Developing Models to Predict Life Expectancy Using Data from the WHO

Chad Bradford
Peyton Lambourne
Anshi Mathur
James Segovia

GROUP FOUR

PROJECT 2 PRESENTATION

March 31, 2025

OUR GOALS:

Our Group Project 2 focuses on harnessing our acquired skill set to assess and clean a rich data set that enables a build of strong predictive models for life expectancy.

Hypothesis: Health, social, and/or economic factors can be used to model life expectancy outcomes with high predictability.

01

IDENTIFY AN INFORMING DATA SET

- Relevance
- Data set quality
- Size and balance
- Feature diversity (without leakage)
- Current
- Ethics

02

ASSESS AND CLEAN THE DATA SET

- Routine cleaning
- Identified and dropped leaking features
- Used lag features to model temporal trends
- Imputed missing values

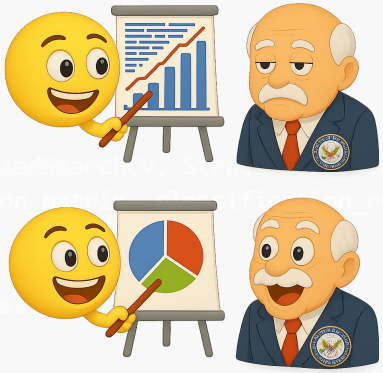
03

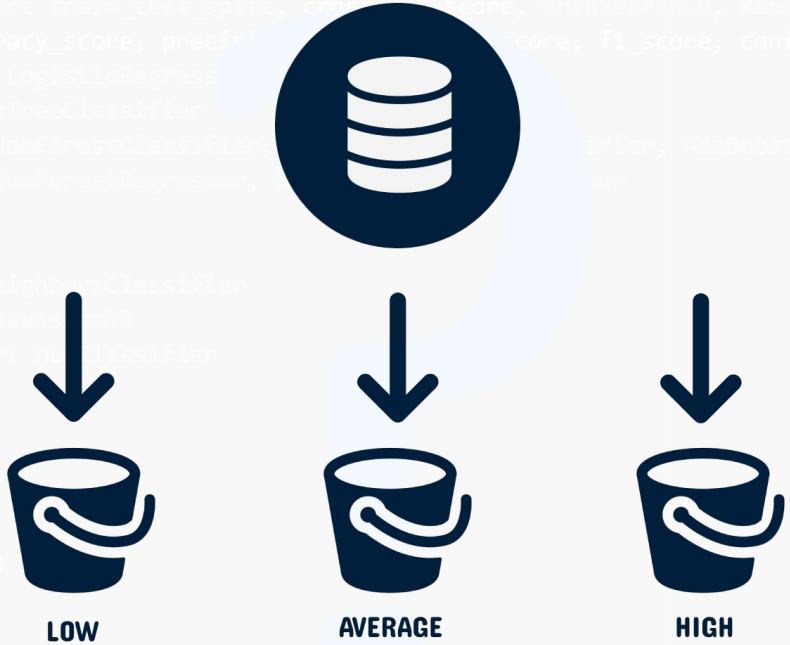
MODEL BUILDING AND TESTING

- Model build and testing
  - Classification
  - Regression

Classify Life Expectancy

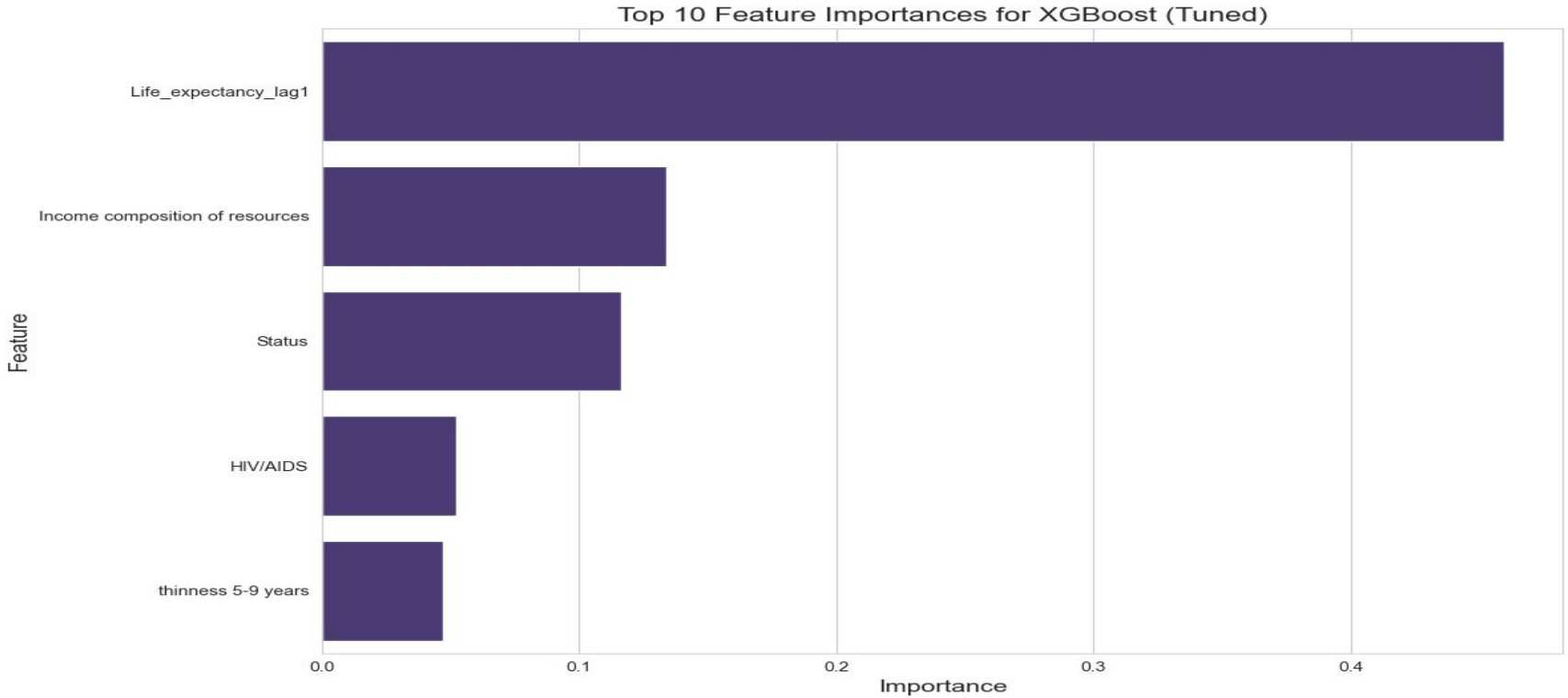
Let's get the boring stuff out of the way





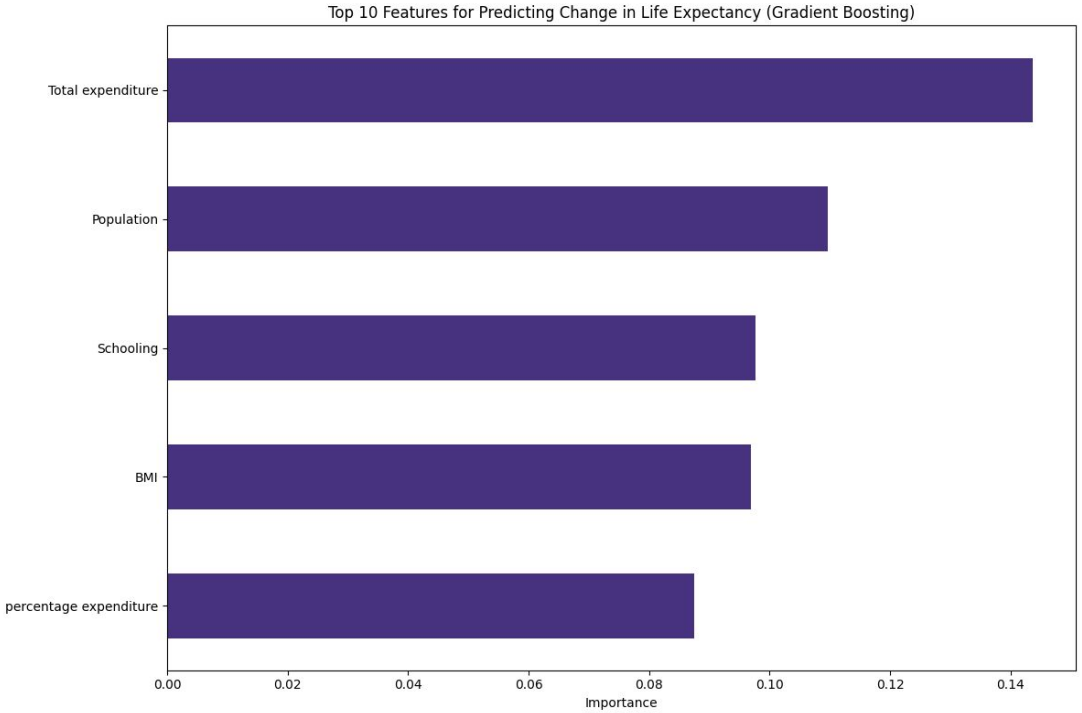
The graph made me do it…

Wow!! Who could have predicted this?



