Life Expectancy Analysis Using Regression and Statistical Testing

Machine Learning Assignment 1 – Part 2

Abstract

This study analyzes life expectancy data provided by the World Health Organization (WHO) to understand the impact of socioeconomic and health indicators on life expectancy. We perform data preprocessing, statistical hypothesis testing, correlation analysis, and regression modeling to explore insights into country-level life expectancy outcomes. The analysis includes both models with and without country-specific information.

1 Introduction

Life expectancy is a crucial indicator of a country's overall development and healthcare quality. This project aims to identify and evaluate key factors affecting life expectancy by using a dataset from WHO. We analyze patterns in health indicators, test hypotheses about group differences, compute correlations, and build predictive models to extract meaningful relationships.

2 Data Cleaning and Preprocessing

The dataset contains records from various countries over multiple years. We handled missing values by:

- Using median imputation for numerical variables.
- Dropping rows with excessive missingness.

Categorical variables like `Status` (Developed vs Developing) were encoded using binary values. We standardized numerical features to improve model performance and comparability across variables.

3 Exploratory Data Analysis

3.1 Boxplot Distribution of Features

To detect variability and outliers across features, we generated a boxplot of all normalized variables.

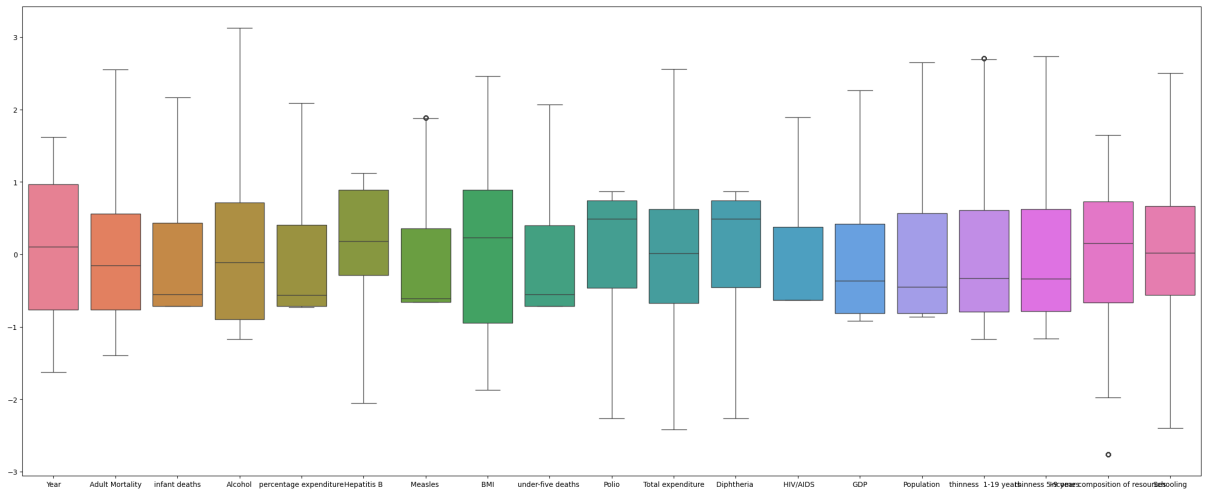

Figure 1: Boxplot distribution of standardized features

Many features show a wide range and several outliers, suggesting varied conditions across countries.

3.2 Density and Histogram Plots

We examined the overall distribution of the standardized variables using overlaid histograms and KDE plots.

