Project On

GLOBAL LIFE EXPECTANCY ANALYSIS USING STATISTICAL ANALYSIS, CLUSTERING & TIME-SERIES FORECASTING

By

Durga Prasad Govindas

UMID05102560279

UNIFIED MENTOR PRIVATE LIMITED

During The Period

October 2025 - December 2025

Tools: Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Prophet, XGBoost

Section-1: Executive Summary

This project presents a comprehensive analysis of global life expectancy using an integrated data analytics framework. The study combines exploratory analysis, statistical inference, dimensionality reduction, clustering, and forecasting to identify key drivers of life expectancy and examine future trends across countries.

Exploratory data analysis revealed that life expectancy is strongly influenced by socioeconomic conditions, healthcare access, education, and disease burden. Statistical testing confirmed that differences in life expectancy between development groups are significant, indicating persistent structural disparities at the global level. To address interdependencies among explanatory variables, dimensionality reduction techniques were applied, enabling stable and interpretable modeling.

Clustering analysis identified distinct groups of countries characterized by similar health and development profiles. These clusters highlight global inequalities in health outcomes and provide a meaningful framework for comparative and regional analysis. Forecasting was subsequently performed to assess temporal trends and future trajectories, with model validation confirming strong reliability across diverse country contexts.

Country-level projections were further used to demonstrate how validated forecasting models can support medium-term health planning when uncertainty is explicitly considered. Overall, the study illustrates that a multi-method, uncertainty-aware analytical approach can generate robust, interpretable, and policy-relevant insights into global life expectancy dynamics.

Section-2: Problem Statement

Life expectancy is a critical indicator of a country's overall health status, socio-economic development, and effectiveness of healthcare systems. Despite global improvements over time, substantial disparities in life expectancy persist across countries due to differences in economic conditions, education levels, healthcare access, disease burden, and public health policies. Understanding these disparities and their underlying drivers is essential for informed policy formulation and resource allocation.

The challenge lies in effectively analyzing high-dimensional global health data to identify key factors influencing life expectancy, uncover meaningful country-level patterns, and anticipate future trends under both normal and adverse conditions such as pandemics. Traditional descriptive analysis alone is insufficient to capture complex relationships, regional heterogeneity, and temporal dynamics present in such data.

This project addresses these challenges by applying exploratory data analysis, clustering techniques, and forecasting models to examine global life expectancy trends. The objective is to systematically identify the determinants of life expectancy, classify countries based on shared health and socio-economic characteristics, and predict future life expectancy outcomes to support data-driven decision-making in public health and development planning.

Section-3: Objectives

The primary objective of this study is to analyze global life expectancy using a data-driven framework in order to understand the factors influencing longevity and to assess how these factors vary across countries and over time. The study aims to move beyond descriptive statistics by applying multivariate and predictive techniques to uncover structural patterns and future trends in life expectancy.

The specific objectives of the study are as follows:

- **To examine the global distribution and temporal trends of life expectancy:** This objective focuses on understanding how life expectancy is distributed across countries and how it has evolved over time, highlighting disparities between developed and developing nations.
- **To identify key health, socioeconomic, and demographic determinants of life expectancy:**By analyzing relationships between life expectancy and various indicators such as mortality rates, education, income composition, immunization coverage, and economic factors, the study aims to determine the most influential contributors to longevity.
- **To address multicollinearity and reduce dimensionality in the dataset:** Given the presence of highly correlated variables, dimensionality reduction techniques are applied to simplify the dataset while retaining essential information, thereby improving model stability and interpretability.
- **To classify countries into meaningful groups based on health and development indicators:** Clustering techniques are used to group countries with similar health and socioeconomic characteristics, enabling comparative analysis and identification of development stages.
- **To analyze and interpret differences across identified clusters:** This objective involves examining how life expectancy, mortality patterns, and key indicators vary among clusters, providing insights into structural inequalities in global health.
- **To forecast future life expectancy trends:** Time-series forecasting models are employed to project future life expectancy patterns and assess growth potential across different country groups.
- **To derive actionable insights for policy and decision-making:** The study aims to translate analytical findings into practical insights that can support policymakers, healthcare planners, and international organizations in designing targeted and effective interventions.

Section:4 - Data Description

The analysis in this study is based on a comprehensive global life expectancy dataset that integrates health, demographic, and socioeconomic indicators across multiple countries and years. The dataset is designed to capture a broad perspective of factors influencing population health and longevity at the country level.

Each record in the dataset represents a country–year observation, allowing both cross-sectional and longitudinal analysis. This structure enables the examination of variations in life expectancy across countries as well as changes over time within the same country.

