Life Expectancy Prediction

LINKEDIN

GITHUB

PRESENTED BY: MOHAMMAD AMIL KHAN

Project Overview & Problem Statement

PROBLEM STATEMENT:

LIFE EXPECTANCY IS A CRUCIAL MEASURE OF A NATION'S HEALTH AND DEVELOPMENT. NUMEROUS FACTORS SUCH AS GDP, HEALTHCARE INVESTMENT, EDUCATION, DISEASE PREVENTION, AND NUTRITION CONTRIBUTE TO LIFE EXPECTANCY ACROSS COUNTRIES. DESPITE RICH HISTORICAL HEALTH DATASETS, MANY COUNTRIES STRUGGLE TO USE THIS DATA EFFECTIVELY FOR PREDICTIVE INSIGHT. THIS PROJECT AIMS TO BUILD ACCURATE MODELS TO PREDICT LIFE EXPECTANCY AND HELP GUIDE POLICY DECISIONS USING DATA-DRIVEN INSIGHTS.

OBJECTIVE:

- ANALYZE THE RELATIONSHIPS BETWEEN SOCIO-ECONOMIC AND HEALTH INDICATORS AND LIFE EXPECTANCY.
- BUILD REGRESSION MODELS TO PREDICT LIFE EXPECTANCY USING THESE FEATURES.
- COMPARE MODEL PERFORMANCE TO DETERMINE THE MOST EFFECTIVE APPROACH.
- VISUALIZE ACTUAL VS PREDICTED OUTCOMES TO EVALUATE MODEL ACCURACY.
- RECOMMEND DATA-DRIVEN STRATEGIES FOR IMPROVING GLOBAL HEALTH FORECASTING.

Dataset Overview

Column Name	Description
Country	Name of the country
Year	Year of observation
Status	Development status (Developed / Developing)
Life expectancy	Average number of years a person is expected to live
Adult Mortality	Deaths between ages 15–60 per 1,000 people
Infant deaths	Infant deaths per 1,000 live births
Alcohol	Per capita alcohol consumption (litres)
Hepatitis B	% of 1-year-olds immunized against Hepatitis B
Measles	Number of reported measles cases
BMI	Average Body Mass Index
Under-five deaths	Deaths of children under age five per 1,000 live births

Column Name	Description
Polio	% of children immunized against polio
Total expenditure	Health expenditure (% of total government spending)
Diphtheria	% of children immunized against diphtheria
HIV/AIDS	Deaths due to HIV/AIDS per 1,000 people
GDP	Gross Domestic Product per capita
Population	Total population
Thinness 1-19 years	% of thinness among youth (ages 1–19)
Thinness 5-9 years	% of thinness among young children (ages 5–9)
Income composition of resources	Composite index of income and development
Schooling	Average years of schooling

Process

1. Imported libraries for data handling, visualization, and modeling (e.g., pandas, scikit-learn, XGBoost).
2. Loaded the dataset and displayed the first few rows.
3. Handled missing values with mean imputation and encoded the categorical Status column.
4. Dropped non-numeric columns like Country and Year.
5. Analyzed correlations, visualized with a heatmap to identify top influencing features.
6. Split data into training and test sets for modeling.
7. Trained models: Linear Regression, Random Forest, and XGBoost.
8. Evaluated performance using RMSE and R^2 Score and selected the best model.
9. Visualized predictions and compared actual vs predicted life expectancy values.
10. Insight and Conclusion.

ALL MODEL RESULT

	Model	RMSE	R^2 Score
1	Random Forest	1.65	0.9686
2	XGBoost	1.76	0.9643
0	Linear Regression	3.90	0.8241

1. **Random Forest** delivered the best performance with an RMSE of **1.65** and R^2 score of **0.9686**.
2. **XGBoost** closely followed with RMSE **1.76** and R^2 **0.9643**, showing strong predictive power.
3. **Linear Regression** lagged behind significantly (RMSE **3.90**, R^2 **0.8241**), indicating poor fit.
4. Both tree-based models captured the data's complexity far better than the linear model.

Higher R^2 and lower RMSE from ensemble methods indicate strong model generalization.

