DATA 119 Final Project

Predicting Life Expectancy

– by Anna Lévay and Will Yan –

Winter 2025

Overview and Motivation

The primary goal of our project is to identify the most significant predictors of Life Expectancy across countries over time. We were inspired by the UN 17 Sustainable Development Goals, the shared 2030 agenda for sustainable development adopted by all UN member states in 2015 (The 17 goals). Life expectancy is a key indicator of SDG 3 – Good Health and Wellbeing. While great advancements have been made since the adoption of the goals, progress on SDG 3 has been slowing year by year, and major challenges still remain in many regions around the world, as shown by Figure 1 (Sustainable Development Report).

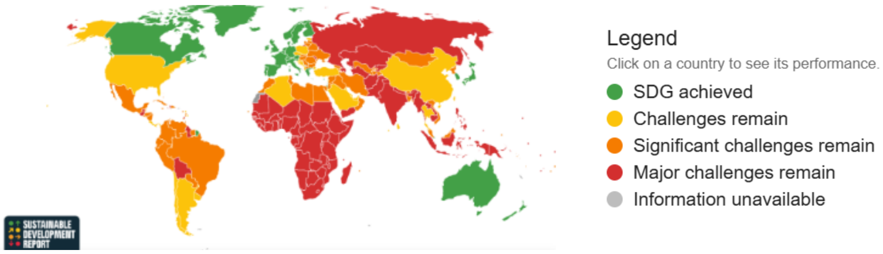

Legend
Click on a country to see its performance.
● SDG achieved
● Challenges remain
● Significant challenges remain
● Major challenges remain
● Information unavailable
SUSTAINABLE DEVELOPMENT REPORT

Figure 1: Sustainable Development Report - Life Expectancy at Birth (Years)

To achieve progress in these areas, it is crucial to understand what are the key drivers of life expectancy, and how they differ region by region. Our exploratory analysis illuminated the fact that life expectancy and several explanatory variables also differ significantly based on economic development status, thus our final research question is: **What are the key drivers of life expectancy, and how do they differ based on economic development status and geographical region?**

Data source, cleaning, and wrangling

We are using the Life Expectancy (World Health Organization) 2024 dataset found on Kaggle. The dataset originally includes health factors for 193 countries collected from the Global Health Observatory (GHO) data repository under the WHO, as well as economic data collected from the UN website, a total of 22 variables for the years 2000-2015.

Before starting the analysis, we had to handle missing values in the data: 1) We identified the columns containing missing values. 2) We assessed if those variables were likely to be important for the analysis, using a correlation matrix. We got rid of variables with very low correlation to Life Expectancy, as well as those with multicollinearity issues – Adult Mortality, Hepatitis B, Total Expenditure, and Income Composition of Resources. 3) We identified and dropped 16 countries for which information was consistently missing in significant variable columns. 4) Values for alcohol consumption per capita were missing for all countries in 2015. We imputed these using the mean alcohol consumption for each country across available years. 5) GDP, Population, and average number of years of Schooling were missing for 10-37 countries. We recovered these using external datasets from Kaggle, the World Bank, and Global Data Lab. Due to frequent mismatch between the precise way country names were written in different datasets, we manually created a dictionary that helped us match the missing values.

After eliminating all missing values, we were left with 179 countries, across 15 years, and 14 explanatory variables: **Schooling** (average number of years), **GDP** (per capita, in $), **BMI**, **Alcohol** (per capita consumption of pure alcohol among 15+ population, in liters), **Diphtheria** (Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)), **Polio** (Polio (Pol3) immunization coverage among 1-year-olds (%)), **HIV/AIDS** (Deaths per 1,000 live births due to HIV/AIDS (0-4 years)), **Measles** (Number of reported cases per 1000 population), **Under-five deaths** (per 1000 population), **Thinness 1-19 years** (Prevalence of thinness (low BMI) among children and adolescents 1–19 years old (%)), **Percentage expenditure** (Health expenditure as a share of total government expenditure per year (%)), **Year**, **Development Status** (binary, Developed or Developing), and **Population** (Total population). Our dependent variable is Life Expectancy at birth, expressed in number of years.

