Regression Analysis Report: Life Expectancy Prediction

1. Project Overview

The objective of this analysis was to develop a predictive model that estimates **Life Expectancy** based on historical data and health indicators. The model uses **Multiple Linear Regression**, a supervised learning algorithm that models the relationship between two or more features and a single target variable.

2. Dataset Summary & Descriptive Statistics

The dataset consists of **2,204 records** representing various countries.

Metric	Year	Life Expectancy (Target)	Infant Mortality (Scaled)
Mean	2007.6	67.85 Years	~0.00
Std Dev	4.61	8.72 Years	1.00
Minimum	2000	42.30 Years	-1.24
Maximum	2015	89.00 Years	1.80

Key Insight: The target variable (Life Expectancy) ranges from 42.3 to 89.0 years, showing significant global health disparities.

3. Correlation Analysis

Correlation measures the strength and direction of the linear relationship between variables (from -1 to +1).

- **Year vs. Life Expectancy (0.166):** A weak positive correlation. This suggests that over time, life expectancy has a slight upward trend globally.

- **Infant Mortality vs. Life Expectancy (-0.537):** A **moderate negative correlation**. This is the strongest predictor in our current model. As infant mortality decreases, life expectancy significantly increases.

4. Model Composition

The Multiple Linear Regression equation derived from the data is:

{Life Expectancy} = -469.49 + (0.268 \times \text{Year}) - (4.56 \times \text{Infant Mortality})$$

Breakdown of Coefficients:

1. **Intercept (-469.49):** The theoretical value of life expectancy if all features were zero (often just a mathematical anchor).
2. **Year (0.268):** For every year that passes, life expectancy is predicted to increase by approximately **0.27 years**, holding other factors constant.
3. **Infant Mortality (-4.56):** For every 1-unit increase in the infant mortality index, life expectancy is predicted to drop by **4.56 years**.

5. Performance Metrics

To evaluate how well the model predicts life expectancy on unseen data (the Test Set), we use the following metrics:

Metric	Value	Interpretation
MAE (Mean Absolute Error)	**5.74**	On average, the model's predictions are off by about 5.7 years.
RMSE (Root Mean Squared Error)	**7.15**	Penalizes larger errors; shows the typical magnitude of the error.
R^2 Score (R-Squared)	**0.343**	The model explains **34.3%** of the variation in Life Expectancy.

6. Conclusions & Recommendations

Conclusions:

- The model successfully identifies **Infant Mortality** as a primary driver of Life Expectancy.
- The current model explains roughly one-third of the health outcome variance.
- The positive coefficient for "Year" confirms a global trend of improving health over the 2000-2015 period.

Recommendations for Improvement:

1. **Feature Addition:** The R^2 score (34%) suggests many factors are missing. Adding variables like GDP, Schooling, Alcohol consumption, or Immunization rates would likely increase accuracy.
2. **Handling Heterogeneity:** Since the dataset contains many different countries, creating regional sub-models (e.g., Africa vs. Europe) might provide better localized predictions.
3. **Non-Linearity:** If the relationship isn't perfectly straight, exploring **Polynomial Regression** or **Random Forest** models could capture more complex patterns.