3 GOOD HEALTH AND WELL-BEING

# Predicting Life Expectancy

Winter 2025

**DATA - 119 - final project**

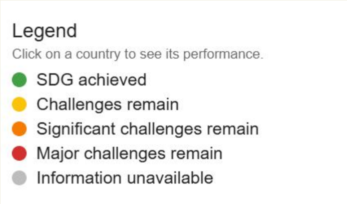
Presented By

**Anna Lévay, Will Yan**

# Report Summary

| | |
|---|---|
| **KEY QUESTIONS** | Stage 1: What are the most important predictors of life expectancy?<br>Stage 2 & 3: How do these differ for developed/developing countries, and across regions? |
| **DATA** | Life Expectancy (WHO) data from Kaggle<br>Recovering missing values from outside data sources |
| **METHODOLOGY** | Comparing Multiple Linear Regression, Ridge, Lasso, Random Forest models<br>Selecting the best model through cross validation, hyperparameter tuning |
| **MAJOR FINDINGS** | Developed countries: model has low explanatory power<br>Developing countries: most important variables are HIV/AIDS and Schooling<br>Regions: Africa and Americas: HIV, Asia: Schooling, Oceania: Alcohol |
| **IMPLICATIONS, LIMITATIONS** | Data cleaning method excludes small island states and micronations<br>Limited scope of explanatory variables: maternal mortality, environmental factors are not included |

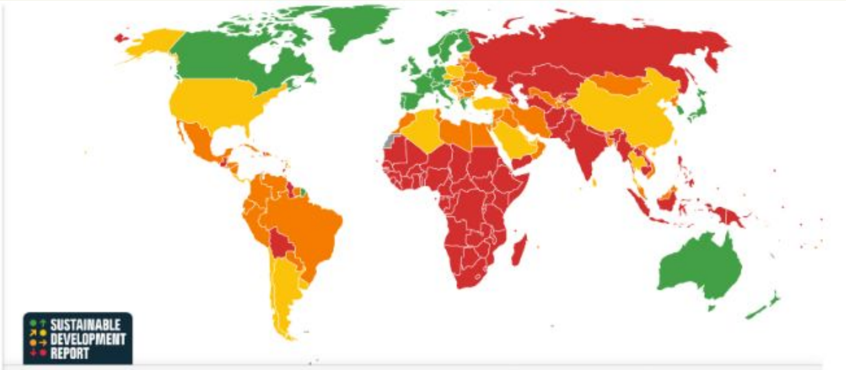
THE GLOBAL GOALS



Legend

Click on a country to see its performance.

- 🟢 SDG achieved
- 🟡 Challenges remain
- 🟠 Significant challenges remain
- 🔴 Major challenges remain
- ⚪ Information unavailable



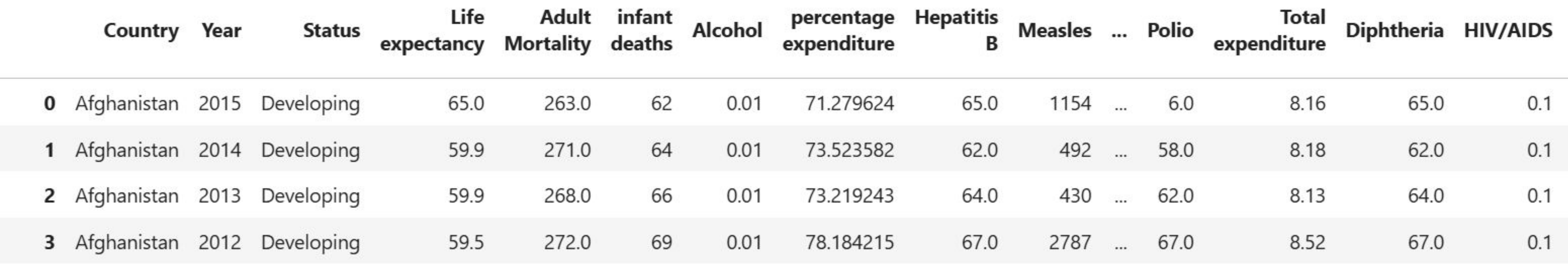
[Sustainable Development Report ](#)- Life Expectancy at Birth (Years)

# Motivation - SDG3

“Ensure healthy lives and promote well-being for all at all ages”

- Slowing progress since 2015
- Significant regional differences
- Different subgoals – which ones to focus on?