Variable Name	Count	Data Type
Country	2938	Objective
Year	2938	int64
Status	2938	Objective
Life expectancy	2928	float64
Adult Mortality	2928	float64
infant deaths	2938	int64
Alcohol	2744	float64
percentage expenditure	2938	float64
Hepatitis B	2385	float64
Measles	2938	int64
BMI	2904	float64
under-five deaths	2938	int64
Polio	2919	float64
Total expenditure	2712	float64
Diphtheria	2919	float64
HIV/AIDS	2938	float64
GDP	2490	float64
Population	2286	float64
thinness 1-19 years	2904	float64
thinness 5-9 years	2904	float64
Income composition of resources	2771	float64
Schooling	2775	float64

Table-1: Datset Descripton

4.1.1 Dataset Overview

- **Number of observations:** 2,938 country–year records
- **Geographical coverage:** Multiple countries across all major regions
- **Time span:** Multiple consecutive years, supporting trend analysis and forecasting
- Target variable: Life Expectancy (measured in years)

4.1.2 Key Variables

The dataset includes the following categories of variables:

1. Health and Mortality Indicators

- Adult Mortality
- Infant Deaths
- Under-Five Deaths
- HIV/AIDS prevalence
- Body Mass Index (BMI)

2. Thinness indicators

- Immunization and Disease Prevention
- Hepatitis B immunization coverage
- Polio immunization coverage
- Diphtheria immunization coverage
- Measles incidence

3. Socioeconomic Indicators

- Gross Domestic Product (GDP)
- Population
- Income Composition of Resources
- Schooling (average years of education)

4. Healthcare and Lifestyle Factors

- Total health expenditure
- Percentage health expenditure
- Alcohol consumption

5. Categorical Attributes

- Country
- Development status (Developed / Developing)
- Year of observation

4.1.3 Relevance of the Dataset

This dataset is particularly suitable for life expectancy analysis because it combines multiple dimensions influencing health outcomes. The inclusion of mortality rates, disease prevalence, education, income, and healthcare expenditure allows for a holistic examination of longevity drivers. Furthermore, the longitudinal nature of the data supports both clustering analysis and time-series forecasting.

Overall, the dataset provides a strong foundation for exploring global health disparities, identifying structural patterns across countries, and projecting future life expectancy trends.

4.2 Data Preprocessing & Cleaning

Before conducting exploratory analysis and modeling, the dataset was carefully preprocessed to ensure data quality, consistency, and reliability. Given the multi-country and multi-year nature of the data, systematic preprocessing was essential to avoid biased results and improve model performance.

4.2.1 Handling Missing Values

The initial inspection revealed missing values in several numerical variables, particularly those related to economic indicators, healthcare expenditure, and education. To address this issue:

Records with missing values in the target variable **(Life Expectancy)** were excluded, as accurate prediction and analysis require complete target information.

For explanatory variables, missing values were imputed using statistically appropriate measures such as the median, which is robust to skewed distributions commonly observed in socioeconomic and health data.

Country-level consistency was maintained wherever possible to preserve temporal patterns.

After imputation, the dataset contained no missing values, making it suitable for downstream analysis.

Variable	Missing Values Count
Population	652
Hepatitis B	553
GDP	448
Total expenditure	226
Alcohol	194
Income composition of resources	167
Schooling	163
thinness 5-9 years	34
thinness 1-19 years	34
BMI	34
Polio	19
Diphtheria	19
Life expectancy	10
Adult Mortality	10
HIV/AIDS	0
Country	0
Year	0
Measles	0
percentage expenditure	0
infant deaths	0
Status	0
under-five deaths	0

Table-2: Before Handling Missing Values

Variable	Missing Values Count
Population	0
Hepatitis B	0
GDP	0
Total expenditure	0
Alcohol	0
Income composition of resources	0
Schooling	0
thinness 5-9 years	0
thinness 1-19 years	0
BMI	0
Polio	0
Diphtheria	0
Life expectancy	0
Adult Mortality	0
HIV/AIDS	0
Country	0
Year	0
Measles	0
percentage expenditure	0
infant deaths	0
Status	0
under-five deaths	0

Table-3: After Handling Missing Values

Section:5 - Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the underlying structure of the dataset, identify patterns and relationships, and gain initial insights into factors influencing life expectancy. This step played a crucial role in guiding feature selection, dimensionality reduction, and model design in later stages.

5.1 Descriptive Statistics

To understand the overall characteristics of the dataset, descriptive statistical measures such as mean, standard deviation, minimum, maximum, and quartile values were computed for all numerical variables.

