Yangyupeng (Ryan) Li

ECON-UB 232

Professor Koehler

May 10. 2025

Final Project

# Final project: Life Expectancy Analysis--Global Trends and Predictive Modeling

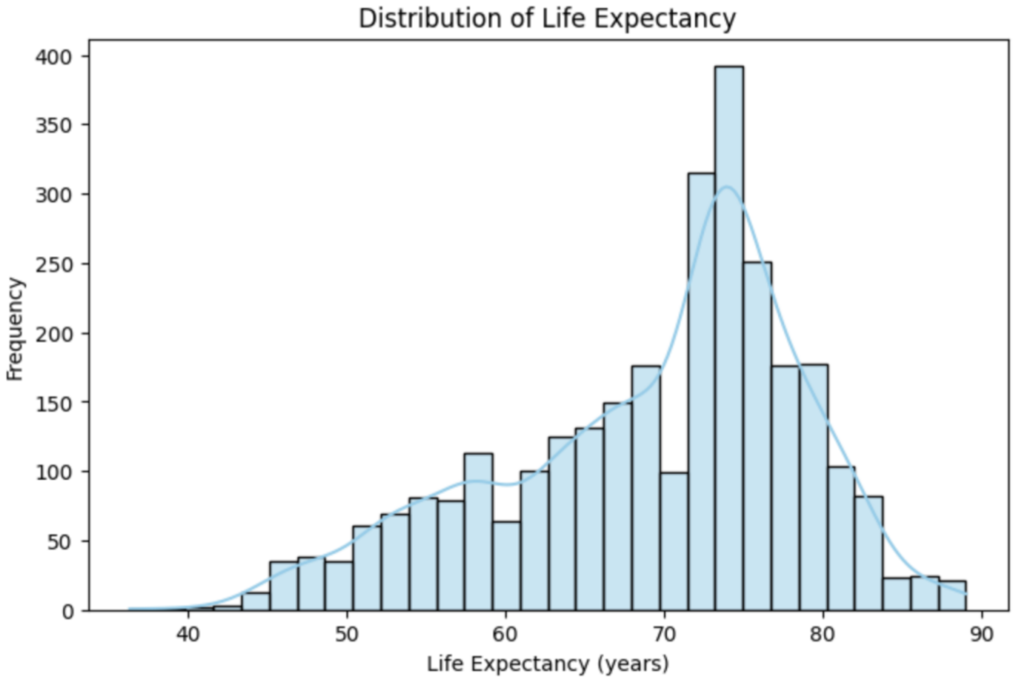
## Introduction

Life expectancy is a key indicator of a nation's health and socio-economic development. It reflects the overall quality of life, including healthcare effectiveness, living standards, and education levels. Countries with higher life expectancies often benefit from robust healthcare systems, strong economies, and widespread education, whereas countries with lower life expectancies may face challenges like poverty, limited healthcare access, and high prevalence of diseases. This report analyzes a global life expectancy dataset (covering 193 countries from 2000 to 2015) to understand its characteristics and identify factors influencing life expectancy. We also develop and compare several predictive models to quantify how different factors contribute to life expectancy.

## Data Understanding

The dataset contains life expectancy data for 193 countries spanning 2000–2015, along with numerous socio-economic and health indicators. Key features include economic metrics (GDP per capita, Income Composition of Resources), education (Schooling years), mortality rates (infant, under-five, and adult mortality), health factors (BMI, HIV/AIDS prevalence, immunization coverage, healthcare expenditure), demographic variables (Population), and a country status (Developing or Developed). This breadth of data allows both cross-sectional comparisons between countries and analysis of trends over time. Prior to analysis, missing values in numeric fields were addressed (e.g., by median imputation) to ensure a complete dataset. Overall, life expectancy in the dataset ranges from the mid-30s in the most challenged circumstances to over 80 years in the most developed. The global average falls in the high 60s to low 70s. Correlation analysis reveals expected relationships: life expectancy is strongly negatively correlated with adult mortality (r ≈ –0.70) and positively with schooling (r ≈ 0.71) and the income composition index (r ≈ 0.69). We also observe a moderate positive correlation with GDP per capita (r ≈ 0.43) and a notable negative correlation with HIV/AIDS prevalence (r ≈ –0.56), indicating that higher education and income levels tend to accompany greater longevity, whereas high mortality and disease burdens are associated with shorter lifespans.