# Data source, data cleaning



| | Country | Year | Status | Life expectancy | Adult Mortality | infant deaths | Alcohol | percentage expenditure | Hepatitis B | Measles | ... | Polio | Total expenditure | Diphtheria | HIV/AIDS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Afghanistan | 2015 | Developing | 65.0 | 263.0 | 62 | 0.01 | 71.279624 | 65.0 | 1154 | ... | 6.0 | 8.16 | 65.0 | 0.1 |
| **1** | Afghanistan | 2014 | Developing | 59.9 | 271.0 | 64 | 0.01 | 73.523582 | 62.0 | 492 | ... | 58.0 | 8.18 | 62.0 | 0.1 |
| **2** | Afghanistan | 2013 | Developing | 59.9 | 268.0 | 66 | 0.01 | 73.219243 | 64.0 | 430 | ... | 62.0 | 8.13 | 64.0 | 0.1 |
| **3** | Afghanistan | 2012 | Developing | 59.5 | 272.0 | 69 | 0.01 | 78.184215 | 67.0 | 2787 | ... | 67.0 | 8.52 | 67.0 | 0.1 |

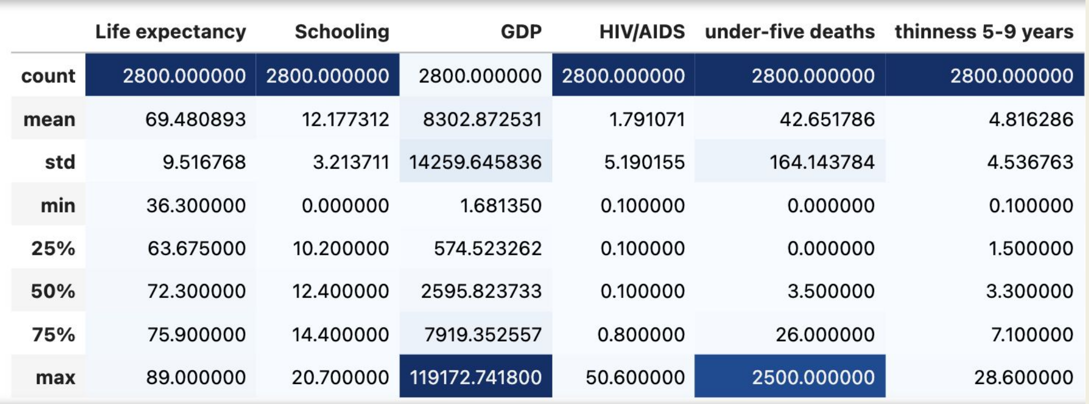
Data overview

- Source: Kaggle, WHO, UN
- Life Expectancy, Economic, Demographic, Health indicators
- Initial: 193 countries, 22 variables, 15 years
- Final: 179 countries, 14 variables

Cleaning process – Handling Missing Values

- Dropped insignificant columns
- Excluded 16 countries – microstates, small island nations
- Recovered missing values from external data sources

