Introduction

In this project, we analyze the relationship between various socioeconomic factors and life expectancy across 192 countries. The dataset, collected by the World Health Organization (WHO), spans from the years 2000 to 2015 and includes key variables such as GDP (gross domestic product) per capita, alcohol consumption, healthcare spending, and mortality rates. Our goal is to build a neural network model to predict life expectancy based on these features and understand the most influential factors affecting longevity.

Hypothesis and Initial Exploration

One of the primary assumptions is that GDP per capita has a strong positive correlation with life expectancy. It is logical to assume that wealthier nations have better healthcare systems, higher standards of living, and improved access to medical services, leading to longer life spans.

GDP Analysis

An initial analysis of GDP distribution (Figure 1) shows that most countries have relatively low GDP per capita, with only a few outliers reaching extreme wealth levels. This observation aligns with economic disparities worldwide, where developing nations make up a significant portion of the dataset

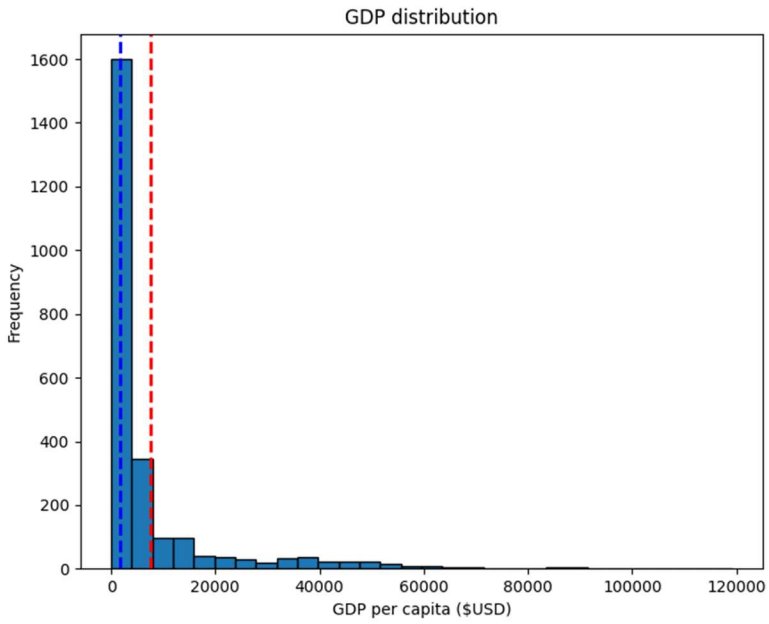

Figure 1. Distribution of GDP per capita for different countries.

Analyzing the Relationship Between GDP and Life Expectancy

It is expected that different features contribute with varying degrees of influence on life expectancy across countries. To begin our analysis, we examine the behavior of a few key variables to understand how they impact life expectancy. One of the most logical variables to analyze is Gross Domestic Product (GDP) per capita, as higher economic development is often associated with better healthcare, improved living conditions, and longer life spans.

We hypothesize that there is a positive correlation between GDP and life expectancy, meaning that countries with higher GDP per capita should have a longer average life expectancy. Our findings support this hypothesis, as shown in Figure 2, where we observe that as GDP increases, life expectancy also tends to rise.