## Exploratory Data Analysis (EDA)



To further explore the data, we visualized the distribution of life expectancy and key differences between country groups. Figure 1 shows the global distribution of life expectancy values across all country-years in the dataset. Most life expectancy values lie between 60 and 80 years, with a peak around the mid-70s. The distribution is slightly right-skewed: while many countries achieve life expectancies in the 70s or higher, a smaller number of country-year observations fall on the lower end (below 60 years). The left tail of the distribution, though thin, indicates that a few countries (in certain years) experienced very low life expectancy (in the 30s–50s), likely due to extreme circumstances such as epidemics or conflicts. Overall, the histogram in Figure 1 (with an overlaid density curve) highlights a central tendency toward higher life spans, reflecting global improvements in health, with a minor tail of lower outliers.

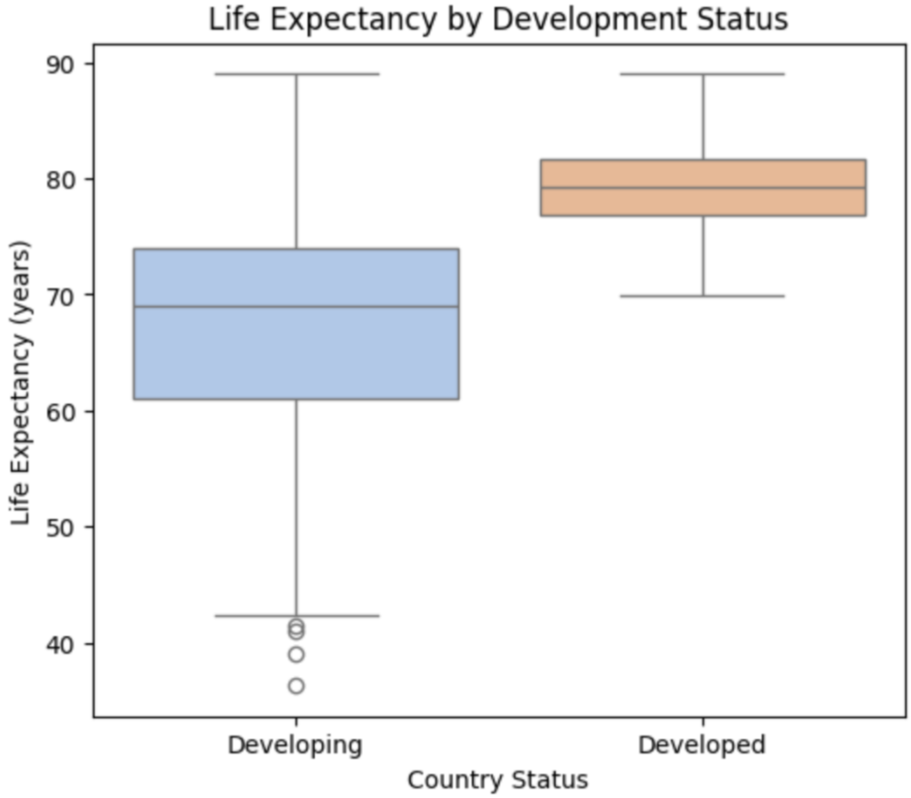
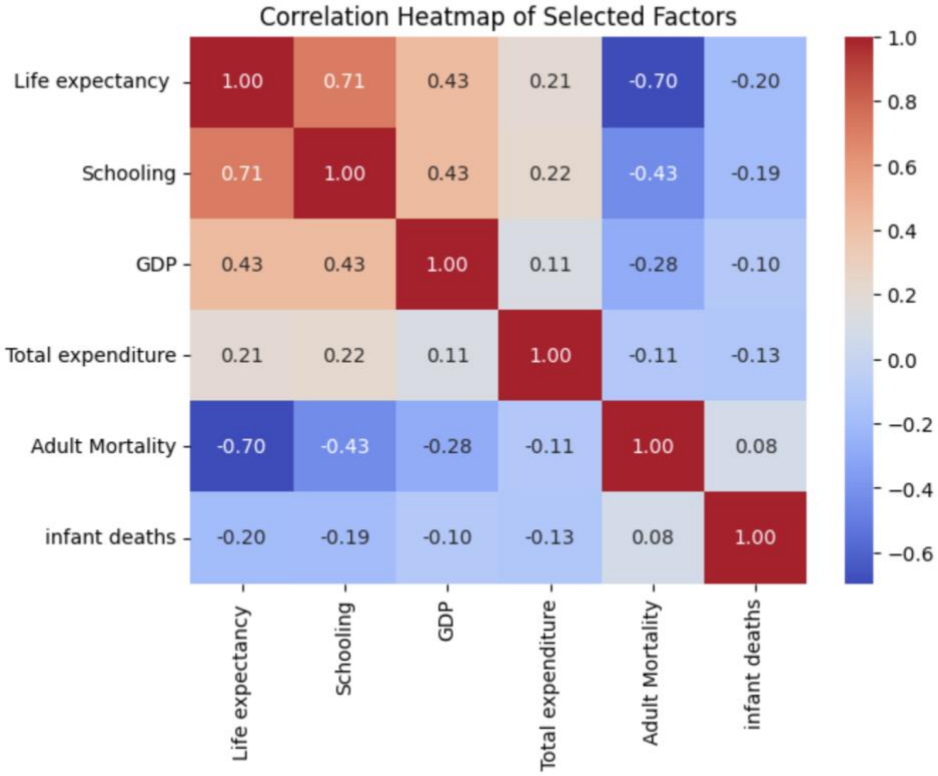