# Summary Statistics



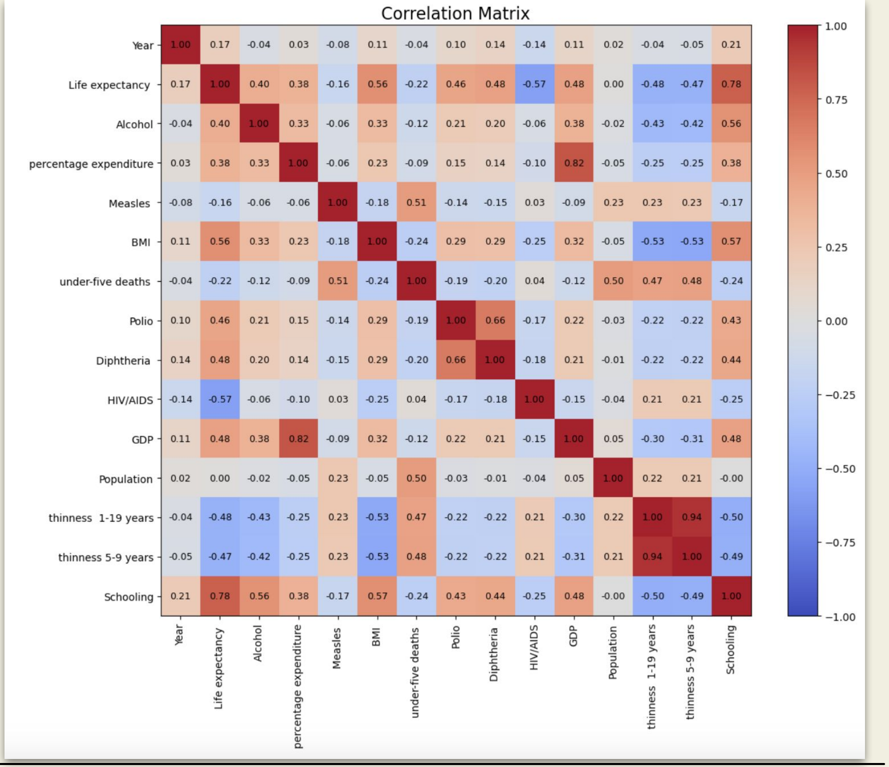
| | Life expectancy | Schooling | GDP | HIV/AIDS | under-five deaths | thinness 5-9 years |
|---|---|---|---|---|---|---|
| **count** | 2800.000000 | 2800.000000 | 2800.000000 | 2800.000000 | 2800.000000 | 2800.000000 |
| **mean** | 69.480893 | 12.177312 | 8302.872531 | 1.791071 | 42.651786 | 4.816286 |
| **std** | 9.516768 | 3.213711 | 14259.645836 | 5.190155 | 164.143784 | 4.536763 |
| **min** | 36.300000 | 0.000000 | 1.681350 | 0.100000 | 0.000000 | 0.100000 |
| **25%** | 63.675000 | 10.200000 | 574.523262 | 0.100000 | 0.000000 | 1.500000 |
| **50%** | 72.300000 | 12.400000 | 2595.823733 | 0.100000 | 3.500000 | 3.300000 |
| **75%** | 75.900000 | 14.400000 | 7919.352557 | 0.800000 | 26.000000 | 7.100000 |
| **max** | 89.000000 | 20.700000 | 119172.741800 | 50.600000 | 2500.000000 | 28.600000 |

Only a summary of some variables is displayed here

- **Life Expectancy**: Mean **69.48 years**, Range **36.3 – 89 years**
- **Schooling**: Mean **12.18 years**, higher in developed countries
- **GDP**: Highly skewed, log transformation applied
- **HIV/AIDS**: Wide variation (**0.1 – 50.6**), major impact on life expectancy
- **Under-Five Deaths**: High disparity between developed & developing countries
- **Thinness (5-9 years)**: Negatively correlated with life expectancy

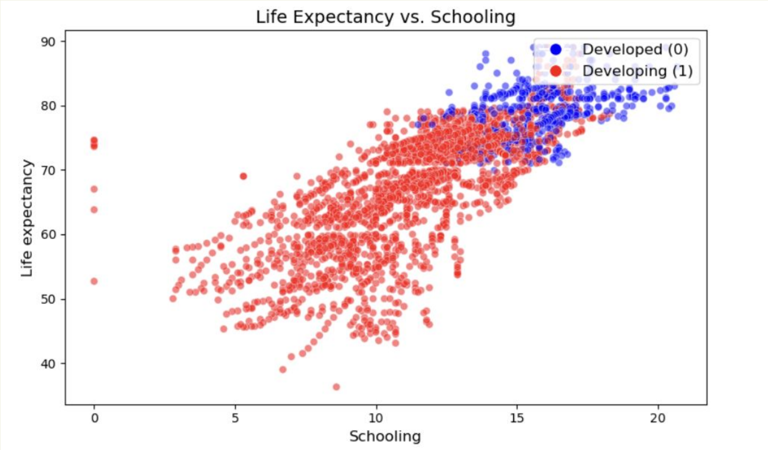
# Correlation Matrix

- **Highly correlated variables** with Life Expectancy:
  a. **Positive correlation:** Schooling (0.78), BMI (0.56), GDP (0.48).
  b. **Negative correlation:** HIV/AIDS (-0.57), Thinness 1-19 years (-0.48), Under-five deaths (-0.22).
- **Surprising low correlation:** Percentage expenditure (0.38) → **health spending alone does not directly predict Life Expectancy.**