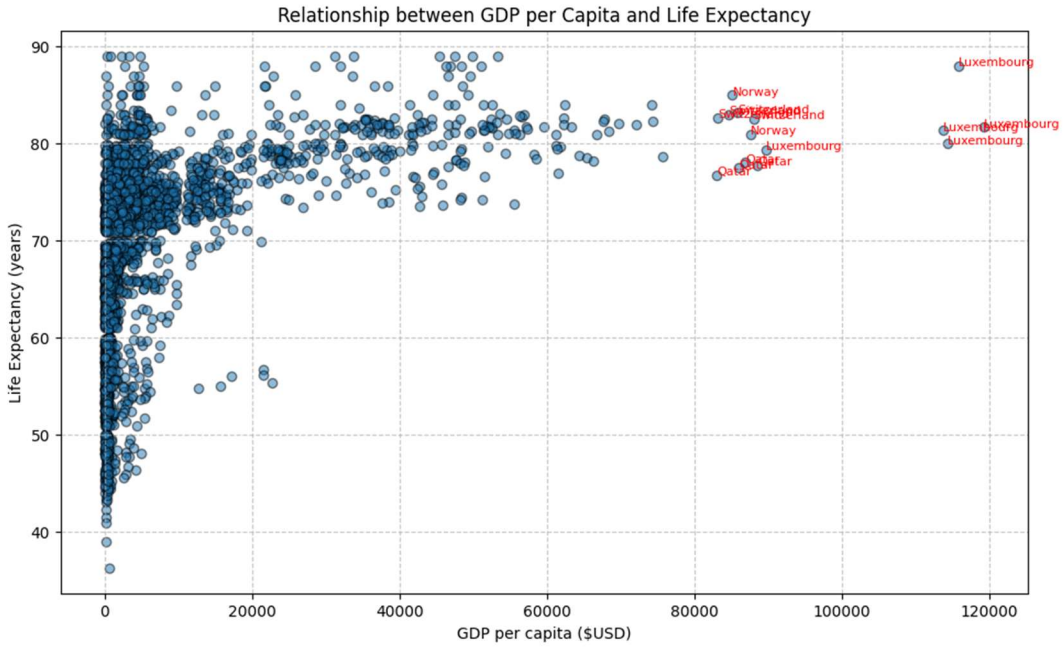

Figure 2. Relation between GDP and life expectancy.

However, this relationship is not strictly linear. From the scatter plot in Figure 2, we can identify a plateau effect occurring around $30,000 GDP per capita, where additional increases in GDP no longer contribute significantly to higher life expectancy. This suggests that beyond a certain economic threshold, other factors such as healthcare systems, lifestyle choices, and environmental conditions may play a more dominant role in determining longevity.

Additionally, we identify outliers (the points on the far right), notably Luxembourg, Switzerland, and Qatar, which exhibit exceptionally high GDP values (~100,000 USD+) but only marginal improvements in life expectancy compared to countries with significantly lower GDP. These cases suggest that once basic healthcare, nutrition, and living conditions are met, further economic gains may not substantially increase life expectancy.

Some of these rich countries may have small populations, which skews GDP per capita and most countries have GDP per capita below 60,000 USD.

We also found cluster of countries with Low GDP (~0 to 10,000 USD)

Most of the countries are in this range, indicating that many developing nations have lower GDP per capita. Life expectancy for these countries varies between ~40 and 75 years.

This analysis demonstrates that while economic prosperity is a strong predictor of life expectancy, its effects diminish at higher income levels, suggesting a diminishing return on health improvements beyond a certain GDP threshold.

Education and Life Expectancy: Correlation or Causation?

After analyzing GDP, another key variable to explore is education, specifically years of schooling. Education is often considered a fundamental factor in improving overall quality of life, but does it directly influence life expectancy?

Strong Positive Correlation

As observed in our data, countries with higher average years of schooling tend to have longer life expectancy. The scatter plot (Figure X) visually demonstrates this trend—countries with more schooling generally cluster in the higher life expectancy range (75-85 years).

However, while this suggests a correlation, it does not necessarily imply causation. The link between schooling and longevity could be indirect, driven by other socioeconomic factors such as:

- Economic Development: Wealthier nations tend to invest more in education, healthcare, and social services.
- Healthcare Access: More educated populations might make better health decisions, but access to quality healthcare plays a more direct role.
- Nutritional Awareness: Education may contribute to better nutrition choices, but food availability depends on economic stability.

Clustered Patterns

From the dataset, we observe distinct clusters:

Low schooling (~0-5 years): Countries in this group show a wide range of life expectancies (~40-75 years), suggesting that schooling alone does not determine longevity.

