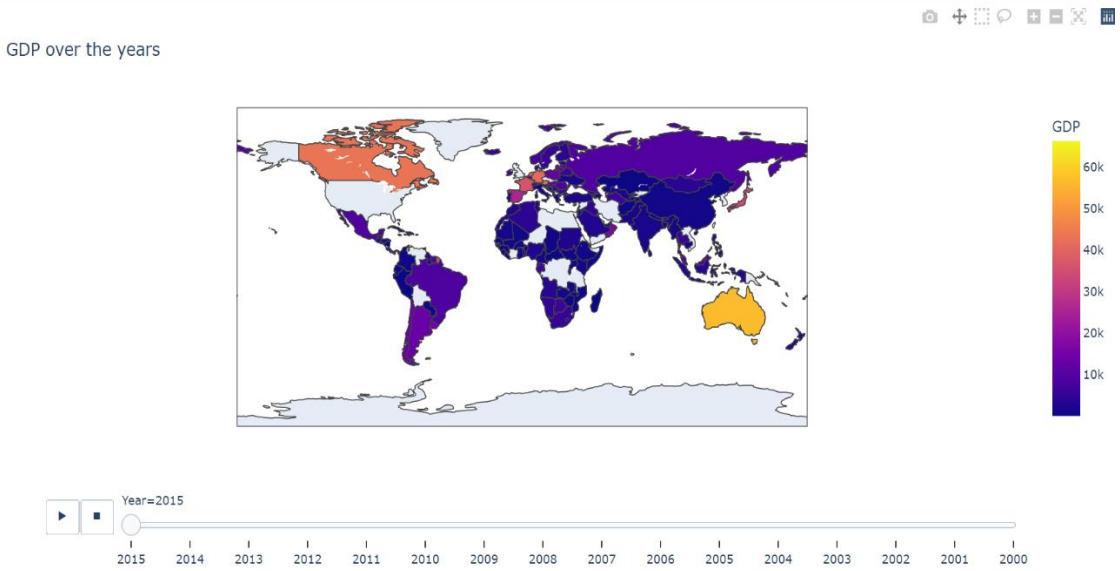


Developing Chloropleth maps based on different years

A particularly effective method for plotting the data points on a map is to use Chloropleth maps or graphs. It offers improved graphics and facilitates our understanding of the material.



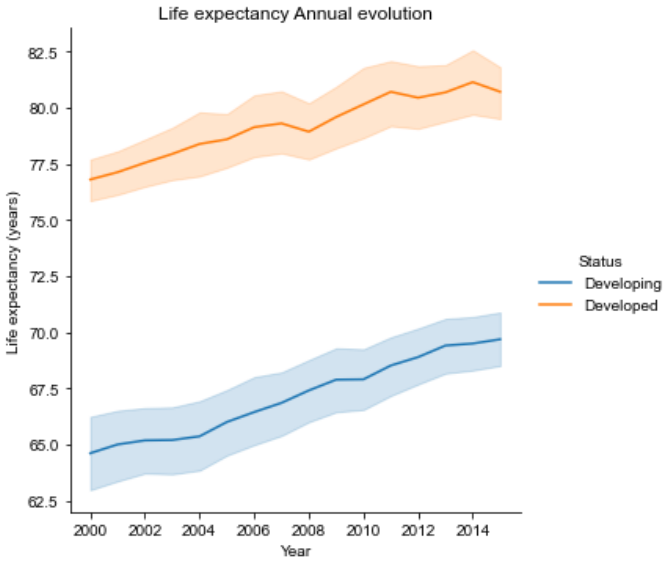
Animated hyperlink: https://skundu01.github.io/