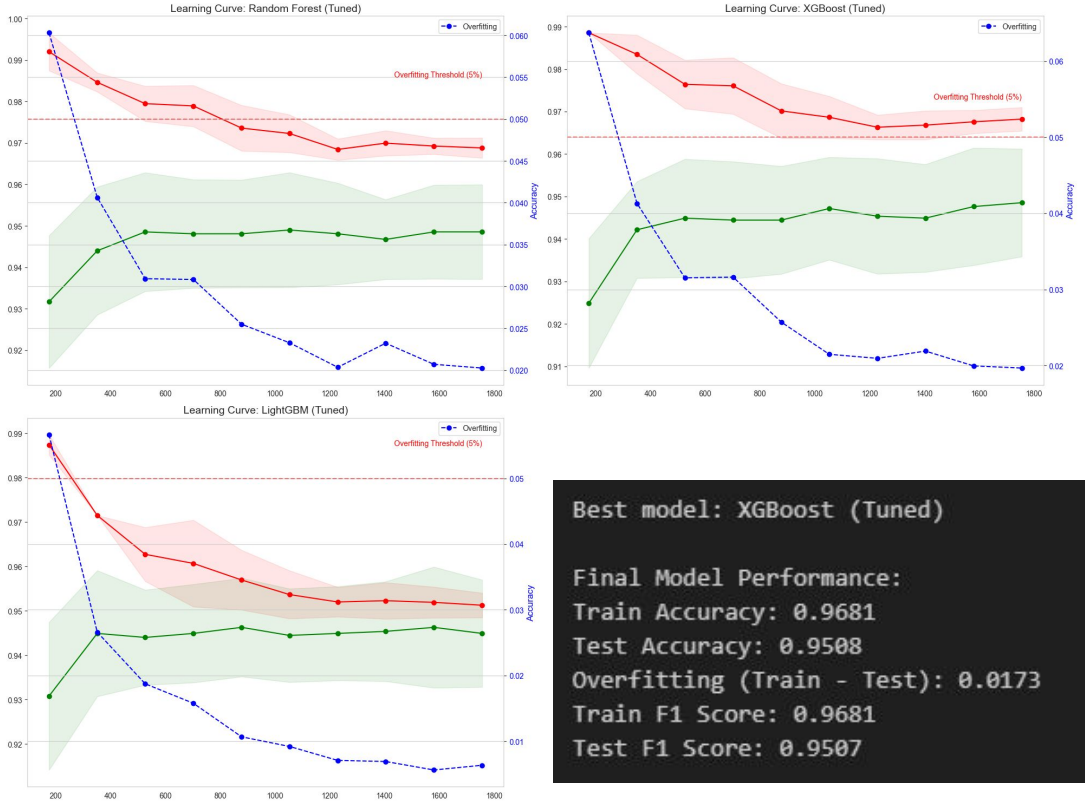
Delta!!!

I made the spirit airlines of machine learning code



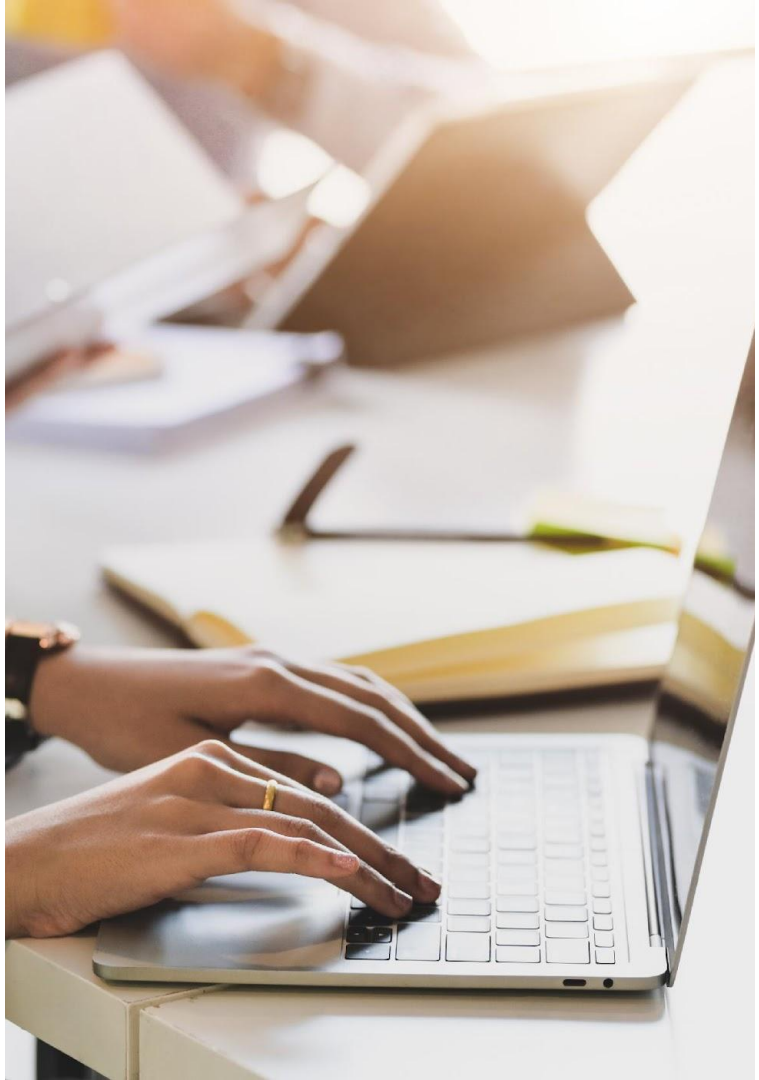
So what does this mean?