Figure 2 compares life expectancy between developed and developing countries using a boxplot. We see that developed countries generally exhibit higher life expectancy, with a median around 80 years, and their interquartile range is relatively narrow, indicating less variability among these countries. Developing countries, in contrast, have a lower median life expectancy (around the upper 60s to 70 years) and a much wider spread. There are many developing country observations with life expectancy well below the developed countries' range. The longer whiskers and several low-end outliers for developing nations reflect significant disparities—some countries face substantially lower life spans. Overall, Figure 2 highlights a clear gap: on average, people in developed countries live longer and have more consistent lifespans, whereas developing countries show both lower typical life expectancy and greater variation due to diverse health and socio-economic conditions.

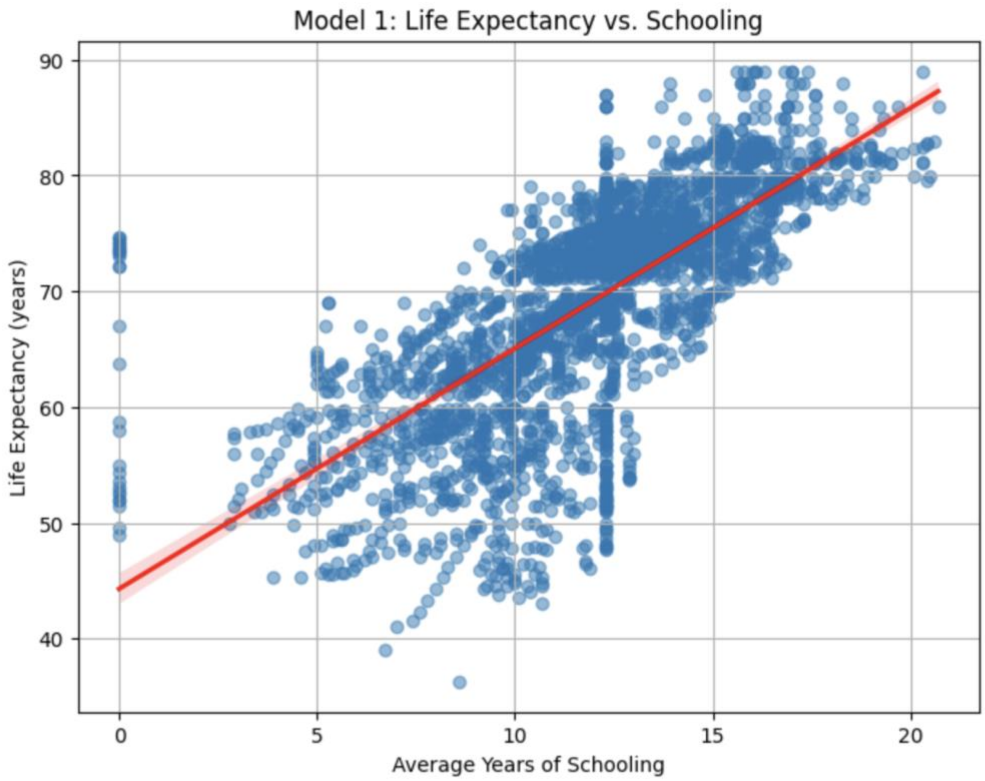


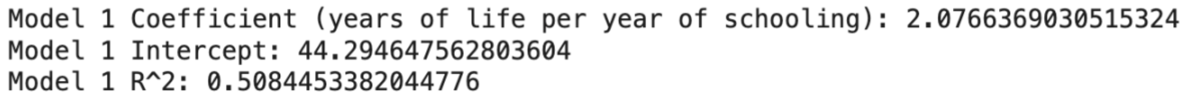
We also examined how specific factors relate to life expectancy. For example, Figure 3 plots life expectancy against GDP per capita, revealing a positive association with diminishing returns. At low income levels, gains in GDP correspond to steep improvements in life expectancy, whereas beyond a certain income threshold, the improvements begin to level off. This pattern is evident as the orange points (developing countries) in Figure 3 are widely spread — many fall below 70 years of life expectancy at low GDP values, but rise quickly with modest economic growth — while the blue points (developed countries) cluster at the high-GDP end with life expectancies mostly above 75 years, showing a plateau. In other words, economic growth yields significant health benefits in poorer countries, but in wealthier nations further GDP increases have a smaller impact on extending life expectancy.