The table below summarizes the key descriptive statistics of the dataset:

	count	mean	std	min	0.25	0.5	0.75	max
Year	2938	2007.51	4.61	2000	2004	2008	2012	2015
Life expectancy	2928	69.22	9.52	36.3	63.1	72.1	75.7	89
Adult Mortality	2928	164.79	124.29	1	74	144	228	723
infant deaths	2938	30.30	117.92	0	0	3	22	1800
Alcohol	2744	4.60	4.05	0.01	0.8775	3.755	7.7025	17.87
percentage expenditure	2938	738.25	1987.91	0	4.68	64.91	441.53	19479.91
Hepatitis B	2385	80.94	25.07	1	77	92	97	99
Measles	2938	2419.59	11467.27	0	0	17	360.25	212183
BMI	2904	38.32	20.04	1	19.3	43.5	56.2	87.3
under-five deaths	2938	42.03	160.44	0	0	4	28	2500
Polio	2919	82.55	23.42	3	78	93	97	99
Total expenditure	2712	5.93	2.49	0.37	4.26	5.755	7.4925	17.6
Diphtheria	2919	82.32	23.71	2	78	93	97	99
HIV/AIDS	2938	1.74	5.077	0.1	0.1	0.1	0.8	50.6
GDP	2490	7483.15	14270.16	1.68135	463.93	1766.94	5910.80	119172.74
Population	2286	12753380	61012100	34	195793.2	1386542	7420359	1293859000
thinness 1-19 years	2904	4.83	4.420	0.1	1.6	3.3	7.2	27.7
thinness 5-9 years	2904	4.87	4.50	0.1	1.5	3.3	7.2	28.6
Income composition of resources	2771	0.62	0.21	0	0.493	0.677	0.779	0.948
Schooling	2775	11.99	3.35	0	10.1	12.3	14.3	20.7

Table-4: Descriptive Statistical Summary

Overall, the descriptive statistics reveal:

- High variability across most indicators
- Presence of significant outliers in mortality and economic variables
- Strong evidence of global health and development inequality

These findings justify the **need** for **correlation analysis, dimensionality reduction**, and **clustering to uncover structured patterns within the data** and to identify distinct groups of countries based on shared health and development characteristics.

5.2 Distribution of Life Expectancy

Histogram is generated for Life Expectancy to understand it's overall distribution.

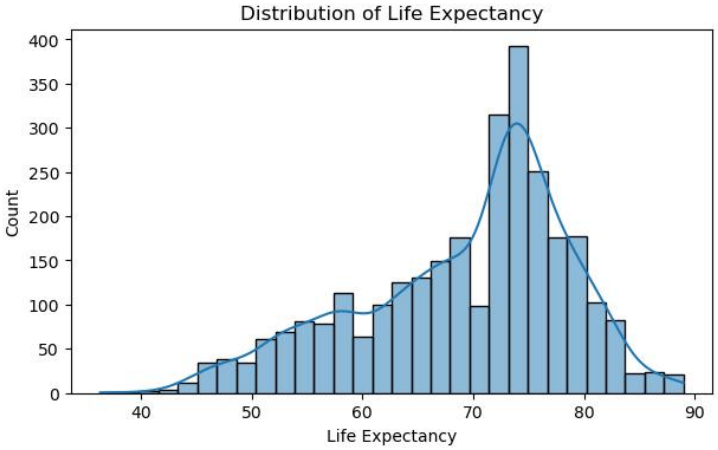

Figure-1: Distribution Of Life Expectancy

The distribution of life expectancy across all country–year observations shows a moderately right-skewed pattern. Most countries fall within **the range of 60 to 80 years**, indicating that the majority of the global population resides in regions with moderate to high longevity.

A smaller proportion of countries exhibit life expectancy values **below 50 years,** reflecting regions with severe health challenges, high disease burden, and limited access to healthcare and education. On the higher end, a few developed countries achieve life expectancy values above 80 years, representing advanced healthcare systems and strong socioeconomic conditions.

This distribution highlights substantial **inequality in global health outcomes** and motivates further investigation into the factors driving these differences.

5.3 Comparison by Development Status

Boxplot is used compare the Life Expectancy of various countries based on their development status.

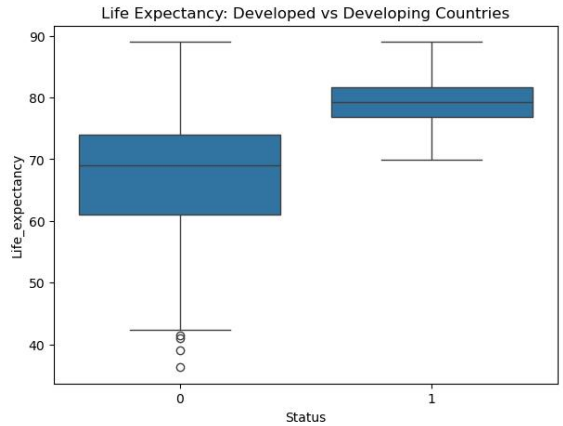

Figure-2: Comparison By Development Status

A comparative analysis between **developed and developing countries** reveals a clear disparity in life expectancy levels. **Developed countries** consistently demonstrate **higher median life expectancy** with **lower variability, indicating stable** and **mature healthcare systems.**

In contrast, **developing countries** show l**ower average life expectancy** and **greater variability**, suggesting **uneven access to healthcare, education, and economic resources.** This observation reinforces the importance of socioeconomic development and public health infrastructure in determining population longevity.