Higher schooling (~15-20 years): These countries generally exhibit higher life expectancy (~75-85 years).

Outliers and Interesting Cases

Some countries report 0 years of schooling but still have a life expectancy above 70 years. This could be due to strong informal education systems or robust healthcare policies despite limited formal education.

A few nations exceed 20 years of schooling, likely representing highly developed countries (e.g., Norway, Germany, USA).

While education strongly correlates with life expectancy, it is likely a reflection of broader economic development and healthcare infrastructure rather than a sole determining factor. Further statistical analysis, such as multiple regression models, could help isolate the effects of schooling while controlling for GDP and healthcare access.

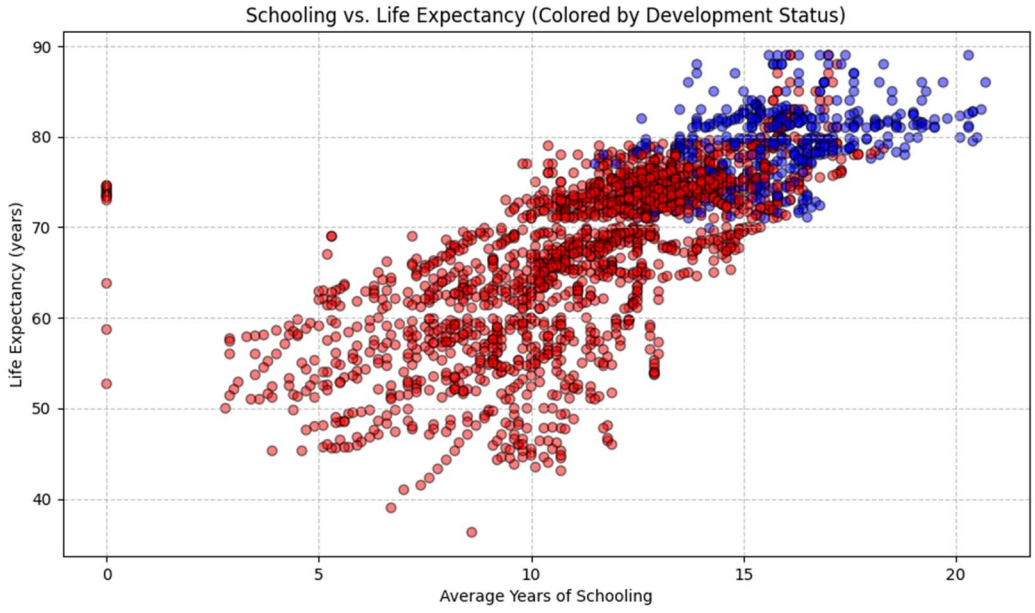

Figure 3. Correlation between years of schooling and life expectancy for different countries.

Top 10 higher education by country and its life expectancy

Country	Years of Schooling	Life Expectancy
Australia	20.7	86.0
Australia	20.6	83.0
Australia	20.5	79.9
Australia	20.4	82.8
Australia	20.4	79.5
Australia	20.4	82.7
New Zealand	20.3	89.0
Australia	20.3	82.5
Australia	20.3	81.2
Australia	20.3	81.0

Neural Network Predicting Life Expectancy

After analyzing individual variables and their correlations with life expectancy across different countries, we now take a more comprehensive approach by considering multiple factors simultaneously. Instead of evaluating each variable separately, we will use a neural network model to integrate all relevant features and predict life expectancy more accurately.

A neural network is a machine learning model inspired by the human brain, designed to recognize patterns in data and make predictions. By training on historical data, the model can learn complex relationships between different features and their impact on life expectancy. This approach allows us to capture nonlinear interactions that may not be evident when analyzing single variables independently.