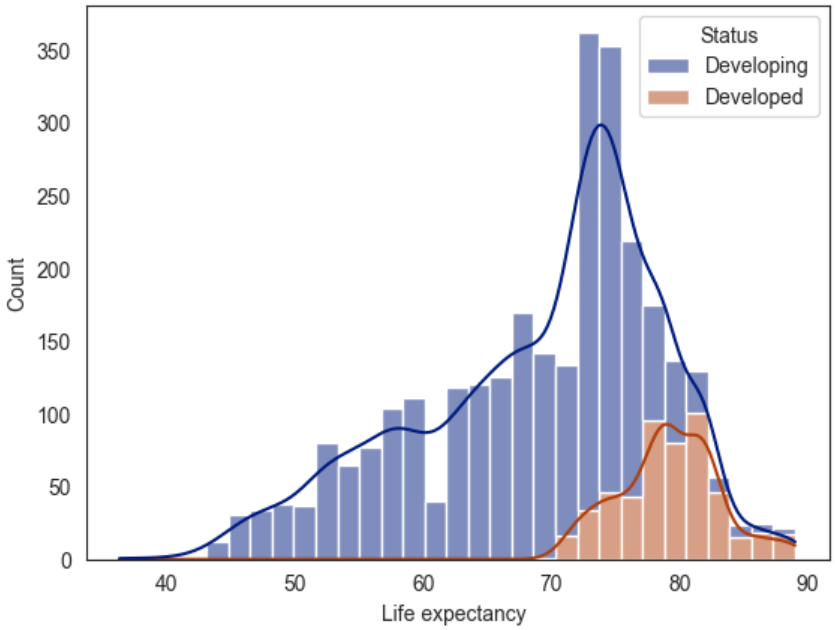
Animated hyperlink: https://skundu01.github.io/index1.html

Annual Evolution of Life Expectancy over the years

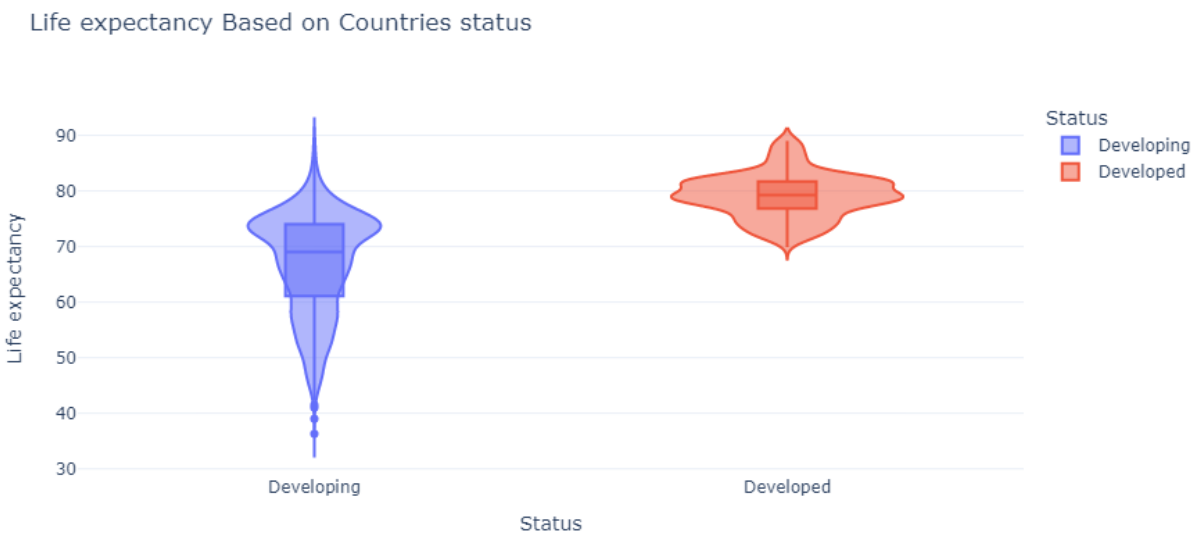
The following graph shows the annual evolution of life expectancy in the developing vs developed nation. It is evident from the above plot that irrespective of the status life expectancy has had a positive trend over the years. Developed countries have also had more years of life expectancy throughout the 15 years.