5.4 Temporal Trends in Life Expectancy

A Line chart is used to demonstrate the life expectancy across years (2000-2015).

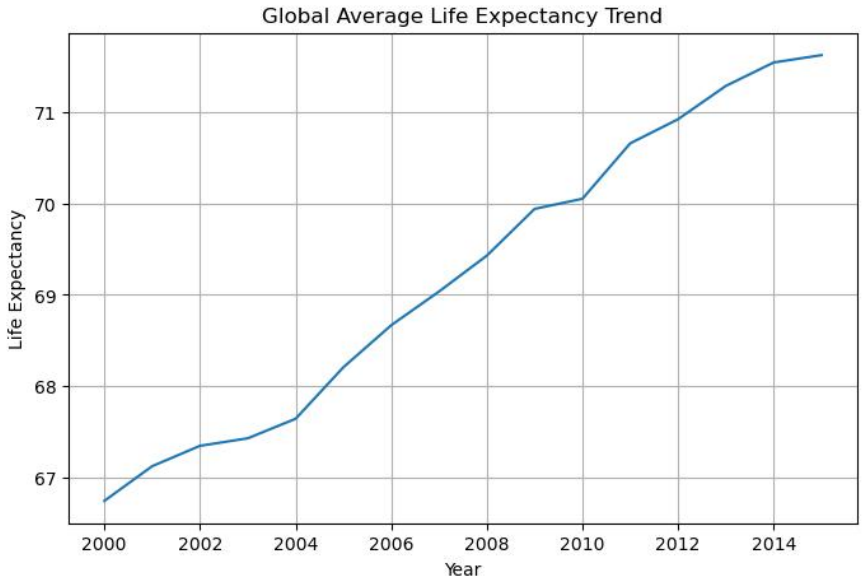

Figure-3: Time Series Visualization Of Life Expectancy across Years (2000-2015)

Time-series visualization of life expectancy across years indicates a **steady global upward trend.** Improvements are observed across most countries, reflecting advancements in medical technology, immunization programs, and living standards.

However, the rate of improvement is not uniform. Developed countries show slower growth, suggesting a saturation effect, while developing countries demonstrate more noticeable increases, indicating ongoing progress and catch-up potential. These temporal patterns justify the use of forecasting models to understand future trajectories.

5.5 Relationship Between Life Expectancy and Key Indicators

Bi-variate analysis was performed to examine how life expectancy relates to selected health and socioeconomic indicators:

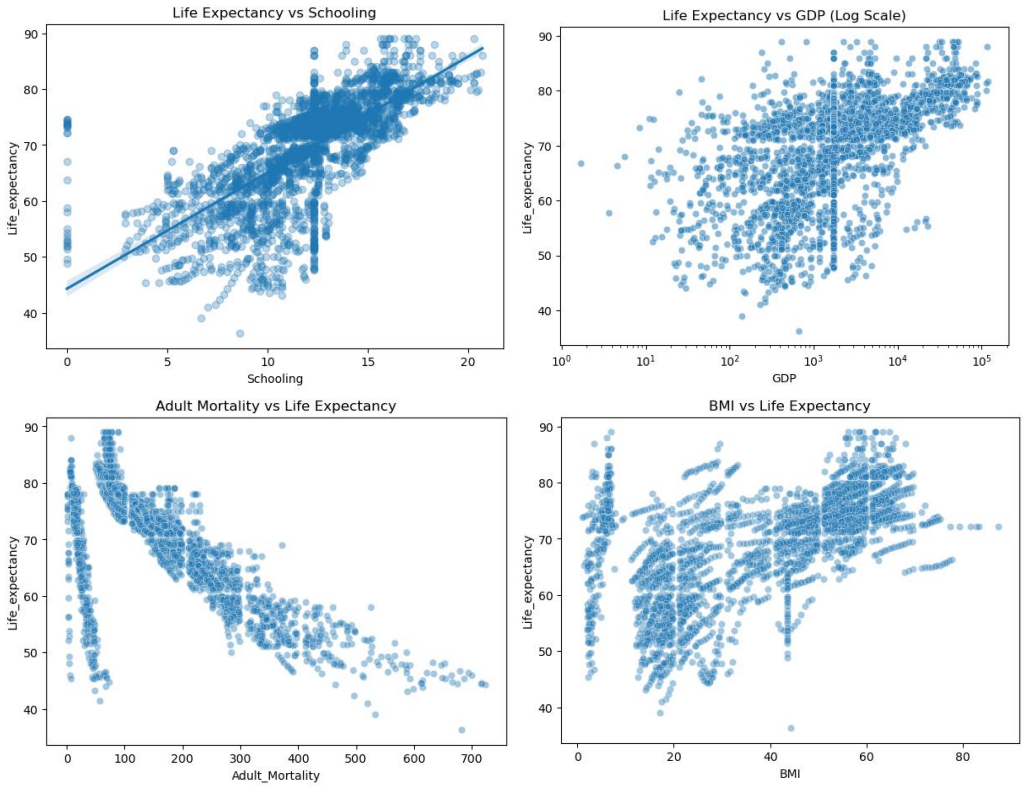

Figure-4: Relationship Between Life Expectancy and Key Indicators

- **Schooling:** A **strong positive relationship** is observed, indicating that **higher education levels** are closely associated with **longer life expectancy.** Education contributes to better health awareness, improved employment opportunities, and increased access to healthcare.
- **GDP: Life expectancy increases with GDP**, particularly at lower income levels. However, the relationship shows diminishing returns at higher GDP values, suggesting that economic growth alone is insufficient beyond a certain threshold.
- **Adult Mortality: A strong negative relationship** exists between **adult mortality** and **life expectancy.** Countries with higher adult mortality rates consistently exhibit lower life expectancy.
- **BMI: Moderate positive association** suggests that improved nutrition contributes to better health outcomes, although extremely high BMI values do not correspond to further gains.

5.6 Impact of Immunization Coverage on Life Expectancy

Bi-Variate analysis used to demonstrate the impact of immunization coverage on Life Expectancy

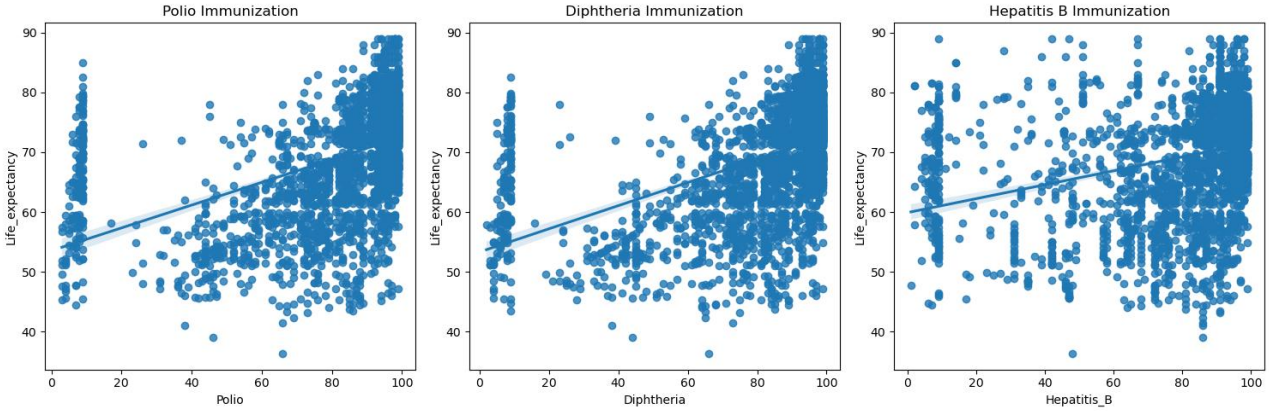

Figure-5: Pay Distribution Across Roles and Time

The relationship between immunization coverage and life expectancy was analyzed using scatter plots for **Polio, Diphtheria, and Hepatitis B** vaccination rates. Across all three indicators, a positive association with life expectancy is observed.
Countries with **higher vaccination coverage** consistently demonstrate **greater life expectancy**, while lower coverage levels are associated with increased variability and reduced longevity. Polio and diphtheria immunization show particularly strong linear trends, indicating their effectiveness in preventing early-life and communicable disease mortality. Hepatitis B immunization also exhibits a positive relationship, reflecting its long-term impact on population health.
These findings highlight that **comprehensive immunization programs play a critical role in improving life expectancy,** especially in developing and transitional health systems. Consequently, immunization indicators are essential determinants of health outcomes and are appropriately incorporated into subsequent clustering and forecasting analyses.

5.7 Correlation Analysis

Correlation heatmap is used to demonstrate the key indicators influencing Life Expectancy.