Correlation Matrix

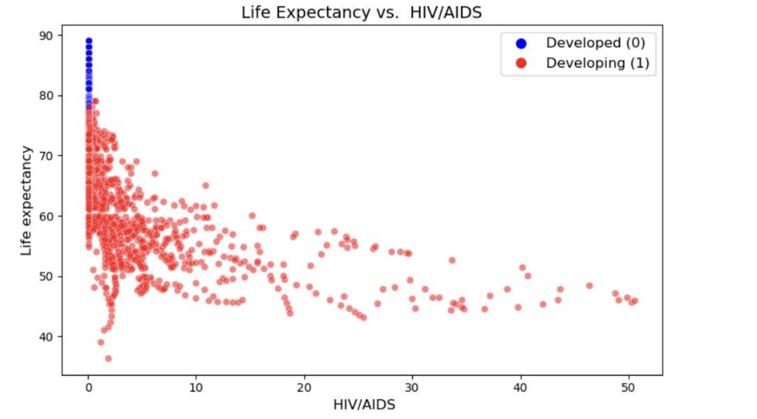
| | Year | Life expectancy | Alcohol | percentage expenditure | Measles | BMI | under-five deaths | Polio | Diphtheria | HIV/AIDS | GDP | Population | thinness 1-19 years | thinness 5-9 years | Schooling |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | 1.00 | 0.17 | -0.04 | 0.03 | -0.08 | 0.11 | -0.04 | 0.10 | 0.14 | -0.14 | 0.11 | 0.02 | -0.04 | -0.05 | 0.21 |
| Life expectancy | 0.17 | 1.00 | 0.40 | 0.38 | -0.16 | 0.56 | -0.22 | 0.46 | 0.48 | -0.57 | 0.48 | 0.00 | -0.48 | -0.47 | 0.78 |
| Alcohol | -0.04 | 0.40 | 1.00 | 0.33 | -0.06 | 0.33 | -0.12 | 0.21 | 0.20 | -0.06 | 0.38 | -0.02 | -0.43 | -0.42 | 0.56 |
| percentage expenditure | 0.03 | 0.38 | 0.33 | 1.00 | -0.06 | 0.23 | -0.09 | 0.15 | 0.14 | -0.10 | 0.82 | -0.05 | -0.25 | -0.25 | 0.38 |
| Measles | -0.08 | -0.16 | -0.06 | -0.06 | 1.00 | -0.18 | 0.51 | -0.14 | -0.15 | 0.03 | -0.09 | 0.23 | 0.23 | 0.23 | -0.17 |
| BMI | 0.11 | 0.56 | 0.33 | 0.23 | -0.18 | 1.00 | -0.24 | 0.29 | 0.29 | -0.25 | 0.32 | -0.05 | -0.53 | -0.53 | 0.57 |
| under-five deaths | -0.04 | -0.22 | -0.12 | -0.09 | 0.51 | -0.24 | 1.00 | -0.19 | -0.20 | 0.04 | -0.12 | 0.50 | 0.47 | 0.48 | -0.24 |
| Polio | 0.10 | 0.46 | 0.21 | 0.15 | -0.14 | 0.29 | -0.19 | 1.00 | 0.66 | -0.17 | 0.22 | -0.03 | -0.22 | -0.22 | 0.43 |
| Diphtheria | 0.14 | 0.48 | 0.20 | 0.14 | -0.15 | 0.29 | -0.20 | 0.66 | 1.00 | -0.18 | 0.21 | -0.01 | -0.22 | -0.22 | 0.44 |
| HIV/AIDS | -0.14 | -0.57 | -0.06 | -0.10 | 0.03 | -0.25 | 0.04 | -0.17 | -0.18 | 1.00 | -0.15 | -0.04 | 0.21 | 0.21 | -0.25 |
| GDP | 0.11 | 0.48 | 0.38 | 0.82 | -0.09 | 0.32 | -0.12 | 0.22 | 0.21 | -0.15 | 1.00 | 0.05 | -0.30 | -0.31 | 0.48 |
| Population | 0.02 | 0.00 | -0.02 | -0.05 | 0.23 | -0.05 | 0.50 | -0.03 | -0.01 | -0.04 | 0.05 | 1.00 | 0.22 | 0.21 | -0.00 |
| thinness 1-19 years | -0.04 | -0.48 | -0.43 | -0.25 | 0.23 | -0.53 | 0.47 | -0.22 | -0.22 | 0.21 | -0.30 | 0.22 | 1.00 | 0.94 | -0.50 |
| thinness 5-9 years | -0.05 | -0.47 | -0.42 | -0.25 | 0.23 | -0.53 | 0.48 | -0.22 | -0.22 | 0.21 | -0.31 | 0.21 | 0.94 | 1.00 | -0.49 |
| Schooling | 0.21 | 0.78 | 0.56 | 0.38 | -0.17 | 0.57 | -0.24 | 0.43 | 0.44 | -0.25 | 0.48 | -0.00 | -0.50 | -0.49 | 1.00 |