In this study, we will train a supervised learning model using a neural network to predict the life expectancy of a country based on the following features:

Economic factors: GDP per capita, total health expenditure, income composition of resources

Health indicators: Adult mortality, child mortality, under-five deaths, HIV/AIDS prevalence, Hepatitis B, Diphtheria, Polio, Measles, BMI

Lifestyle factors: Alcohol intake, schooling, thinness (1-19 years and 5-9 years)

Demographic data: Population size, development status (developed or developing)

By training our neural network on this dataset, we aim to develop a model that can generalize patterns from the data and provide reliable predictions of life expectancy for different countries. This method will enable us to assess the collective influence of multiple variables and understand their combined impact on global health outcomes.

Data Preprocessing

Before training the neural network, several preprocessing steps were applied to the dataset to ensure consistency and reliability. First, missing data was handled by converting entries labeled as "N/A" or "Unknown" into NaN values, which were then replaced with the column mean to avoid data loss. The categorical variable "Status", which indicates whether a country is developed or developing, was mapped to numerical values, assigning 1 to developed countries and 0 to developing ones. Given that the dataset contained variables with different scales, MinMax Scaling was used to normalize all numerical features to a common range between 0 and 1, preventing any particular variable from disproportionately influencing the model. Additionally, the "Year" column was removed to prevent unintended bias, as it does not directly impact life expectancy.

Neural Network Model

With the data preprocessed, a feedforward neural network was implemented using TensorFlow and Keras. The model consisted of an input layer that takes in the selected features, followed by two hidden layers.

Input Layer: Features representing socioeconomic factors.

Hidden Layers: Layer 1: 64 neurons with ReLU activation.

Layer 2: 32 neurons with ReLU activation.

Output Layer: A single neuron with a linear activation function to predict life expectancy.

Optimization The model was optimized using the Adam optimizer, which adapts the learning rate dynamically and Mean Squared Error (MSE) was chosen as the loss function to minimize the difference between predicted and actual values.

The network was trained for 50 epochs with an 80-20 train-test split, ensuring that the model could generalize well to unseen data.

To evaluate the performance of the model, two key metrics were considered: Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE measures the average squared differences between actual and predicted values, penalizing larger errors more significantly, while MAE

provides a more interpretable measure by averaging the absolute differences between predictions and actual values. A lower MSE and MAE indicate better predictive accuracy, reflecting the model's ability to estimate life expectancy based on the selected socioeconomic and health-related features.

Model Evaluation

To assess the performance of our neural network, we used two key metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE). The MSE measures the average squared difference between actual and predicted values, penalizing larger errors more significantly, while MAE provides a more interpretable measure by averaging the absolute differences between predictions and actual values.

For the test dataset, the model achieved an MSE of **12.5228** and an MAE of **2.5818**, meaning that, on average, our predictions deviate by approximately **2.58 years** from the actual life expectancy. This suggests that the model is reasonably accurate in estimating life expectancy based on socioeconomic and health-related indicators.

The scatter plot in Figure 3 compares actual and predicted life expectancy values, where a strong diagonal pattern indicates a well-fitted model. Most predictions closely align with actual values, though some variance is observed for certain countries. The table below presents a sample of predicted and actual life expectancy values, further illustrating the model's performance of sample predictions.

Sample Country	Predicted	Actual
Country 1	71.448685	71.0
Country 2	68.943245	67.0
Country 3	70.699829	74.1
Country 4	70.558281	72.4
Country 5	56.067501	59.5
Country 6	80.979279	76.9
Country 7	77.946564	81.0
Country 8	60.952351	59.9
Country 9	73.832901	76.7
Country 10	74.081017	75.2