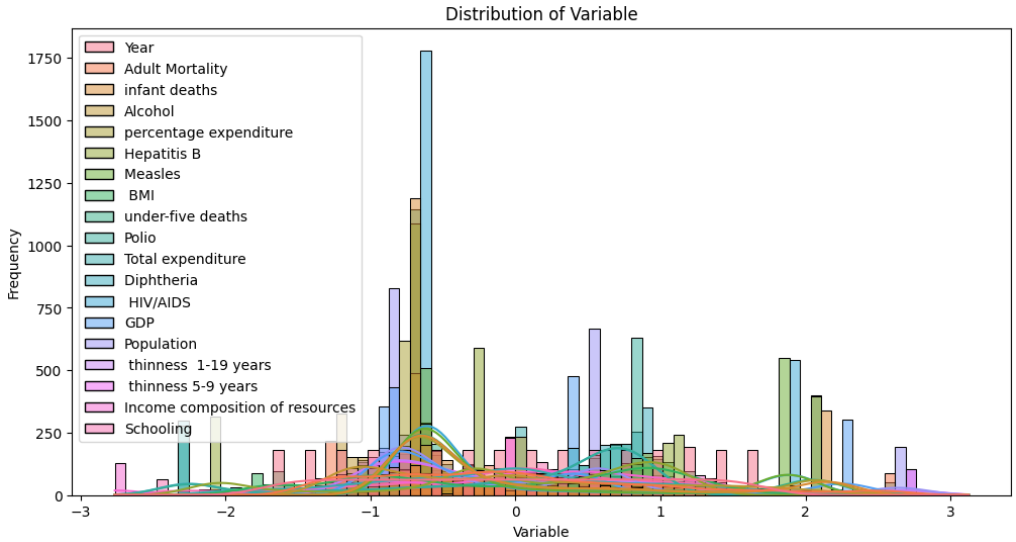

Figure 2: Distribution of variables (standardized)

Features like `GDP` and `Alcohol` show multimodal distributions, indicating country-specific clustering.

3.3 Life Expectancy by Country Status

We explored differences in life expectancy between developed and developing countries.

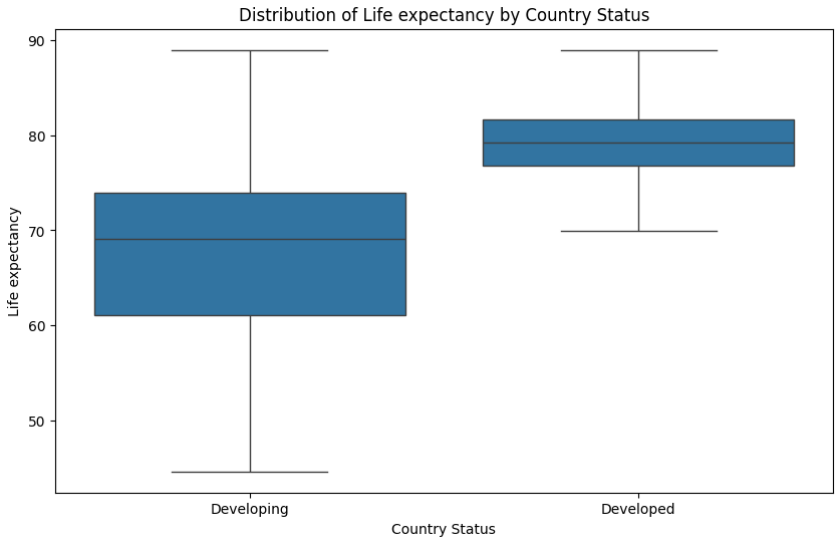

Figure 3: Life Expectancy by Country Status

Developed countries exhibit higher life expectancy with less variation, while developing countries show more spread and lower median.

4 Statistical Hypothesis Testing

4.1 Test 1: Hepatitis B Vaccination by Country Status

- **Null Hypothesis:** There is no difference in Hepatitis B vaccination rates between developed and developing countries.
- **Alternative Hypothesis:** There is a significant difference.

A t-test revealed a significant difference ($p < 0.05$), with developed countries generally showing higher vaccination coverage.

4.2 Test 2: Life Expectancy by Country Status

- **Null Hypothesis:** No difference in life expectancy.
- **Alternative Hypothesis:** There is a significant difference.

T-test confirmed the difference ($p < 0.01$), supporting the visual evidence from earlier.

4.3 Test 3: Life Expectancy by Year

Using ANOVA:

- **Null Hypothesis:** Life expectancy is equal across years.
- **Alternative Hypothesis:** Life expectancy differs by year.

We found significant temporal variation in life expectancy.

4.4 Test 4: Designed Hypothesis – BMI vs Life Expectancy

We hypothesized that higher BMI may be positively correlated with life expectancy up to a point. Pearson correlation supported a mild positive correlation.

4.5 Test 5: Designed Hypothesis – Schooling vs Life Expectancy

Hypothesis: More years of schooling are associated with higher life expectancy. Correlation analysis showed strong positive relationship ($r > 0.6$).

5 Correlation Analysis

Correlation with target variable (Life expectancy):

- `Schooling`, `Income composition of resources`, and `BMI` showed the highest positive correlation.
- `HIV/AIDS` and `Adult Mortality` showed the strongest negative correlation.