# Life Expectancy vs. Schooling



- **Higher education = Longer life expectancy**
  a. **Developed Countries** → Generally **higher schooling (10+ years) and higher life expectancy (~75-90 years).**
  b. **Developing Countries** → More spread out, with many having **low schooling (<8 years) and lower life expectancy (<70 years).**

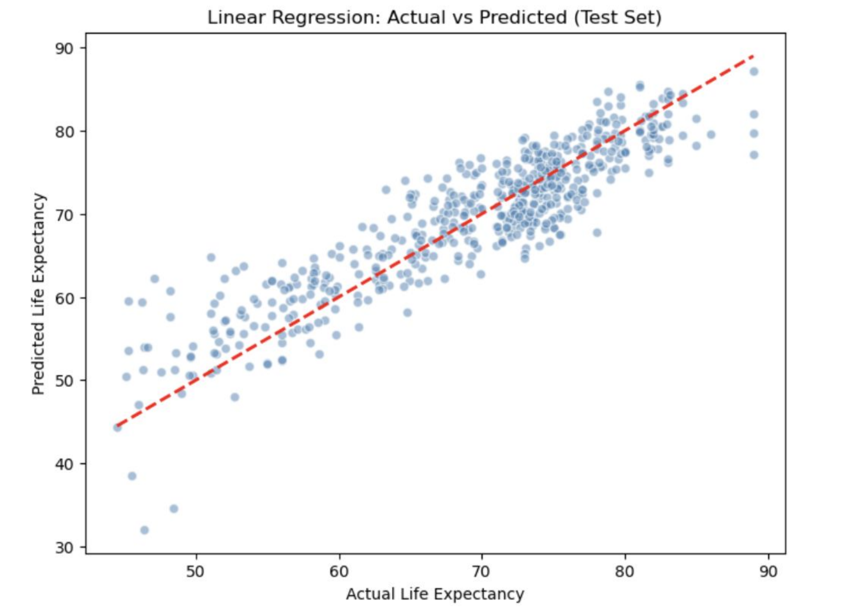
# Life Expectancy vs. HIV/AID



- **Higher HIV/AIDS rate = Lower life expectancy**
  a. **Developed Countries** → Almost all have **near-zero HIV/AIDS rates** and **higher life expectancy (~80+ years)**.
  b. **Developing Countries** → Many have **HIV/AIDS rates >10%** and **lower life expectancy (<60 years)**.
  c. **Extreme Cases** → Countries with **HIV/AIDS rates above 30% show drastic reductions in life expectancy (~40-50 years)**.

# Model Building: Predicting Life Expectancy

**Goal: Identify the strongest predictors of Life Expectancy and find the most accurate model.**



Linear Regression (Actual vs. Predicted):

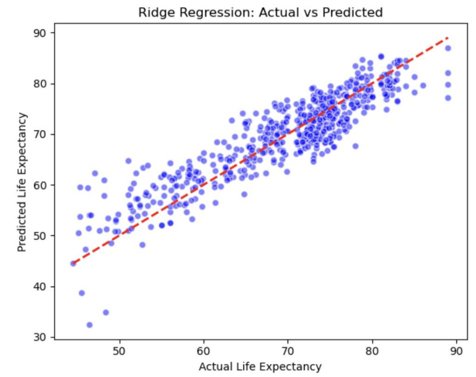
- Strong linear trend, but clear dispersion around the regression line.
- Prediction errors increase for extreme values.
- High variance suggests the need for better model (Lasso & Ridge).