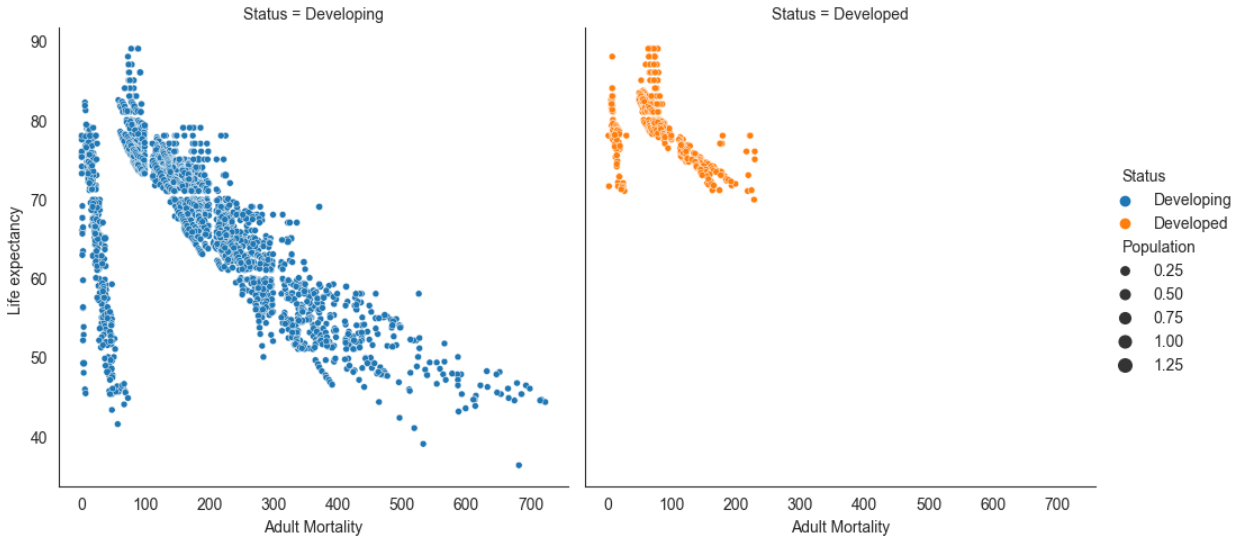
Distribution of Life Expectancy in developing vs developed nation



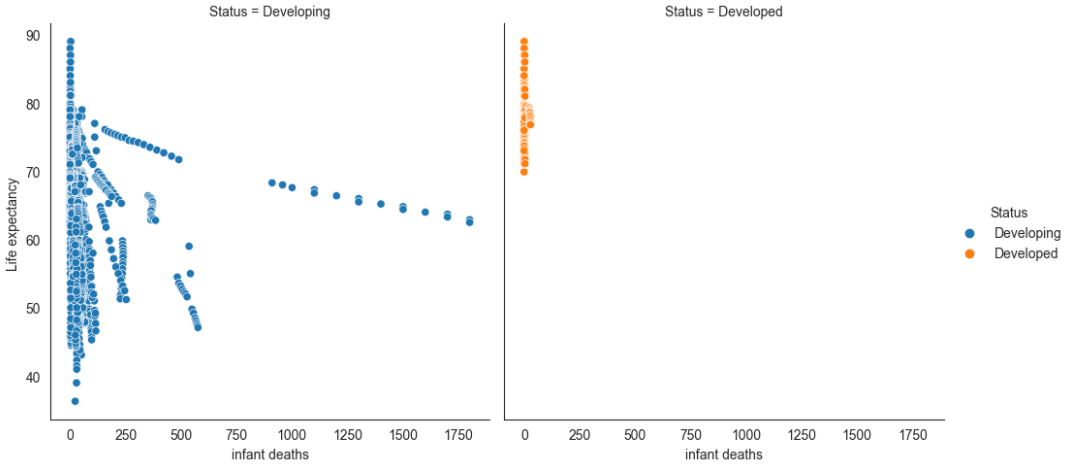
Violin plot of Life expectancy based on status