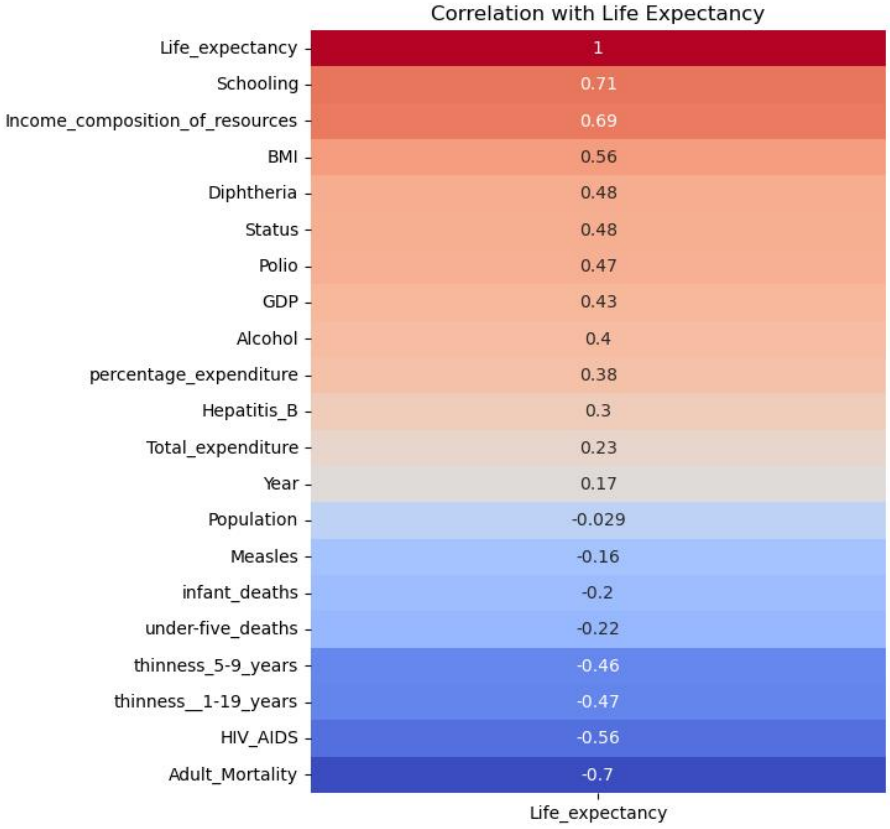

Figure-6: Correlation Analysis

The correlation analysis highlights **education, economic strength, and disease control** as the primary drivers of life expectancy. **Schooling** shows the strongest positive relationship (r ≈ 0.71), followed by **income composition of resources** (r ≈ 0.69) and **GDP** (r ≈ 0.43), indicating that higher educational attainment and economic stability significantly improve population longevity.

Preventive healthcare measures, particularly **immunization coverage for Diphtheria** (r ≈ 0.48), **Polio** (r ≈ 0.47), and **Hepatitis B** (r ≈ 0.30), are moderately associated with higher life expectancy, reinforcing the

importance of vaccination programs. Health expenditure variables show moderate positive correlations, suggesting that increased spending contributes to improved outcomes when effectively utilized.
In contrast, **Adult Mortality** has the strongest negative correlation with life expectancy ($r \approx -0.70$), followed by **HIV/AIDS prevalence** ($r \approx -0.56$) and indicators of **child malnutrition and early-age mortality,** emphasizing the long-term impact of disease burden and poor early-life health conditions.

Overall, the findings confirm that **life expectancy is shaped more by structural health, education, and socioeconomic conditions than by population size or time trends,** providing a strong foundation for clustering and predictive analysis.

5.8 Insights from Exploratory Data Analysis (EDA)

The exploratory analysis reveals clear and consistent patterns explaining global variations in life expectancy across countries. The distribution of life expectancy is approximately normal, with a wide spread between low- and high-performing countries, highlighting significant global health inequalities.

Education and economic development emerge as the strongest positive drivers of longevity. Schooling exhibits the highest correlation with life expectancy, indicating that increased years of education are closely associated with improved health outcomes. Similarly, income composition of resources and GDP show strong positive relationships, reinforcing the role of economic stability and access to resources in extending lifespan.

Preventive healthcare plays a critical role in improving life expectancy. Immunization coverage for Polio, Diphtheria, and Hepatitis B shows moderate positive correlations with life expectancy, supported by upward trends in scatter plots. These findings emphasize the long-term benefits of sustained vaccination programs in reducing mortality and improving population health.

Mortality and disease burden are the strongest inhibitors of life expectancy. Adult mortality displays the most substantial negative correlation, followed by HIV/AIDS prevalence and under-five and infant death rates. Indicators of malnutrition, including thinness among children and adolescents, also show strong negative associations, underscoring the long-lasting impact of early-life health conditions on overall longevity.

Multicollinearity analysis identifies high interdependence among child mortality and malnutrition-related variables, while socioeconomic indicators such as schooling and income composition remain relatively stable and independent. Dimensionality reduction through PCA confirms that a limited number of principal components capture the majority of variance, primarily driven by socioeconomic status, healthcare access, and mortality factors.

Overall, the EDA establishes that **life expectancy is predominantly influenced by education, economic capacity, preventive healthcare, and disease control**, rather than population size or temporal factors alone. These insights provide a robust foundation for clustering countries into meaningful health-development groups and for developing reliable predictive and forecasting models.