Evaluation:

- MAE: 3.14
- RMSE: 4.01
- $R^2$ Score: 0.8175

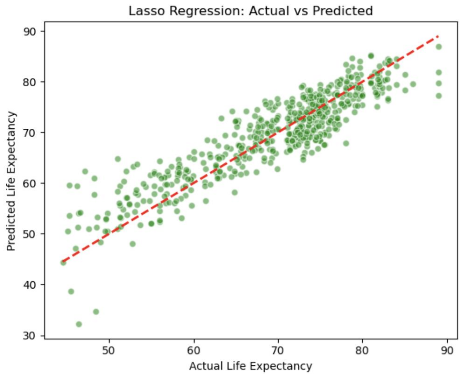
# Lasso & Ridge

## Why: Handling Multicollinearity



Ridge:

- R² = 0.8167, slightly worse than before
- Retains all features, but does not significantly improve predictions.
- Education (Schooling = 4.19) & Health (HIV/AIDS = -3.55) remain top predictors.

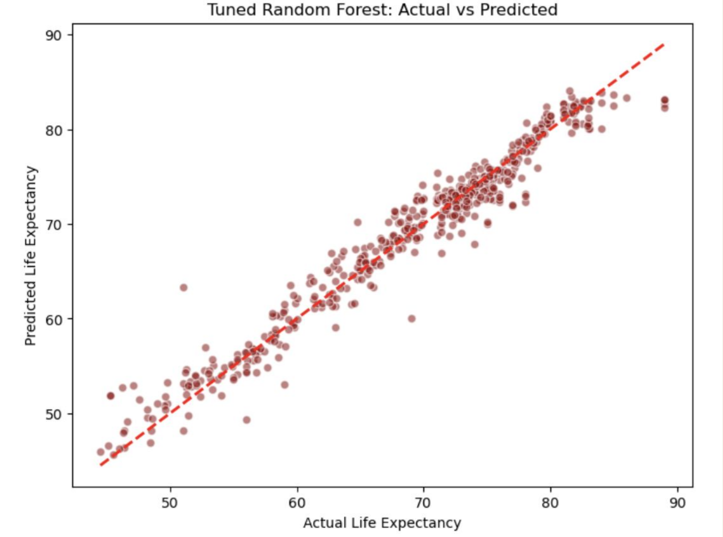


Lasso:

- R² = 0.8168 → no improvement
- Removes weaker predictors, but does not increase accuracy
- Non-linearity exists in the data → more complex model

# Random Forest

**Why: Captures non-linear relationship in Life Expectancy**



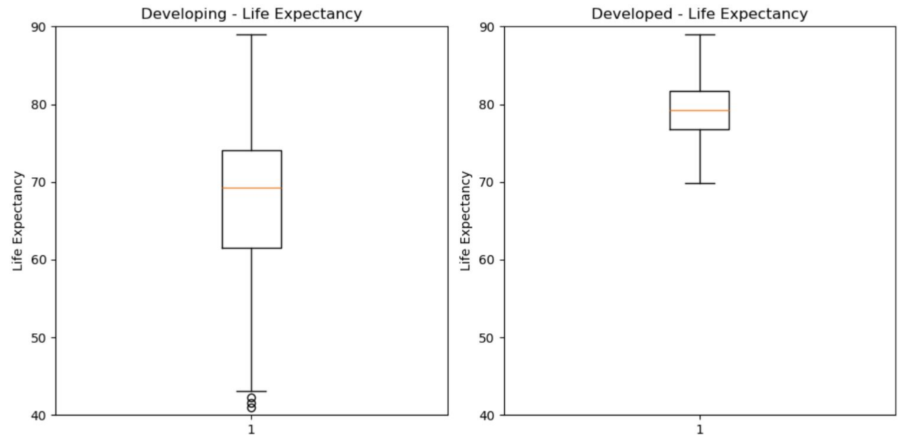
Evaluation:

- Random Forest significantly outperforms all other models ($R^2 = 0.985$).

So what drives Life Expectancy?

- Result:
  - HIV/AIDS (0.6308) → lower life expectancy
  - Schooling (0.1704) → More schooling, better healthcare knowledge, economic opportunities, and access to medical care.
  - Thinness 1-19 years (0.0448) and BMI (0.0336)
  - log_GDP (0.0132) has lower importance than expected.