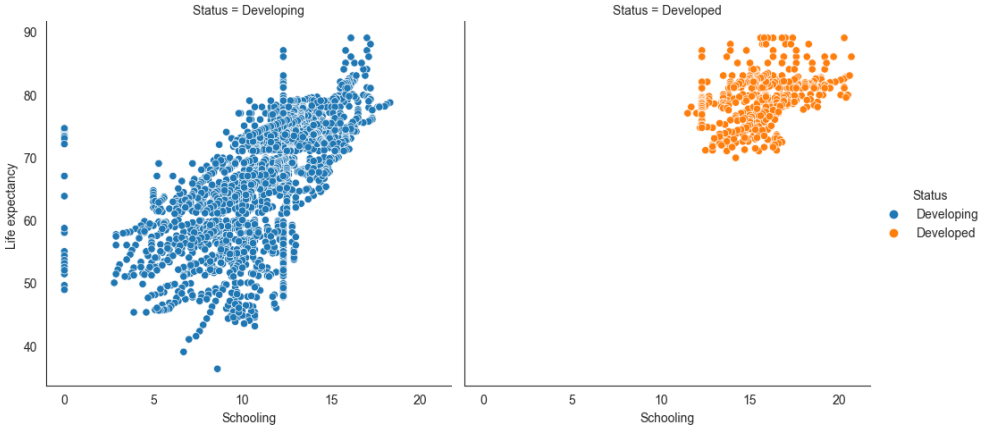
Life Expectancy vs Adult Mortality



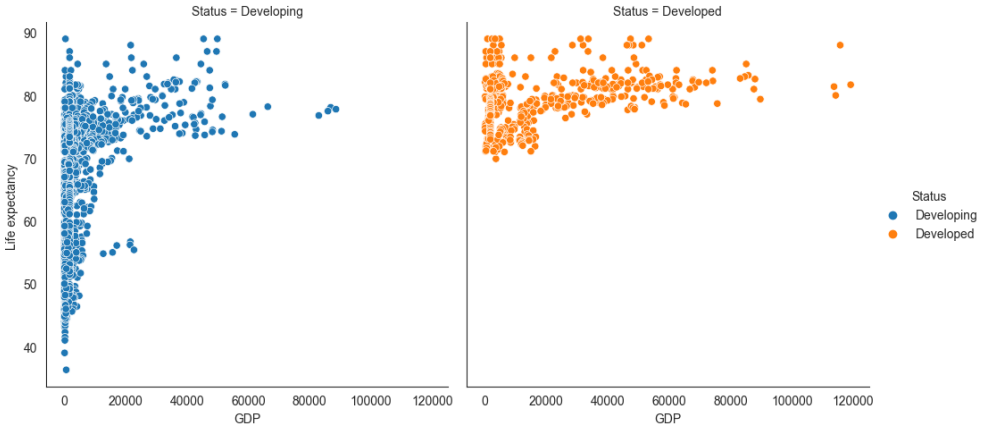
Life Expectancy vs Infant deaths



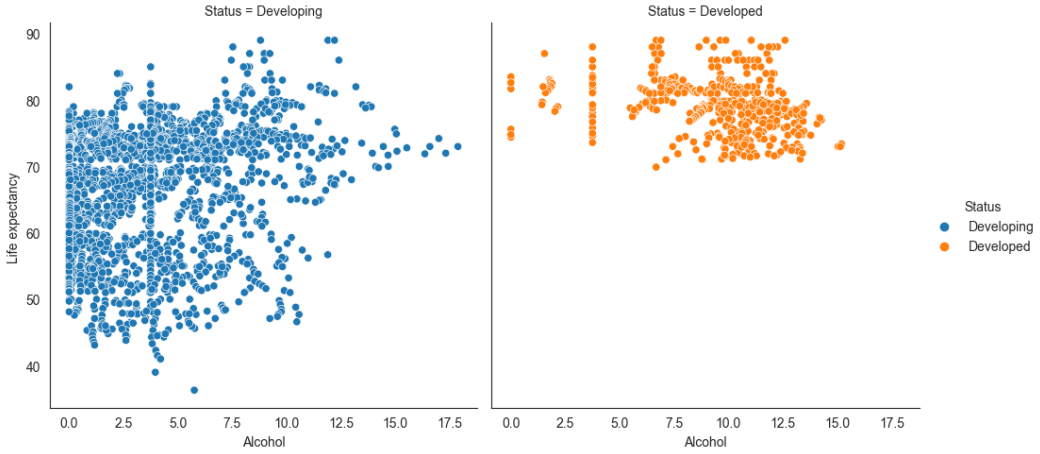
Life Expectancy vs Schooling



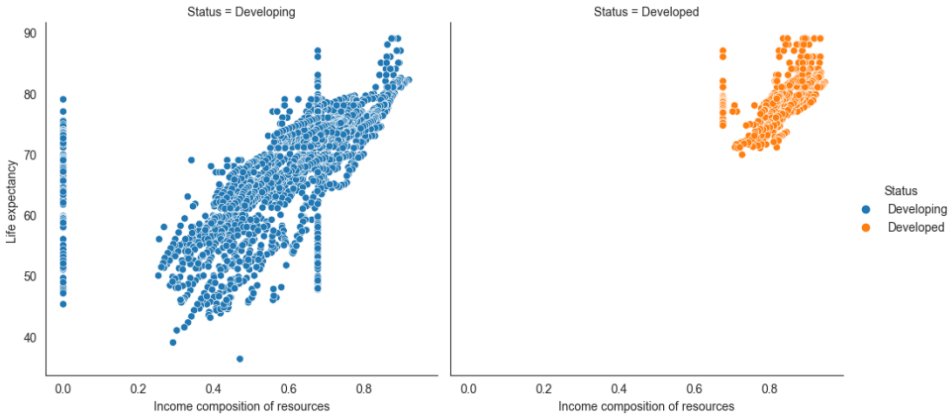
Life expectancy vs GDP



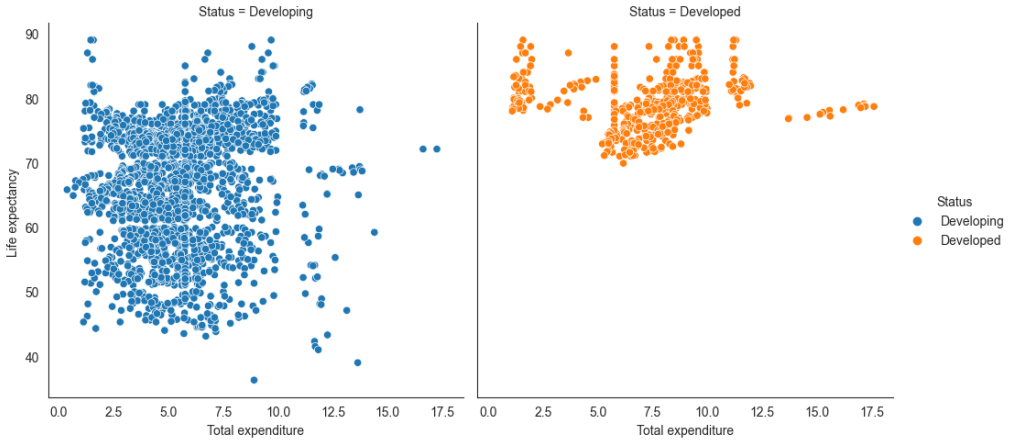
Life Expectancy vs Alcohol Consumption



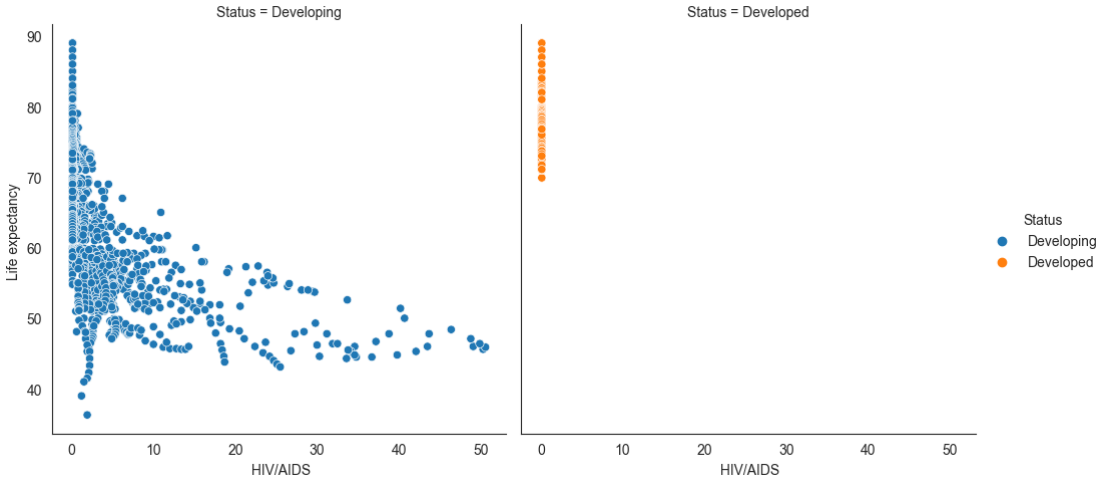
Life Expectancy vs Total Income