Exploratory Data Analysis and Feature Engineering

To better visualize the relationship among our numeric variables, we created a correlation matrix and then plotted it with a heatmap and annotated each cell with its correlation coefficient. We discover that Schooling exhibited the strongest positive correlation with Life Expectancy (0.78) and HIV/AIDS had the strongest negative correlation with Life Expectancy (-0.57). We also log-transformed GDP to address its skewed distribution and included this log_GDP feature in all subsequent analyses. After confirming that the dataset included both numeric and categorical features, we converted the country's development status into a dummy variable so that Developing status would be represented numerically.

Then, we created scatter plots for each selected feature against Life Expectancy. Although we produced many such scatter plots for all "interesting features", the two that stood out were: **1) Schooling vs. Life Expectancy (Figure 2).** A clear upward trend shows that higher average years of Schooling align with higher Life Expectancy. Developed countries (blue) cluster in the upper right, reflecting better education and longer lifespans, while developing countries (red) show a broader spread out but generally remain below the developed cluster. **2) HIV/AIDS vs. Life Expectancy (Figure 3).** A strong inverse relationship emerges, higher HIV/AIDS prevalence corresponds to lower life expectancy. Developing countries often appear toward the lower right (high HIV/AIDS, lower lifespans), whereas developed countries cluster at near-zero HIV/AIDS and higher life expectancy.

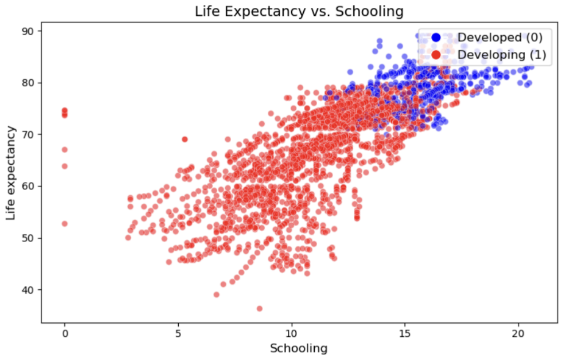


Figure 2: Life Expectancy vs. Schooling

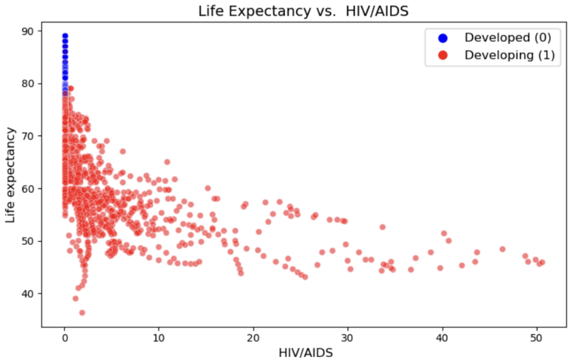


Figure 3: Life Expectancy vs. HIV/AIDS

Stage 1: Identify the Strongest Predictors of Life Expectancy

After assembling our dataset, we began our predictive modeling by fitting an ordinary linear regression on the cleaned dataset. This model achieved a mean $R^2 \approx 0.79$ with 5-fold cross-validation, and an $R^2 \approx 0.82$ on the test set, with average errors of about 3-4 years. This provides a good baseline, but the model cannot address the multicollinearity that arises when features (e.g., Population, Year) overlap. In addition, OLS assumes a strictly linear relationship, so it may struggle with nonlinear patterns of extreme borderline cases – such as countries with particularly high Population yet relatively low BMI, or those with anomalous Year and Schooling combinations. Due to these limitations, we next explored Ridge and Lasso, which penalize large coefficients to improve stability.

After applying Ridge and Lasso regressions to mitigate multicollinearity and shrink less important features, the models still yielded roughly the same R^2 as our baseline linear regression, around 0.79–0.82 overall. While both methods confirmed Schooling as a highly positive variable (coefficient near 4.2–4.3) and HIV/AIDS as strongly negative (–3.5), they did not significantly increase the model's predictive accuracy. Consequently, we moved to Random Forest to better account for these nonlinear relationships.