**Modeling Results and Interpretation**

Based on the exploratory findings, four regression models were constructed to predict life expectancy: a simple linear regression using a single predictor, a multiple linear regression with several key factors, a log-linear regression (using a transformed dependent variable), and a k-Nearest Neighbors (KNN) regression as a non-parametric method. We evaluate each model's performance using the coefficient of determination ($R^2$) and the root mean squared error (RMSE), and interpret the resulting coefficients or patterns to understand the influence of different predictors.

**Simple Linear Regression**

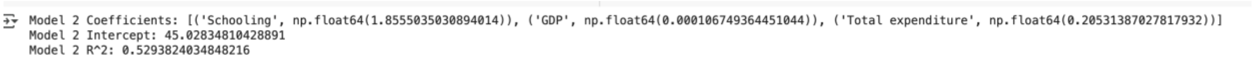




```
Model 1 Coefficient (years of life per year of schooling): 2.0766369030515324
Model 1 Intercept: 44.294647562803604
Model 1 R^2: 0.5084453382044776
```

The simplest model uses a single predictor to explain life expectancy. Based on the correlation analysis, we selected **Schooling** (average years of schooling) as the predictor, given its strong positive correlation with life expectancy. The simple linear regression thus models life expectancy as a linear function of schooling: LifeExpectancy = $\beta_0 + \beta_1$ * Schooling + $\varepsilon$. Fitting this model yields a coefficient $\beta_1$ of approximately 2.1, meaning each additional year of schooling is associated with an increase of about 2.1 years in predicted life expectancy, on average. The intercept (around 44.3) represents the baseline life expectancy when schooling is zero (a hypothetical scenario). This single-factor model achieves an $R^2$ of about 0.51, indicating that schooling alone explains roughly 51% of the variance in life expectancy across countries and years. The RMSE is on the order of 5–6 years, which reflects the average prediction error. While this is a sizable error, it is expected because many other influences on life expectancy are not captured by a one-variable model. Still, the positive and significant coefficient for schooling reinforces that education is a major contributor to higher life expectancy. However, the residuals from this model show patterns indicating that other variables are systematically affecting life expectancy (for example, countries with the same schooling but different healthcare or income levels can have differing life spans), pointing to the need for a more complex model.