Life Expectancy vs Total Expenditure



Life Expectancy vs HIV/AIDS



8. Modeling

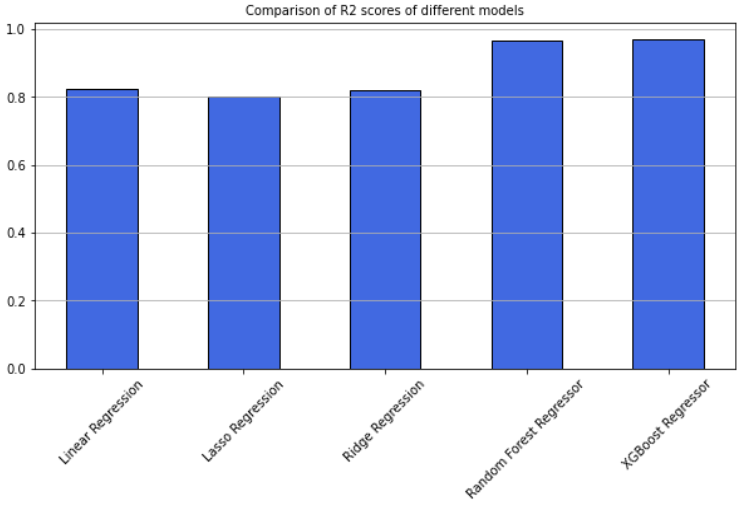
This is a regression problem in supervised learning. The following regression models were used:

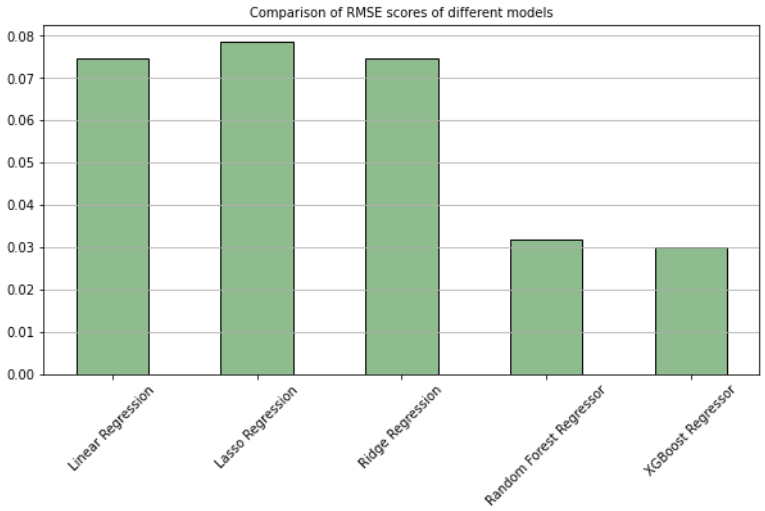
- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regressor
- XG Boost Regressor

The following is a table showing the comparison of the metrics between different models. For the regression metrics, the Root Mean Squared Error (RMSE) and R-squared (R2) were used. Both RMSE and R2 metrics are quite popular metrics and can be utilized to assess model performance. The RMSE metric is easily optimizable with most well-known algorithms. The smaller the RMSE score the better the model performance. The R2 metric is an equally popular metric. The higher the R2 score the better the model performance. It also makes the comparison between models easier and consistent.