Sub text go brrr




```
Best model: XGBoost (Tuned)

Final Model Performance:
Train Accuracy: 0.9681
Test Accuracy: 0.9508
Overfitting (Train - Test): 0.0173
Train F1 Score: 0.9681
Test F1 Score: 0.9507
```



Regression Analysis

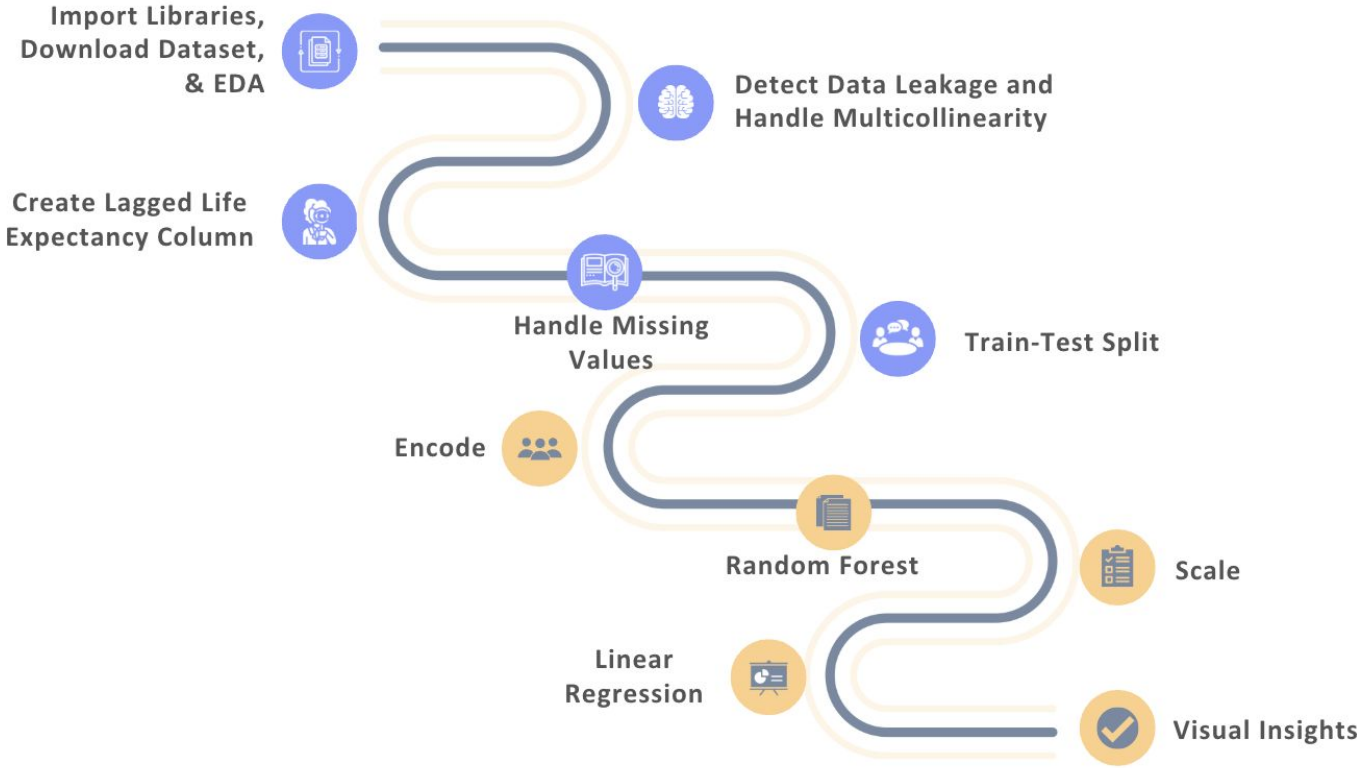
The Problem Statement: What are the factors that impact Life Expectancy?