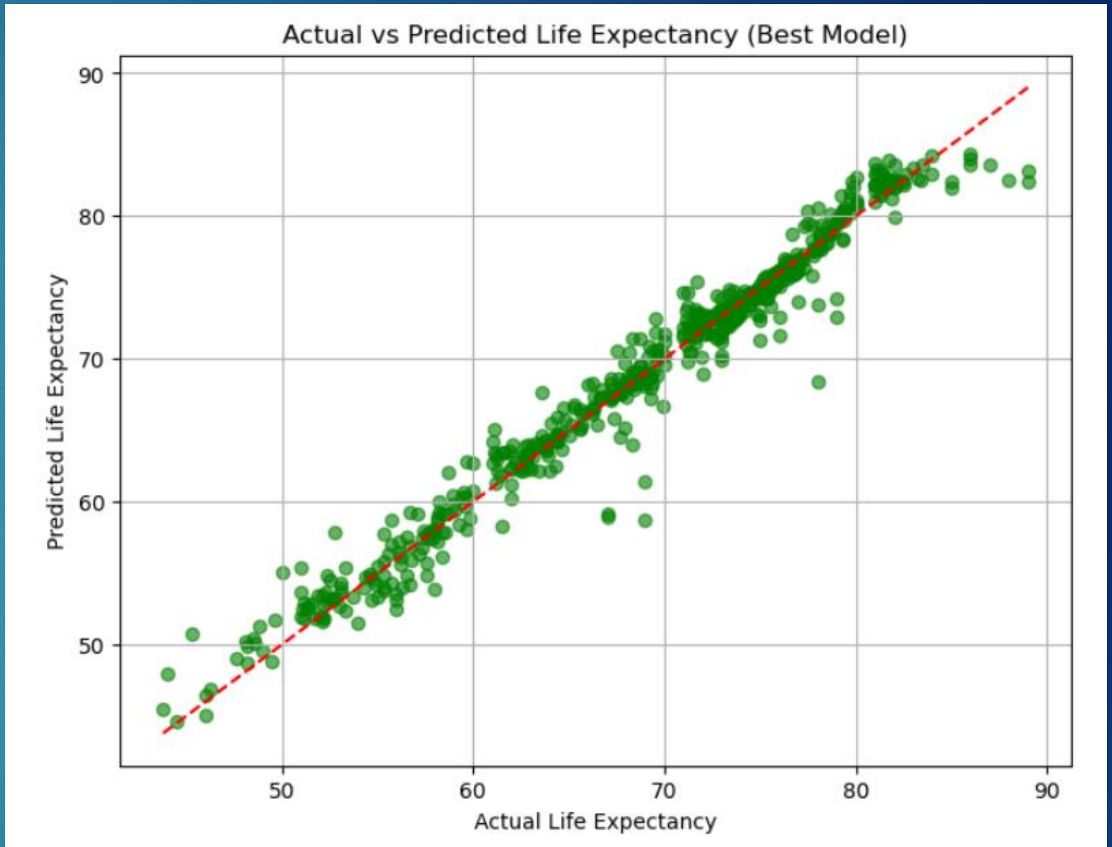
Conclusion

Tree-based ensemble models like Random Forest and XGBoost are far more effective for predicting life expectancy, with Random Forest being the top-performing model due to its superior accuracy and stability.

Actual vs Predicted value

CONCLUSION

THE SCATTER PLOT OF ACTUAL VS PREDICTED LIFE EXPECTANCY SHOWS A STRONG LINEAR ALIGNMENT ALONG THE RED DIAGONAL, INDICATING THAT THE MODEL'S PREDICTIONS CLOSELY MATCH THE TRUE VALUES. MOST DATA POINTS CLUSTER TIGHTLY AROUND THE LINE, DEMONSTRATING HIGH ACCURACY AND MINIMAL PREDICTION ERROR. THIS CONFIRMS THAT THE CHOSEN MODEL—LIKELY RANDOM FOREST—IS HIGHLY EFFECTIVE AT CAPTURING THE UNDERLYING PATTERNS IN THE DATA. OVERALL, THE MODEL GENERALIZES WELL AND IS SUITABLE FOR REAL-WORLD LIFE EXPECTANCY FORECASTING.



Final Conclusion

1. TREE-BASED MODELS LIKE RANDOM FOREST AND XGBOOST SIGNIFICANTLY OUTPERFORM LINEAR REGRESSION IN PREDICTING LIFE EXPECTANCY.

2. RANDOM FOREST IS THE MOST RELIABLE MODEL, OFFERING BOTH LOW ERROR AND HIGH CONSISTENCY.

3. THESE MODELS EFFECTIVELY CAPTURE COMPLEX RELATIONSHIPS, MAKING THEM SUITABLE FOR REAL-WORLD FORECASTING BASED ON DEMOGRAPHIC AND HEALTH INDICATORS.