After tuning key hyperparameters via GridSearchCV, our Random Forest substantially improved predictive accuracy, attaining a final test-set $R^2 \approx 0.96$. This model consistently captured nonlinear relationships among features, showing a near-perfect diagonal in the plot **(Figure 4)**. Examining feature importances revealed HIV/AIDS as the most important factor (important ≈ 0.63), followed by Schooling (≈ 0.17). Factors such as Thinness 1-19 years, BMI, and Under-five deaths also contributed notably **(Table 1)**. As a result, the Random Forest outperformed linear methods and underscored both health and socioeconomic factors as key drivers of Life Expectancy worldwide.

Feature Name	Feature Importance
HIV/AIDS	0.6308
Schooling	0.1708
Thinness 1-19 years	0.0437
BMI	0.0337
Under-five deaths	0.0282
Alcohol	0.0172
Year	0.0154
Log_GDP	0.0134
Percentage Expenditure	0.0117
Diphtheria	0.0088
Population	0.0084
Polio	0.0071
Measles	0.0066
Status_Developing	0.0042

Table 1: Most Important Features for Predicting Life Expectancy

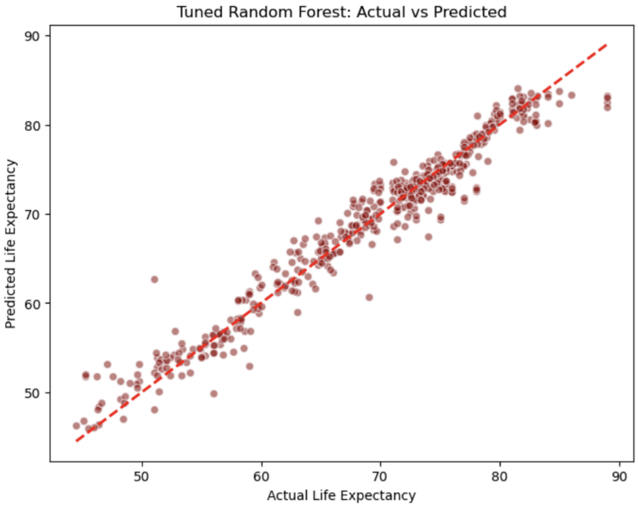


Figure 4: Tuned Random Forest Model: Actual vs. Predicted

Stage 2: Developed vs. Developing Countries

After dividing our data into Developed and Developing subsets, we tested Ridge and Lasso, but the results were still underwhelming, with $R^2 \approx 0.61$ in developed countries and $R^2 \approx 0.75$ in developing ones.

So, we moved on to Random Forest, the model performed substantially better, attaining $R^2 \approx 0.80$ for developed countries and $R^2 \approx 0.94$ for developing nations.

In developed regions, thinness 1-19 years emerged as the most influential feature (≈ 0.38 importance), which is somewhat surprising since thinness is typically linked to malnutrition in lower-income contexts **(Figure 5)**. Here it may reflect different health patterns (e.g., eating disorders) or demographic factors. Alcohol and year also played significant roles, but the model left about 20% of variability unexplained–perhaps due to lifestyle diversity in wealthier nations. Conversely, in developing countries, HIV/AIDS dominated (importance ≈ 0.64), dwarfing even Schooling and BMI **(Figure 6)**. This shows how infectious diseases can severely impact longevity.