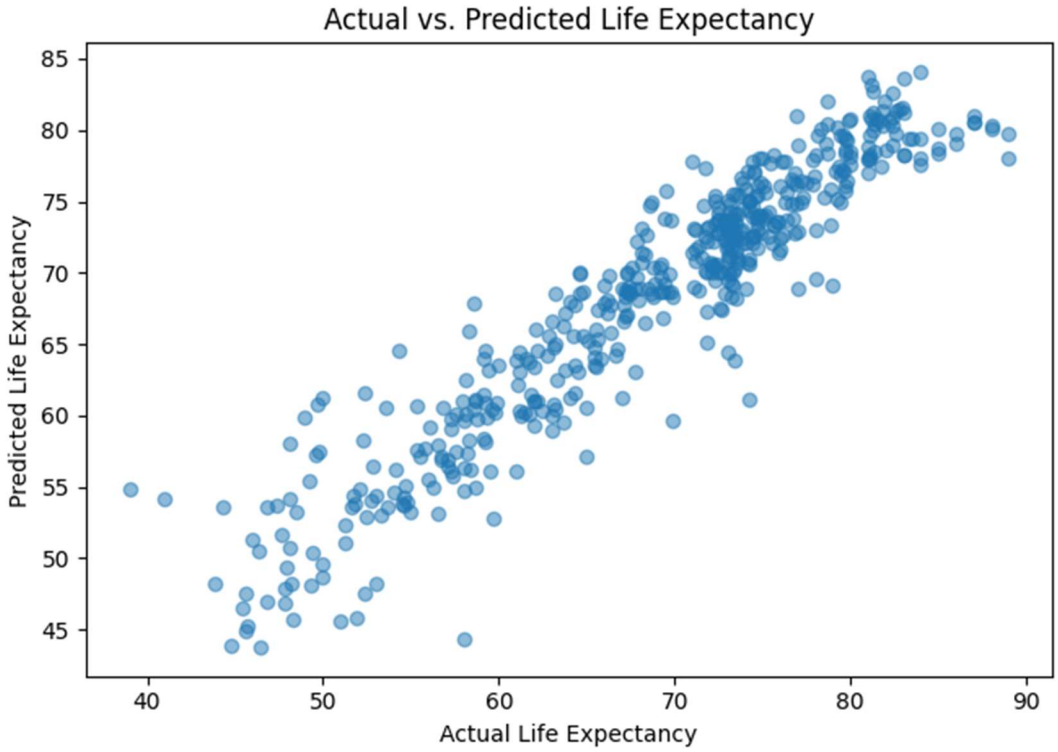

Figure 4. Comparison of Predicted and Actual Life Expectancy Values

The scatter plot in Figure 4 visualizes the relationship between actual and predicted life expectancy values. If the model were making perfect predictions, all points would lie exactly on the diagonal line where predicted values equal actual values). In this case, most points closely follow this diagonal pattern, suggesting that the model is generally performing well. However, some outliers deviate more significantly, indicating instances where the model struggled to predict life expectancy accurately.

A closer look at these deviations suggests that certain factors influencing life expectancy, such as healthcare quality, environmental conditions, or cultural aspects, may not be fully captured by the available features.

Potential Improvements & Future Work

While the model demonstrates a reasonable ability to predict life expectancy based on socioeconomic and health indicators, several improvements could enhance its accuracy:

Feature Engineering:

Adding more relevant features, such as access to clean water, healthcare quality, or disease prevalence, could improve predictions.

Transforming existing features (e.g., applying logarithmic scaling to GDP) might help capture nonlinear relationships.

Hyperparameter Tuning:

Adjusting the neural network's architecture (e.g., adding more layers or neurons).

Experimenting with different activation functions, learning rates, and optimizers.

Exploring Alternative Machine Learning Models:

While neural networks provide flexibility, simpler models like Random Forests or Gradient Boosting (e.g., XGBoost) might offer competitive performance with greater interpretability.

Final Thoughts

Overall, the neural network successfully captures major trends in life expectancy prediction, with most predictions aligning closely with actual values. However, refining feature selection, optimizing the model architecture, and incorporating more data-driven insights could further enhance predictive accuracy. This study provides a solid foundation for future work aimed at better understanding the complex factors influencing life expectancy worldwide.