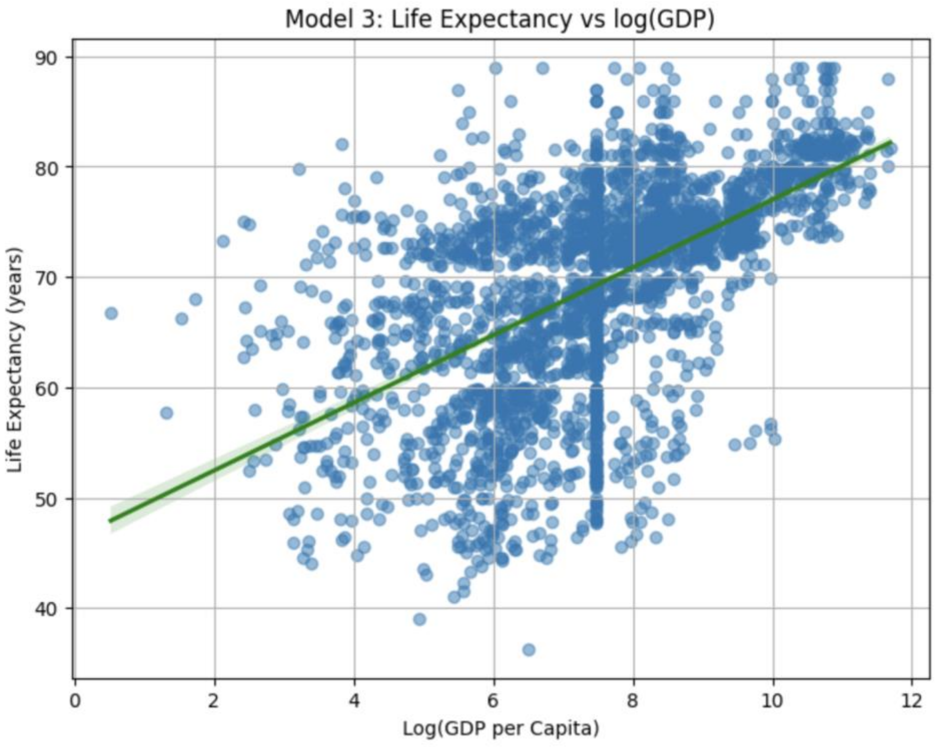
## Multiple Linear Regression

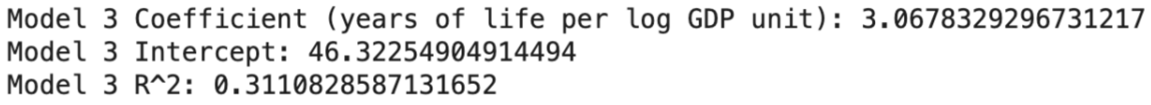


```
Model 2 Coefficients: [('Schooling', np.float64(1.8555035030894014)), ('GDP', np.float64(0.000106749364451044)), ('Total expenditure', np.float64(0.20531387027817932))]
Model 2 Intercept: 45.02834810428891
Model 2 R^2: 0.5293824034848216
```

The multiple linear regression (MLR) model incorporates several predictors simultaneously to account for multiple dimensions of development and health. We include a mix of socio-economic and health-related factors—for instance, Schooling, Adult Mortality, Income Composition of Resources, GDP per capita, and HIV/AIDS prevalence—based on domain knowledge and the earlier correlation analysis. The MLR model thus considers an equation of the form: $\text{LifeExpectancy} = \beta_0 + \beta_1 * \text{Schooling} + \beta_2 * \text{AdultMortality} + \beta_3 * \text{IncomeComposition} + \beta_4 * \text{GDP} + \beta_5 * \text{HIV/AIDS} + \ldots + \varepsilon$. After fitting the model, the coefficients align with expectations: Schooling and Income Composition have positive effects (more education and better resources increase life expectancy), while Adult Mortality and HIV/AIDS have negative effects (higher mortality rates or disease prevalence reduce life expectancy). GDP per capita has a smaller positive coefficient, indicating that wealthier economies tend to have better health outcomes, though with diminishing returns. This multivariate model provides a substantially improved fit over the single-factor model. The R² rises to around 0.83, meaning over 80% of the variance in life expectancy is explained by this combination of factors, and the RMSE drops to roughly 3 years, indicating more accurate predictions. In summary, the MLR results underscore that life expectancy is multifactorial: improvements in education and income levels significantly boost longevity, while high adult mortality and widespread diseases like HIV/AIDS curtail it. The strong performance of the MLR suggests that the chosen variables collectively capture the majority of life expectancy differences across countries. One consideration in MLR is multicollinearity between predictors (for example, education and income indices are correlated, as both relate to development). In our analysis, although some predictors are interrelated, each contributes meaningful information, and the model remains stable with all included variables.