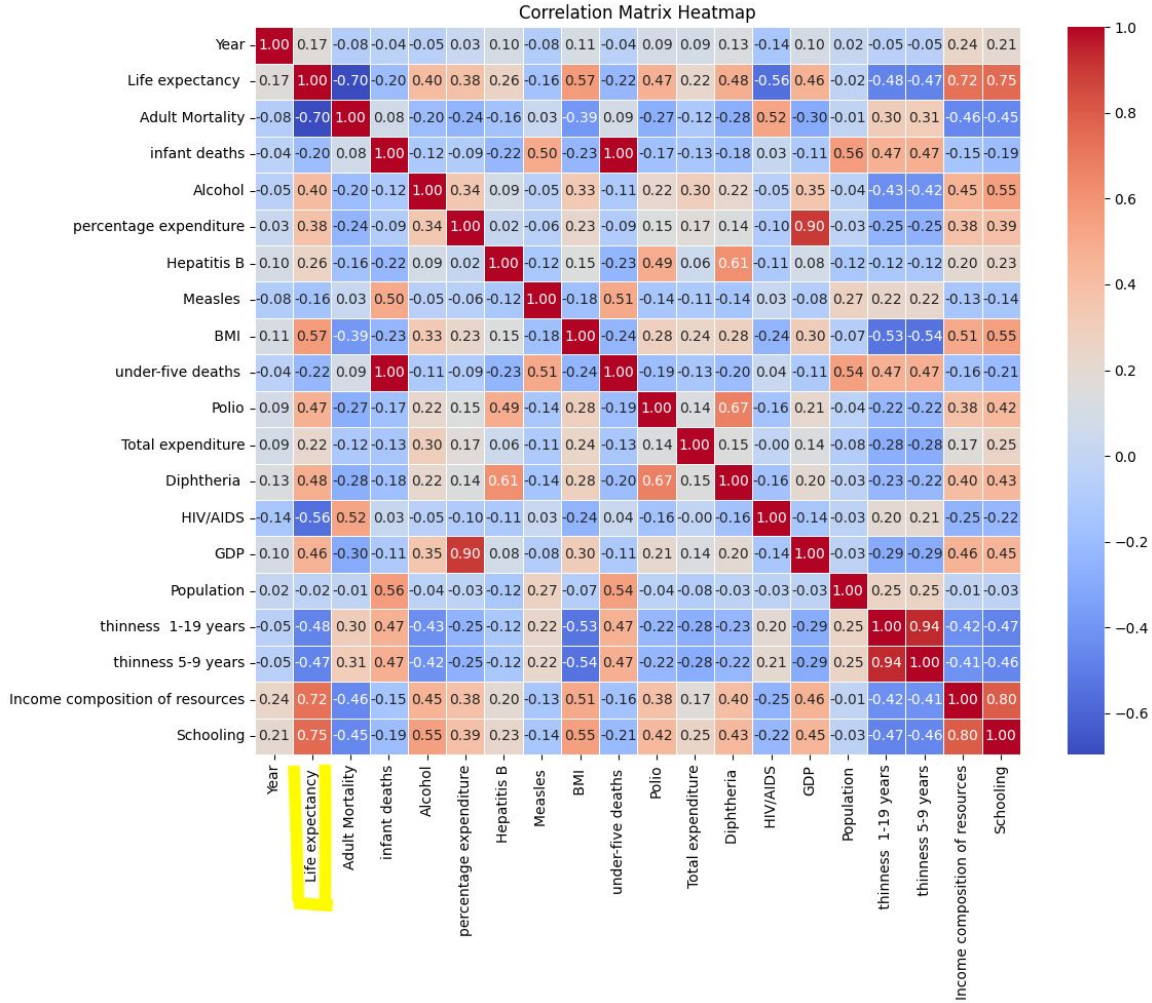
Goal: Predicting life expectancy using a variety of features.

Our Approach: Linear Regression and Random Forest.

Why Regression?: A regression model is ideal for predicting continuous outcomes, such as life expectancy, from various numeric and categorical predictors.