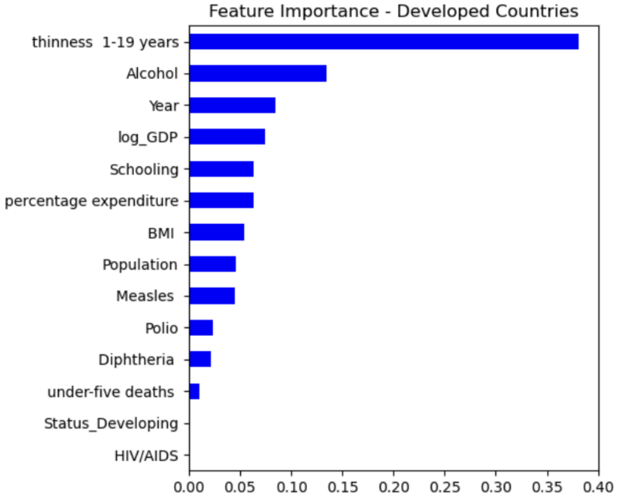


Figure 5: Feature Importance – Developed Countries

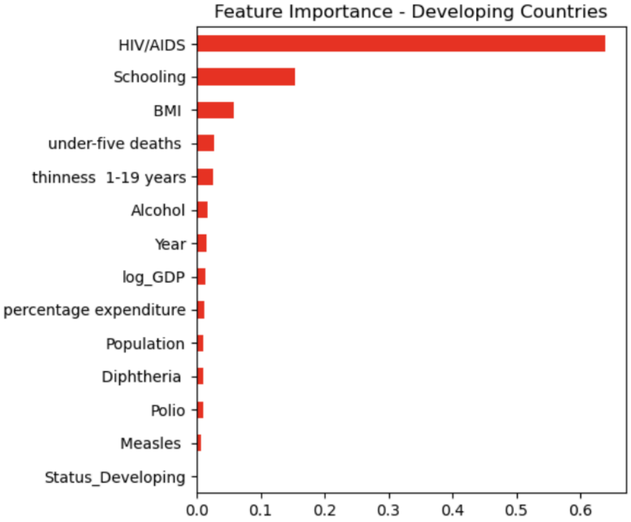
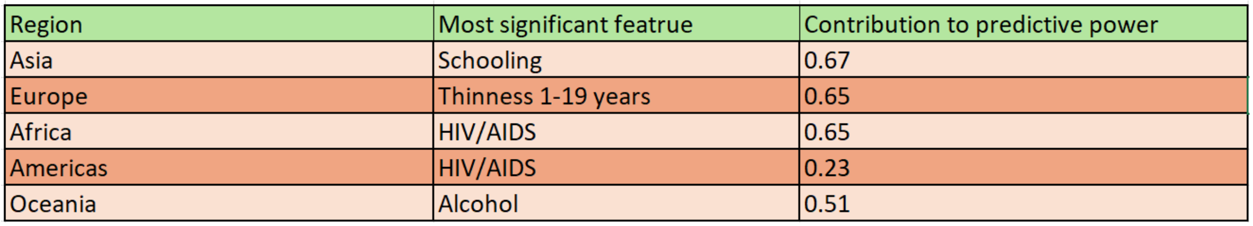


Figure 6: Feature Importance – Developing Countries

Stage 3: Regional Differences

In Stage 3 of the analysis we studied regional differences between the relative importance of explanatory variables in predicting life expectancy. We created a region mapping dictionary to map each country in our dataset to one of 5 regions based on continents: Africa, Americas, Asia, Europe, Oceania. For each region we split the data into train and test sets, scaled numeric variables using train set statistics, then trained 3 models: Ridge, Lasso, and Random Forest, and performed cross validation to evaluate them.

For each continent, the Random Forest model achieved the highest cross-validation R^2 value (Asia: 0.92, Europe: 0.82, Africa: 0.93, Americas: 0.87, Oceania: 0.93) and the lowest error terms. The model for Europe has the weakest explanatory power, but each model across regions tends to generalize well and has a high performance on the test sets. Table 2 shows key regional differences in significant features:



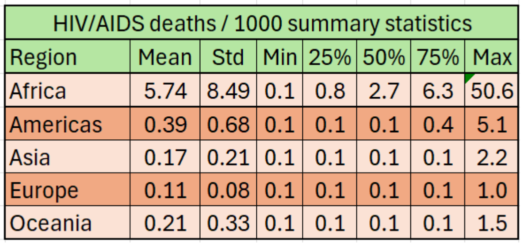
Region	Most significant featrue	Contribution to predictive power
Asia	Schooling	0.67
Europe	Thinness 1-19 years	0.65
Africa	HIV/AIDS	0.65
Americas	HIV/AIDS	0.23
Oceania	Alcohol	0.51

Table 2: Most Significant Features and their Contribution to Predictive Power by Region

In all regions except the Americas, a single variable seems to overwhelmingly influence Life Expectancy predictions. This variable differs for each of these 4 regions, highlighting the heterogeneity across geographical areas. In contrast, in the Americas, the most significant feature only had a 23% contribution to the model's overall predictive power, indicating a lower generalizability of Life Expectancy drivers in this region, perhaps due to the significant differences between North and South American countries.