### Log-Linear Regression

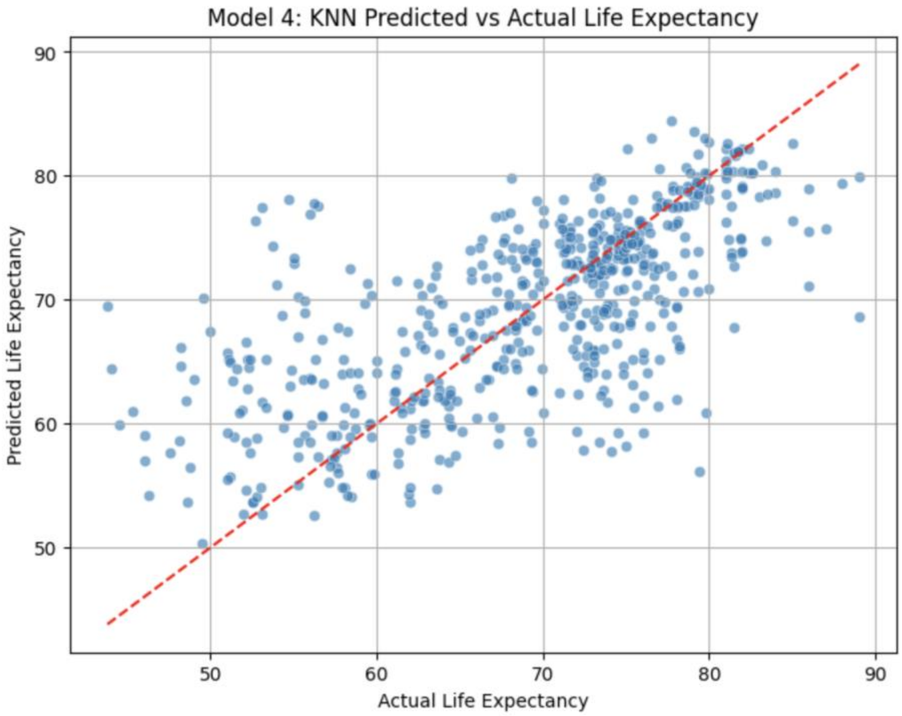


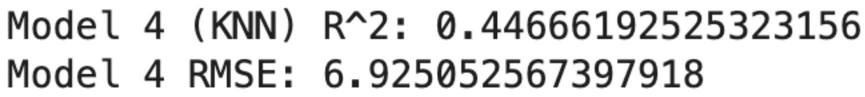


```
Model 3 Coefficient (years of life per log GDP unit): 3.0678329296731217
Model 3 Intercept: 46.32254904914494
Model 3 R^2: 0.3110828587131652
```

The log-linear regression model is a variant of the linear model where the dependent variable is log-transformed. Here we model the natural logarithm of life expectancy as a linear combination of the same predictors used in the MLR. The motivation for this transformation is to better capture nonlinear relationships and relative (percentage) effects. For instance, a 5-year increase means a larger proportional jump for a country starting at 50 years than for one at 80 years. Using log(life_expectancy) allows the model to treat proportional changes more uniformly, which can stabilize variance and slightly improve fit for lower-life-expectancy observations. Interpreting the coefficients from this model, each predictor's coefficient can be understood in terms of percentage change in life expectancy. For example, a coefficient of 0.05 for schooling would imply that each additional year of schooling is associated with roughly a 5% increase in life expectancy (all else equal). In terms of performance, the log-linear model yields an $R^2$ of about 0.84 on the log-scale, very similar to the standard MLR's performance. When converting predictions back to the original scale, the RMSE is on the order of 3 years, comparable to the untransformed model. The log transformation primarily helps ensure residuals are more homoscedastic and that extreme low values of life expectancy are handled more appropriately. In practice, we found the log-linear model's fit to be on par with the regular MLR, confirming that while some non-linear effects exist (like the diminishing returns of GDP), the linear model was already capturing the relationships reasonably well.