Process Workflow



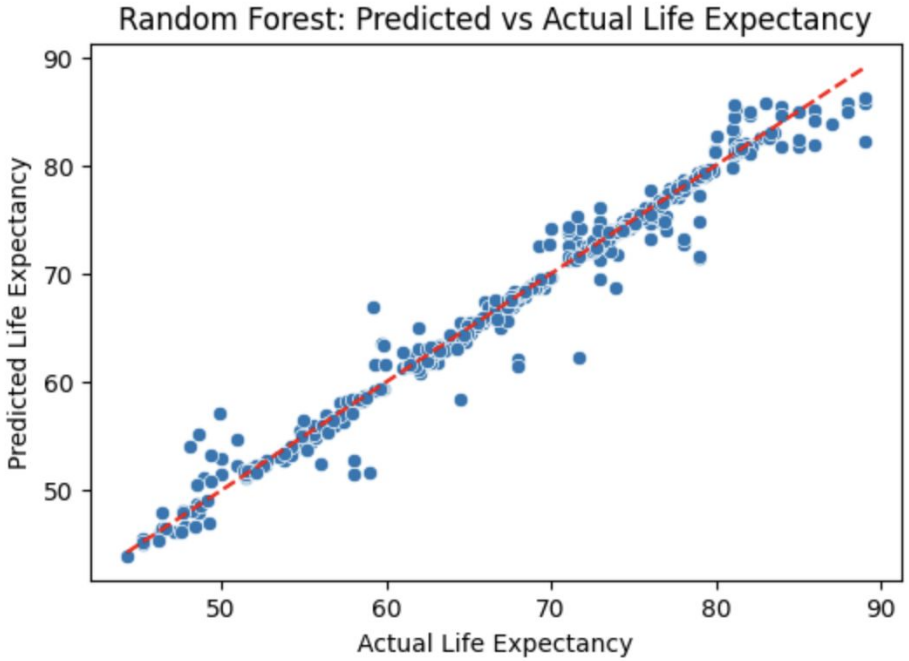

Correlation Matrix Heatmap
Year 1.00 0.17 -0.08 -0.04 -0.05 0.03 0.10 -0.08 0.11 -0.04 0.09 0.09 0.13 -0.14 0.10 0.02 -0.05 -0.05 0.24 0.21
Life expectancy 0.17 1.00 -0.70 -0.20 0.40 0.38 0.26 -0.16 0.57 -0.22 0.47 0.22 0.48 -0.56 0.46 -0.02 -0.48 -0.47 0.72 0.75
Adult Mortality -0.08 -0.70 1.00 0.08 -0.20 -0.24 -0.16 0.03 -0.39 0.09 -0.27 -0.12 -0.28 0.52 -0.30 -0.01 0.30 0.31 -0.46 -0.45
infant deaths -0.04 -0.20 0.08 1.00 -0.12 -0.09 -0.22 0.50 -0.23 1.00 -0.17 -0.13 -0.18 0.03 -0.11 0.56 0.47 0.47 -0.15 -0.19
Alcohol -0.05 0.40 -0.20 -0.12 1.00 0.34 0.09 -0.05 0.33 -0.11 0.22 0.30 0.22 -0.05 0.35 -0.04 -0.43 -0.42 0.45 0.55
percentage expenditure 0.03 0.38 -0.24 -0.09 0.34 1.00 0.02 -0.06 0.23 -0.09 0.15 0.17 0.14 -0.10 0.90 -0.03 -0.25 -0.25 0.38 0.39
Hepatitis B 0.10 0.26 -0.16 -0.22 0.09 0.02 1.00 -0.12 0.15 -0.23 0.49 0.06 0.61 -0.11 0.08 -0.12 -0.12 -0.12 0.20 0.23
Measles -0.08 -0.16 0.03 0.50 -0.05 -0.06 -0.12 1.00 -0.18 0.51 -0.14 -0.11 -0.14 0.03 -0.08 0.27 0.22 0.22 -0.13 -0.14
BMI 0.11 0.57 -0.39 -0.23 0.33 0.23 0.15 -0.18 1.00 -0.24 0.28 0.24 0.28 -0.24 0.30 -0.07 -0.53 -0.54 0.51 0.55
under-five deaths -0.04 -0.22 0.09 1.00 -0.11 -0.09 -0.23 0.51 -0.24 1.00 -0.19 -0.13 -0.20 0.04 -0.11 0.54 0.47 0.47 -0.16 -0.21
Polio 0.09 0.47 -0.27 -0.17 0.22 0.15 0.49 -0.14 0.28 -0.19 1.00 0.14 0.67 -0.16 0.21 -0.04 -0.22 -0.22 0.38 0.42
Total expenditure 0.09 0.22 -0.12 -0.13 0.30 0.17 0.06 -0.11 0.24 -0.13 0.14 1.00 0.15 -0.00 0.14 -0.08 -0.28 -0.28 0.17 0.25
Diphtheria 0.13 0.48 -0.28 -0.18 0.22 0.14 0.61 -0.14 0.28 -0.20 0.67 0.15 1.00 -0.16 0.20 -0.03 -0.23 -0.22 0.40 0.43
HIV/AIDS -0.14 -0.56 0.52 0.03 -0.05 -0.10 -0.11 0.03 -0.24 0.04 -0.16 -0.00 -0.16 1.00 -0.14 -0.03 0.20 0.21 -0.25 -0.22
GDP 0.10 0.46 -0.30 -0.11 0.35 0.90 0.08 -0.08 0.30 -0.11 0.21 0.14 0.20 -0.14 1.00 -0.03 -0.29 -0.29 0.46 0.45
Population 0.02 -0.02 -0.01 0.56 -0.04 -0.03 -0.12 0.27 -0.07 0.54 -0.04 -0.08 -0.03 -0.03 -0.03 1.00 0.25 0.25 -0.01 -0.03
thinness 1-19 years -0.05 -0.48 0.30 0.47 -0.43 -0.25 -0.12 0.22 -0.53 0.47 -0.22 -0.28 -0.23 0.20 -0.29 0.25 1.00 0.94 -0.42 -0.47
thinness 5-9 years -0.05 -0.47 0.31 0.47 -0.42 -0.25 -0.12 0.22 -0.54 0.47 -0.22 -0.28 -0.22 0.21 -0.29 0.25 0.94 1.00 -0.41 -0.46
Income composition of resources 0.24 0.72 -0.46 -0.15 0.45 0.38 0.20 -0.13 0.51 -0.16 0.38 0.17 0.40 -0.25 0.46 -0.01 -0.42 -0.41 1.00 0.80
Schooling 0.21 0.75 -0.45 -0.19 0.55 0.39 0.23 -0.14 0.55 -0.21 0.42 0.25 0.43 -0.22 0.45 -0.03 -0.47 -0.46 0.80 1.00
Year
Life expectancy
Adult Mortality
infant deaths
Alcohol
percentage expenditure
Hepatitis B
Measles
BMI
under-five deaths
Polio
Total expenditure
Diphtheria
HIV/AIDS
GDP
Population
thinness 1-19 years
thinness 5-9 years
Income composition of resources
Schooling
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