# Stage 2: Life Expectancy in Developed vs. Developing Countries



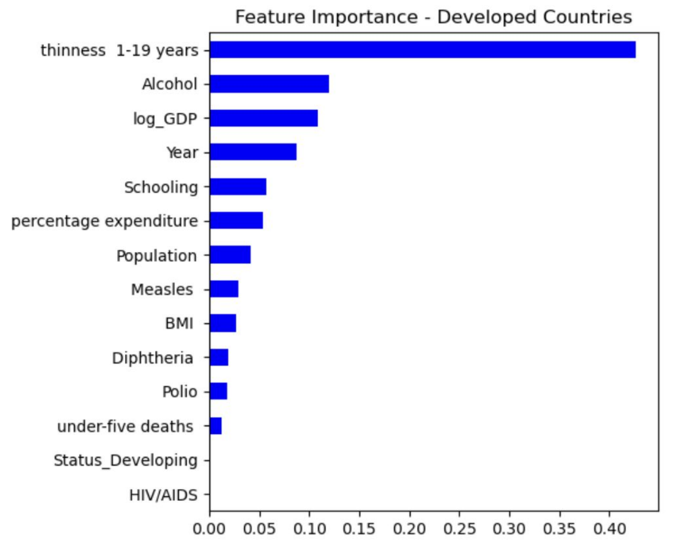
Developing

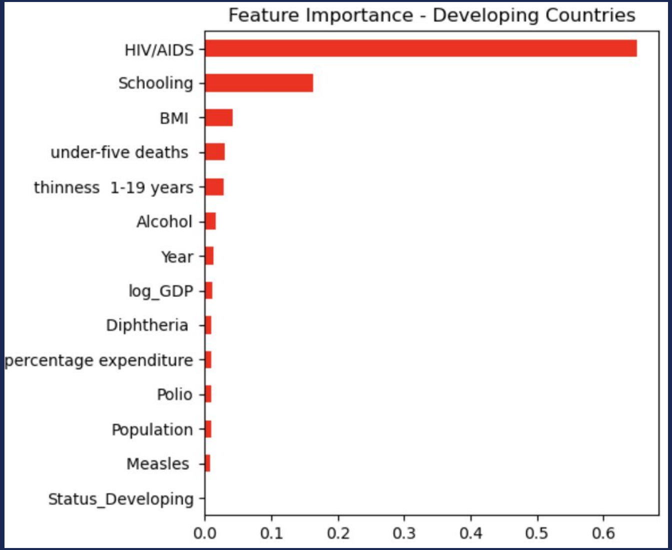
- Wider distribution & more lower-end outliers
  - Many countries below 60 years

Developed

- Higher median & less variation
  - Median life expectancy (~80 years).

# Key Predictors for Developed vs. Developing Countries



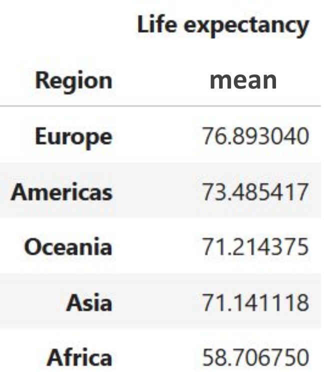


- Model: Random Forest performs best in both groups (R²: 0.68 vs. 0.93).
- Life Expectancy in Developing nations is more predictable
- HIV/AIDS is the strongest negative predictor in developing countries.
- Question the thinness in developed countries

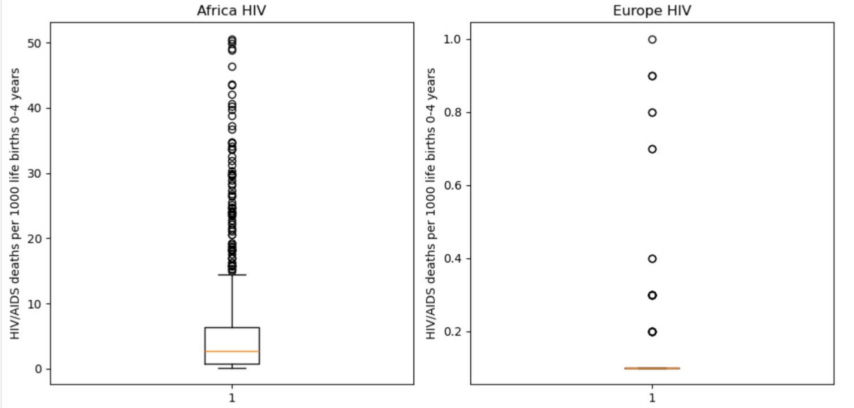
# Stage 3: Life Expectancy across geographical regions