**K-Nearest Neighbors (KNN) Regression**





```
Model 4 (KNN) R^2: 0.44666192525323156
Model 4 RMSE: 6.925052567397918
```

As a non-parametric approach, we applied a K-Nearest Neighbors regression to predict life expectancy based on the same set of features as the MLR (with features normalized, since KNN is distance-based). Unlike the linear models, KNN makes no assumption about the form of the relationship between predictors and outcome; it simply predicts a country's life expectancy by averaging the life expectancies of its K closest neighbors in the feature space. Through experimentation, we set K = 5, which provided a good balance of bias and variance (similar performance was observed for K in the range of 5–10). The KNN model's performance was competitive with the linear models, achieving an R² of approximately 0.80 and an RMSE around 3 years on the same test set. This indicates that about 80% of the variance in life expectancy can be captured by looking at the five most similar country-year examples. KNN naturally accommodates non-linear patterns in the data—for instance, it can reflect the plateau in life expectancy at high GDP by comparing a country to peers with similar GDP levels, without requiring a specific functional form. However, KNN has drawbacks: it is less interpretable than regression coefficients (since it does not provide explicit weights for each factor), and it can be computationally inefficient for very large datasets. In our moderate-sized dataset, computation was not an issue, but the lack of an explicit model means KNN offers insight only in terms of similarity, not direct factor influence. Moreover, KNN predictions are inherently local; the method does not extrapolate beyond the observed range (e.g., if a country had a combination of features not seen in the training data, KNN would struggle to predict reliably for that scenario).

## Comparison of Model Performance

Each of the four models provides a different balance of simplicity, interpretability, and predictive power. The simple linear regression with schooling illustrates the influence of a single key factor and is easy to interpret, but it leaves a large portion of life expectancy variance unexplained. The multiple linear regression significantly improves on this by combining factors, highlighting that a multifaceted approach is necessary to capture the drivers of life expectancy. Its coefficients give direct insights—for instance, quantifying how much an improvement in education or income can offset a higher mortality rate. The log-linear regression offers a nuanced perspective by focusing on relative (percentage) changes; it marginally improves model fit and addresses heteroscedasticity, but in our case its predictive performance was essentially on par with the standard MLR. The KNN regression stands out as a flexible method that can capture complex non-linear relationships without assuming a specific model form. It performed comparably to the linear models in terms of $R^2$ and error, confirming that our chosen features carry sufficient information for accurate predictions even under a non-parametric approach. In terms of model strengths, the linear models (especially MLR) provided both high accuracy and clear interpretability, making them valuable for drawing actionable insights. KNN's strength lies in its ability to fit non-linear patterns, but in this dataset the benefit over linear models was modest. Overall, the multiple linear regression emerged as the most balanced and interpretable model, achieving high explanatory power while clearly quantifying each variable's contribution to life expectancy.

## Conclusion

The analysis of global life expectancy data indicates that longevity is influenced by multiple socio-economic and health factors. Higher average years of schooling and greater GDP per capita are strongly associated with longer life expectancy, while high adult mortality rates and a heavy burden of HIV/AIDS are linked to shorter lifespans. Visual comparisons underscore that developed countries exhibit substantially higher median life expectancies (around 80 years) and less variability than developing countries, which show lower medians and wider dispersion. In predictive modeling, a simple univariate linear regression (e.g., life expectancy versus schooling) explained a moderate proportion of variance but was outperformed by multivariable approaches. The multiple linear regression model (incorporating schooling, GDP, healthcare expenditure, and related predictors) achieved a strong fit and yielded interpretable coefficient aligned with known influences. A log-linear model (using log-transformed GDP) confirmed these trends within a proportional framework. A K-nearest neighbors model delivered similar predictive accuracy without assuming linearity, although it lacked the straightforward interpretability of the linear models. Overall, these findings confirm that broad-based investments in education, healthcare, and economic development are critical to improving life expectancy across nations.

Future work should broaden the scope of predictors and methods to capture complex influences on longevity. Incorporating lifestyle factors (smoking, obesity, and physical inactivity) and environmental measures (air pollution and climate) would address additional known risk factors. Exploring advanced modeling techniques—such as ensemble methods (random forests, gradient boosting) and time-series forecasting (ARIMA, state-space models)—could uncover nonlinear relationships and temporal patterns. Robust validation (using k-fold cross-validation and hold-out test sets) is essential to ensure generalizable results and prevent overfitting. Improved feature engineering (including interaction terms, composite indices, and transformations) could further enhance predictive power. These next steps would provide a more comprehensive understanding of life expectancy drivers and yield more reliable forecasts for policy planning.