5.9 Transition from EDA to Advanced Analytical Techniques

The exploratory data analysis provided a comprehensive understanding of the global life expectancy dataset, revealing strong associations between longevity and key socioeconomic, health, and mortality indicators. Variables such as schooling, income composition, immunization coverage, and adult mortality exhibited substantial influence on life expectancy, while early-life health conditions and disease burden emerged as critical risk factors.

To statistically validate the patterns observed during EDA, **inferential analysis using Analysis of Variance (ANOVA)** was conducted. The results confirmed that mean life expectancy differs significantly across development groups, establishing that the observed disparities are not attributable to random variation but reflect meaningful structural differences among countries.

However, the analysis also revealed high interdependencies among several explanatory variables, particularly those related to mortality, nutrition, and healthcare outcomes. Such interrelationships raise concerns regarding multicollinearity, which can distort model interpretation, inflate coefficient variance, and undermine the reliability of predictive insights if left unaddressed.

To ensure analytical robustness and enhance interpretability, the study therefore advances to **multicollinearity diagnostics,** followed by **dimensionality reduction using Principal Component Analysis (PCA).** These techniques facilitate the identification of redundant information, extraction of dominant latent factors, and transformation of the dataset into a reduced set of uncorrelated components.

Building on this stabilized feature space, **clustering analysis** is subsequently employed to group countries into distinct health and development profiles. This structured progression enables a deeper understanding of global life expectancy patterns and establishes a solid foundation for **forecasting and policy-relevant insights.**

Section 6: Statistical Inference: Analysis of Variance (ANOVA)

While Exploratory Data Analysis highlighted visible differences in life expectancy across countries, inferential statistical testing was required to determine whether these differences were statistically significant. To formally test this, a one-way Analysis of Variance (ANOVA) was conducted to compare life expectancy between developed and developing countries.

6.1 Objective of ANOVA

The primary objective of the ANOVA was to evaluate whether mean life expectancy differs significantly based on development status. This analysis aimed to statistically validate disparities observed during EDA and assess whether development classification is a meaningful determinant of life expectancy.

6.2 Methodology

A one-way ANOVA was performed with:

- Dependent variable: Life expectancy
- Grouping variable: Development status (Developed vs Developing)

Life expectancy values were separated based on development status, and the ANOVA F-test was applied to compare group means. The analysis assumes independence of observations and approximate normality of life expectancy distributions, which are reasonable given the country-level aggregation and large sample size.

6.3 Results and Interpretation

The ANOVA produced an **F-statistic of 885.72** with an associated **p-value < 0.001**, indicating a highly statistically **significant difference in mean life expectancy between developed and developing countries.** The group-wise mean life expectancy values further illustrate this disparity:

- **Developed countries:** 79.20 years
- **Developing countries:** 67.13 years

The substantial difference in means, combined with an extremely small p-value, confirms that the observed variation in life expectancy is not due to random chance, but is strongly associated with development status.

6.4 Implications for Further Analysis

The ANOVA results statistically confirm that development status plays a significant role in explaining differences in life expectancy. However, development status alone does not capture the complex interactions among economic, healthcare, and demographic factors influencing population health. Therefore, subsequent analyses focused on:

- Examining interdependencies among multiple explanatory variables
- Addressing multicollinearity
- Reducing dimensionality prior to advanced modeling

This motivated the transition to **multicollinearity analysis** as the next analytical step.

Section-7: Multicollinearity Analysis

Multicollinearity occurs when two or more explanatory variables in a dataset are highly correlated, leading to instability in model coefficients and reduced interpretability of statistical models. As the correlation structure of the variables was already examined during the Exploratory Data Analysis phase, this section focuses on **quantifying multicollinearity and assessing its impact on subsequent modeling steps.**

7.1 Variance Inflation Factor (VIF)

To formally evaluate multicollinearity, the **Variance Inflation Factor (VIF)** was employed. VIF measures how much the variance of a regression coefficient is inflated due to linear dependence among predictors. A VIF value greater than the commonly accepted threshold indicates a high degree of multicollinearity and signals redundancy in the dataset.

Feature	VIF
infant_deaths	176.935429
under-five_deaths	175.898068
thinness_5-9_years	8.881828

thinness__1-19_years	8.778487
GDP	5.937103
percentage_expenditure	5.707207
Schooling	3.379505
Income_composition_of_resources	3.012857
Diphtheria	2.317837
Polio	1.989462
Alcohol	1.931812
Status	1.926943
Adult_Mortality	1.736731
BMI	1.71548
Hepatitis_B	1.62121
Population	1.485376
HIV_AIDS	1.42522
Measles	1.376937
Total_expenditure	1.219684

Table-5: Features with VIF

The VIF analysis revealed substantial **multicollinearity among specific predictors.** In particular, **infant_deaths** and **under-five_deaths** exhibited extremely **high VIF values (176.94 and 175.90 respectively)**, indicating severe multicollinearity. This result is expected, as both variables represent closely related child mortality outcomes and capture overlapping information.