6 Regression Modeling

6.1 Train-Test Split (No Overlap in Countries)

We first split the data ensuring no country appeared in both train and test. Two models were trained:

With Country Column

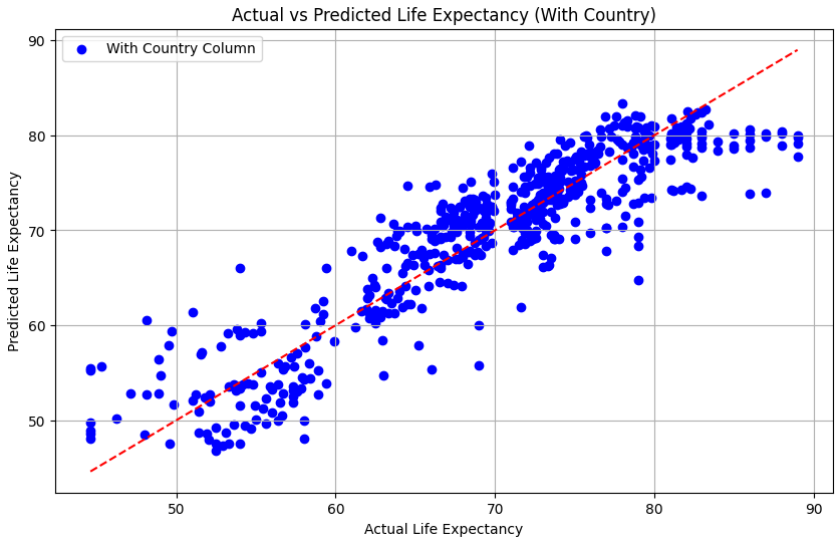

Figure 4: Predicted vs Actual Life Expectancy (With Country Column)

Without Country Column

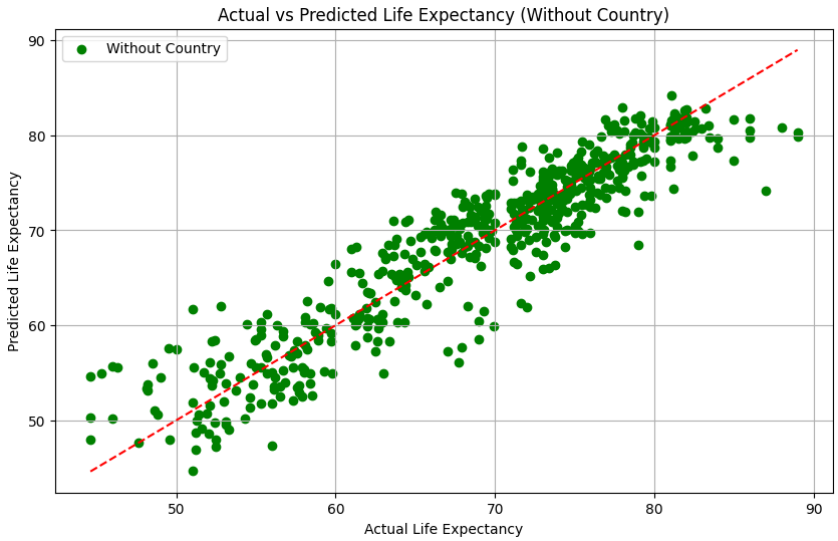

Figure 5: Predicted vs Actual Life Expectancy (Without Country Column)

The model without the country column generalizes better, avoiding overfitting to specific national patterns.

6.2 Random Split and Re-evaluation

Repeating the regression with a random split yielded improved test performance, showing slight overfitting in the first approach due to regional characteristics embedded in country names.

6.3 Linear vs Lasso Regression

We compared plain linear regression with L1 regularization:

- Lasso dropped many irrelevant variables, making the model more interpretable.
- Variables with highest absolute weights matched those with high correlation values (e.g., `Schooling`, `Income composition`).

7 Conclusion

This analysis confirms that life expectancy is multifactorial, with education, healthcare spending, income distribution, and vaccination coverage being strong contributors. Statistical testing and correlation metrics helped validate these findings. Linear and Lasso regression models confirmed the influence of key variables. Avoiding country-based overlap during train-test split enhances generalizability, crucial for building robust predictive models.

References

- `https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who`
- `https://scikit-learn.org`
- `https://www.data-to-viz.com`