Model Name	RMSE	R2
Linear Regression	0.074506	0.822111
Lasso Regression	0.078636	0.801845
Ridge Regression	0.074777	0.820814
Random Forest Regression	0.031716	0.967766
XG Boost Regression	0.030095	0.970976

Model Performance Graphs

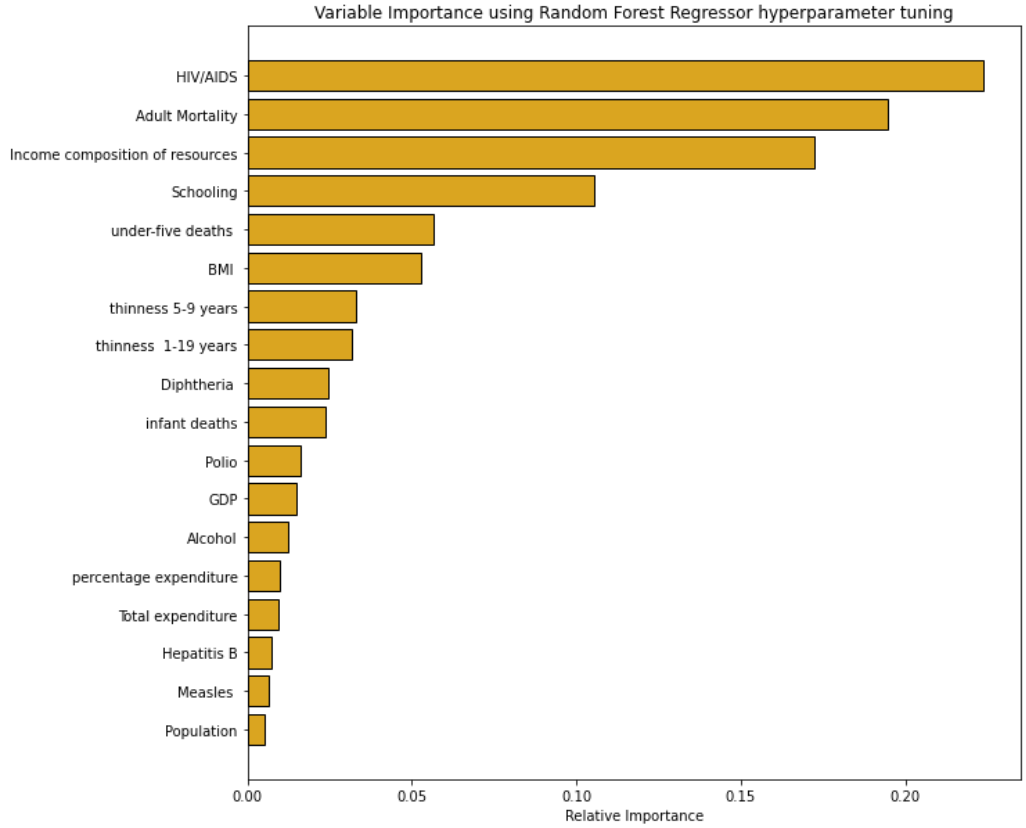


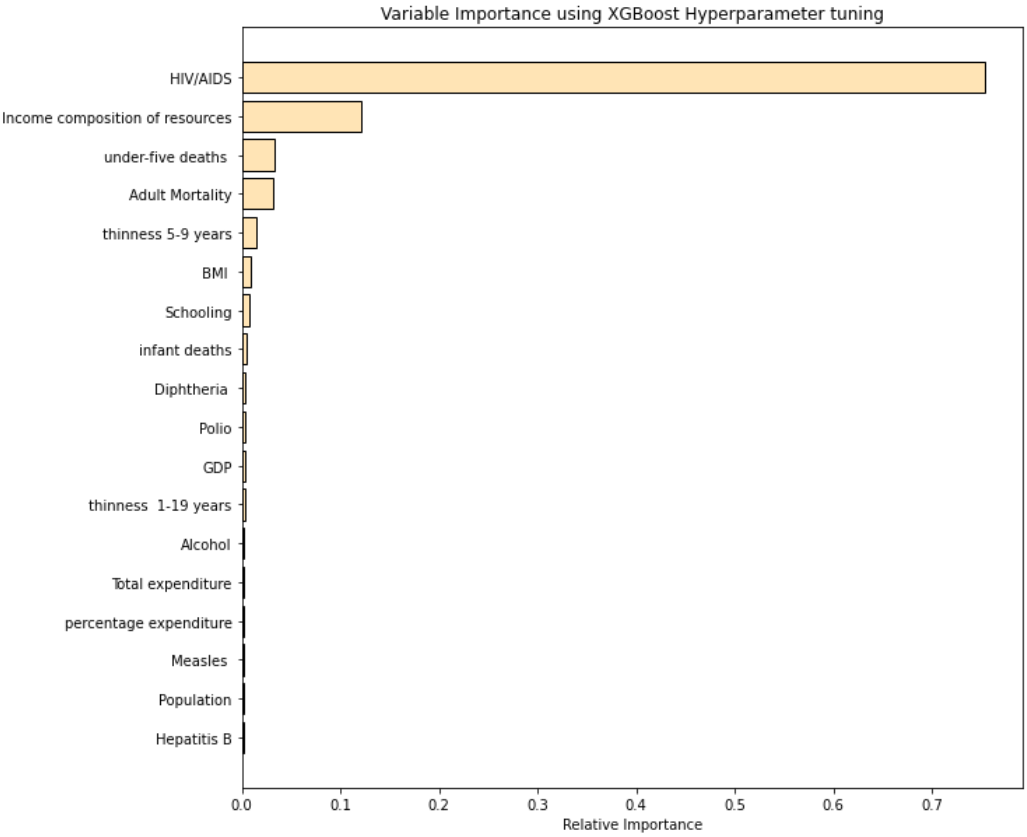


Checking the model performances with both R2 and RMSE scores, it is evident that both Random Forest Regressor and XG Boost Regressor were the best performing models and were selected for hyperparameter tuning and final model building.

9. Hyperparameter Tuning, Feature Selection and Final Modeling

As stated earlier, both the Random Forest model and the XG Boost model were selected for further hyperparameter tuning. The choice of a set of ideal hyperparameters for training a machine learning model is referred to as hyperparameter tuning in machine learning. Cross validation was performed to avoid any kind of overfitting. Both Randomized Search CV and GridSearchCV was used in this process of hyperparameter tuning. Two final models were built, and feature importance was calculated for each of the models.


Variable Importance using Random Forest Regressor hyperparameter tuning
HIV/AIDS
Adult Mortality
Income composition of resources
Schooling
under-five deaths
BMI
thinness 5-9 years
thinness 1-19 years
Diphtheria
infant deaths
Polio
GDP
Alcohol
percentage expenditure
Total expenditure
Hepatitis B
Measles
Population
0.00
0.05
0.10
0.15
0.20
Relative Importance


Variable Importance using XGBoost Hyperparameter tuning
HIV/AIDS
Income composition of resources
under-five deaths
Adult Mortality
thinness 5-9 years
BMI
Schooling
infant deaths
Diphtheria
Polio
GDP
thinness 1-19 years
Alcohol
Total expenditure
percentage expenditure
Measles
Population
Hepatitis B
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Relative Importance

10. Conclusion