Literature showed differences in life expectancy across continents

Our data confirms that



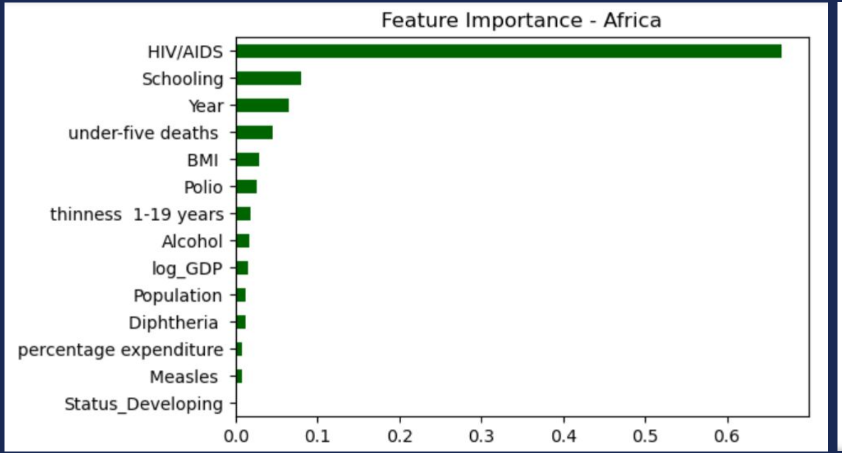
|  | Life expectancy |
|---|---|
| Region | mean |
| Europe | 76.893040 |
| Americas | 73.485417 |
| Oceania | 71.214375 |
| Asia | 71.141118 |
| Africa | 58.706750 |

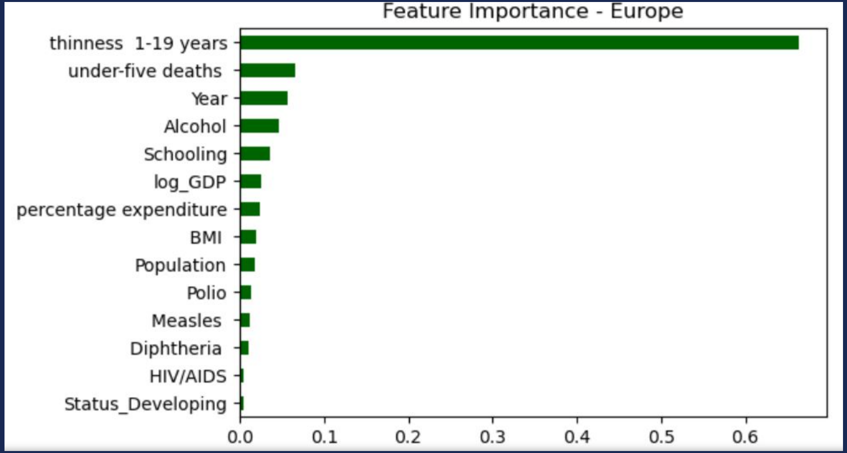


Data shows some variables have different distributions across regions

Question: **Is the difference in life expectancy within a continent explained primarily by the same variables across continents, or do drivers differ by continent?**

# Stage 3: Life Expectancy across geographical regions





Method: **Random Forest model** emerged as the best method for all continents

Result: **Most significant features vary a lot by continent**

- Africa and Americas: HIV/AIDS
- Asia: Schooling
- Oceania: Alcohol
- Europe: Thinness 1-19 years

# Conclusion, limitations

Implications:

- Increasing life expectancy world wide requires differentiated action targeting the most significant features in each region and according to the level of economic development in a given country
- SDG target 3.3: End the epidemics of AIDS is crucially important to increase life expectancy, especially in Africa and the Americas

Limitations:

- Data cleaning method excludes small island states and micronations
- Limited scope of explanatory variables: maternal mortality, environmental factors are not included
- Some findings are hard to explain intuitively

**BEST PREDICTIVE MODEL**

**Random Forest:**

- Non-linear relationships in the data
- $R^2$=0.958

**DEVELOPED / DEVELOPING**

**Key differences:**

- Developed: regressions have little explanatory power
- Developing: HIV, Schooling have the most impact on predictions

**REGIONAL DIFFERENCES**

**Top variables differ:**

- Africa and Americas: HIV/AIDS
- Asia: Schooling
- Oceania: Alcohol
- Europe: Thinness 1-19 years