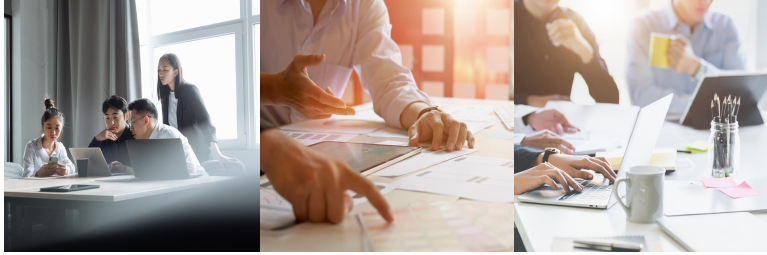

Life Expectancy vs. Adult Mortality
Life expectancy
Adult Mortality
90
80
70
60
50
40
0
100
200
300
400
500
600
700

Results

Linear Regression R^2 Score:	0.9697
Linear Regression MSE:	2.8684
Random Forest R^2 Score:	0.9742
Random Forest MSE:	2.4434

Visual Insights



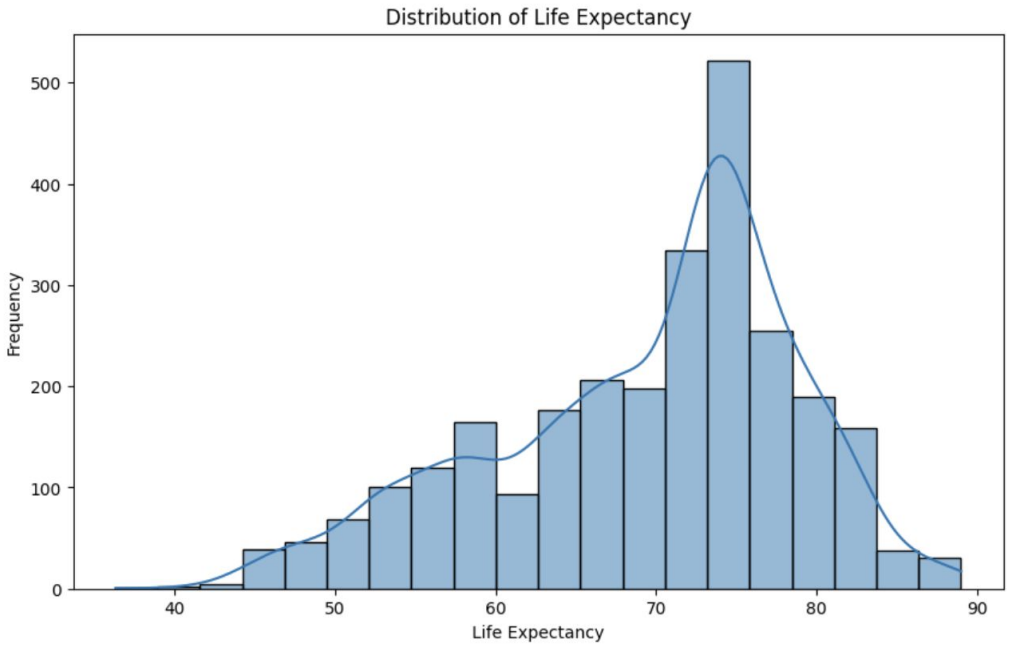


- Strong positive correlation
- Some instances of outliers
- Model Effectiveness



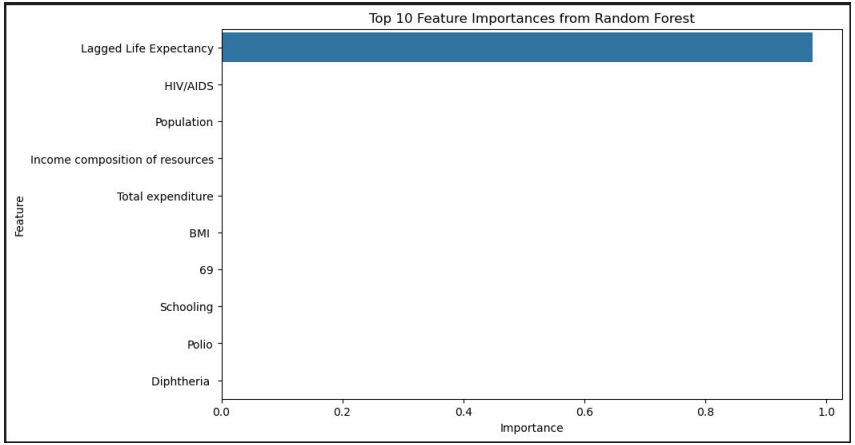
Visual Insights *Continued*

o



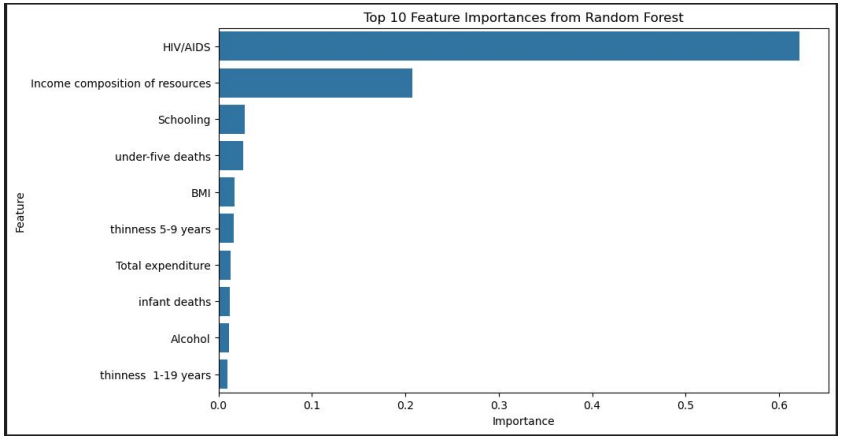
- **Key Implications**
- **Potential Contributing Factors**

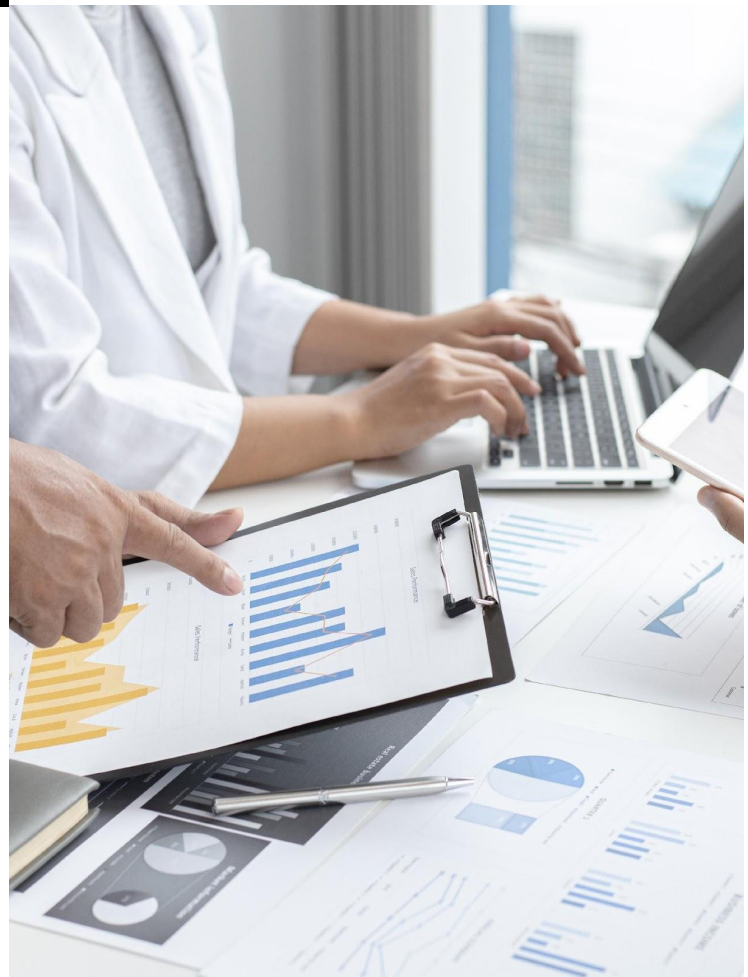




Further Exploration

- **Overfitting**
- **Additional Evaluation and Tuning**





Model	R^2 Score	MSE
Linear Regression	0.679854	30.138812
Ridge Regression	0.680977	30.033117
Lasso Regression	0.614680	36.274393
ElasticNet Regression	0.628472	34.975971
Random Forest	0.933913	6.221531
Gradient Boosting	0.896896	9.706326
Support Vector Regression	0.741046	24.378221
Decision Tree	0.877399	11.541767
K-Neighbors	0.842791	14.799808

SVR Tuning Results:
R^2: 0.8192
MSE: 17.0193

ANALYSIS

- **Random Forest still performs well**
- **SVR improved from .741 to .892 with hyperparameter tuning**

GROUP FOUR

THANK YOU

March 31, 2025