Life expectancy continues to be a key metric to assess the health of the population in countries throughout the world. The federal health departments in different countries and the World Health Organization have a great stake in this and continues important downstream studies based on this statistic. The average life expectancy of a country can aid in evaluating different health related policies in all countries especially the developing countries. This study was done to assess the predictive factors affecting life expectancy. Five different modeling algorithms were performed on the dataset. After evaluating the metrics, it was determined that the Random Forest and the XG Boost models were the best performing models and were selected for hyperparameter tuning. Final models were built for both after hyperparameter tuning. In both the models, HIV/AIDS incidence was the topmost feature for predicting life expectancy. One of the main reasons behind this is because HIV/AIDS was still a pandemic in the early 2000s that reduced the life expectancy in most countries especially the developing nations. In the early 2000s, cures for the disease were sparse and expensive. This led to reduced life expectancy. Some of the other factors predicted to have a high relative importance by both models were adult mortality, income, schooling, infant mortality and childhood thinness and schooling.

11. Future Studies

Life expectancy figures are an important statistic with continuous source of data generation monitored by World Health Organization. This dataset only contains the various predictive factors from 2000-2015. From late 2019, we have endured through the COVID-19 pandemic which had killed more than 5 million people worldwide. A lot of people have also been suffering from long COVID which would reduce life expectancy by a great amount. Therefore, data related to COVID infection and vaccination status will be important in predicting future features. Vaccination status for other diseases would change predictive capabilities. With several other geo-political scenarios like the Russia-Ukraine war and its ancillary consequences, economic factors would be pivotal as predictive features in the new data.