Some of these findings are easily justifiable. For example, it is reasonable that HIV/AIDS emerged as the most significant feature in Africa, since this variable has the highest negative correlation with Life Expectancy, and the mean, range, and standard deviation for it is by far the largest in Africa (Table 3).

Others, such as Thinness in Europe or Alcohol consumption in Oceania are much harder to explain intuitively, and would require further analysis.



HIV/AIDS deaths / 1000 summary statistics							
Region	Mean	Std	Min	25%	50%	75%	Max
Africa	5.74	8.49	0.1	0.8	2.7	6.3	50.6
Americas	0.39	0.68	0.1	0.1	0.1	0.4	5.1
Asia	0.17	0.21	0.1	0.1	0.1	0.1	2.2
Europe	0.11	0.08	0.1	0.1	0.1	0.1	1.0
Oceania	0.21	0.33	0.1	0.1	0.1	0.1	1.5

Table 3: Summary statistics of HIV/AIDS variable by region

Limitations and Implications

Through our analysis we managed to show that increasing life expectancy requires differentiated action, targeting the most significant features by region and according to the level of economic development of the country. Concretely, we can conclude that out of the sub targets of SDG 3, target 3.3: "End the epidemics of AIDS (...)" is crucially important to increase life expectancy, especially in Africa (The 17 goals).

Some limitations constrain our analysis. Firstly, our data cleaning method eliminated 16 countries, all microstates, small island and island nations, and countries that received independence in the past 20 years or have unique sovereignty status. Thus they face special circumstances which could have impacted our analysis, but were left unaccounted for. Secondly, grouping countries into regions produced unbalanced groups, Africa – 800 data points, Oceania – 160, making it harder to compare model performance across regions, since models trained on more data may perform inherently better. Lastly, there might be more variables significant for predicting Life Expectancy, such as environmental factors (air pollution, access to clear water) which our analysis did not include.

Further research could be done with a larger set of explanatory variables and across a more extensive period of time. A time series analysis could be used to study how Life Expectancy has evolved over time and what variable changes influenced it the most.

Member contributions: Coding: All together. Write-up: Anna Lévay – Overview and Motivation, Data source and cleaning, Stage 3, Implications and limitations, Will Yan – Exploratory Analysis and Feature Engineering, Stage 1, Stage 2. Slides: Same content distribution as report.

Works Cited

"The 17 goals | sustainable development." *United Nations*, https://sdgs.un.org/goals. Accessed 06 March 2025.

Global Data Lab. "Subnational HDI (V8.1)." *Global Data Lab*, https://globaldatalab.org/shdi/table/esch/?levels=1&interpolation=0&extrapolation=0. Accessed 06 March 2025.

Ismail, Hina. "Life expectancy (world health organization) 2024." *Kaggle*, 29 May 2024, https://www.kaggle.com/datasets/sonialikhan/life-expectancy-who-2024. Accessed 6 March 2025.

Nitisha. "GDP per capita all countries." *Kaggle*, 28 April 2020, https://www.kaggle.com/datasets/nitishabharathi/gdp-per-capita-all-countries?select=GDP.csv. Accessed 6 March 2025.

"Population, total | Data." *World Bank Open Data*, https://data.worldbank.org/indicator/SP.POP.TOTL. Accessed 6 March 2025.

"Sustainable Development Report 2024." *Sustainable Development Report 2024*, https://dashboards.sdgindex.org/map/indicators/life-expectancy-at-birth. Accessed 6 March 2025.