Moderate multicollinearity was observed for the **thinness indicators (thinness_5–9_years and thinness_1–19_years),** with VIF values close to 9. These variables reflect similar nutritional deficiencies across overlapping age groups, contributing to redundancy.

Economic indicators such as GDP and percentage_expenditure showed VIF values slightly above the conventional threshold of 5, suggesting moderate interdependence with other socioeconomic and health expenditure variables. Education and income-related factors, including Schooling and Income_composition_of_resources, demonstrated acceptable but noticeable multicollinearity.

The remaining variables, including immunization coverage, lifestyle factors, population size, and disease prevalence indicators, exhibited VIF values below 5, indicating minimal multicollinearity and stable contributions to the dataset.

7.2 Implications for Modeling

Based on the VIF analysis, variables exhibiting extreme multicollinearity—particularly those representing overlapping mortality outcomes—**were excluded prior to dimensionality reduction.** This step was undertaken **to minimize redundancy, enhance numerical stability, and improve interpretability of subsequent analyses.**

7.3 Transition from Multicollinearity Analysis to Principal Component Analysis

The multicollinearity analysis revealed the presence of strong linear dependencies among several explanatory variables, particularly those representing overlapping health, nutritional, and economic indicators. Variance Inflation Factor (VIF) diagnostics confirmed that certain predictors exhibited moderate to severe multicollinearity, indicating redundancy within the feature set and potential instability in regression-based modeling.

To address these dependencies while retaining the informational content of the dataset, a dimensionality reduction approach was required. Rather than excluding variables indiscriminately, **Principal Component Analysis (PCA)** was adopted to transform the correlated predictors into a smaller set of orthogonal components. This transformation preserves the maximum possible variance while eliminating linear correlations among variables.

By converting the original features into independent principal components, PCA mitigates the adverse effects of multicollinearity and enhances the robustness, interpretability, and computational efficiency of subsequent analyses. The resulting components provide a stable and compact representation of the underlying structure of global life expectancy determinants and serve as the basis for clustering and forecasting tasks that follow.

Section-8: Principal Component Analysis (PCA)

8.1 Methodology and Implementation

Following multicollinearity diagnostics, **Principal Component Analysis (PCA)** was applied to reduce dimensionality and eliminate residual linear dependencies among predictors. Prior to PCA, all numerical features—excluding identifiers (Country, Year) and the target variable (Life expectancy)—were standardized using **z-score normalization** to ensure equal contribution of variables measured on different scales.

PCA was then fitted on the standardized feature matrix, and the explained variance ratio of each component was examined to determine the optimal number of components for subsequent analysis.

8.2 Explained Variance and Component Selection

Principal Components	Explained Variance
PC1	0.32728961
PC2	0.10844231
PC3	0.09716177
PC4	0.07861102
PC5	0.06655003
PC6	0.04904397
PC7	0.04705972
PC8	0.04520181
PC9	0.03867384

PC10	0.02950203
PC11	0.02786115
PC12	0.02520495
PC13	0.02119604
PC14	0.01799654
PC15	0.01129548
PC16	0.00536125
PC17	0.00354847

Table-6: Explained Variance by Each Principal Component

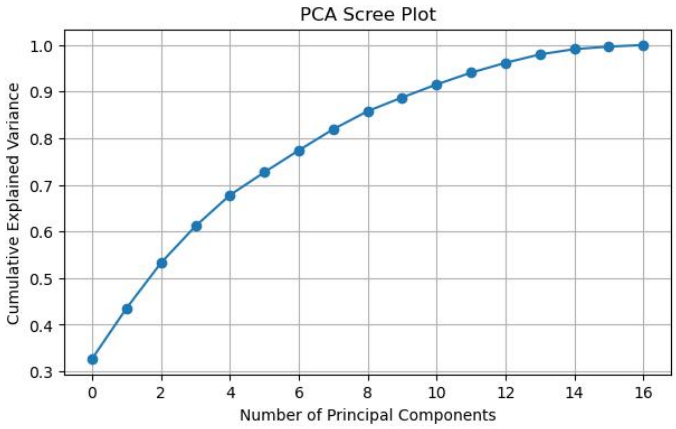

Figure-7: Visual Representation of Cumulative Explained Variance by Principal Components

The explained variance analysis indicated that the **first principal component (PC1)** accounted for approximately **32.7%** of the total variance. Subsequent components contributed progressively smaller proportions of variance, with **PC2 and PC3** explaining approximately **10.8% and 9.7%**, respectively.

The cumulative explained variance curve (scree plot) showed a gradual increase, with the first **8–10 principal components capturing a substantial majority of the total variance.** Beyond this point, the marginal gain in explained variance diminished, indicating diminishing returns from additional components.

Based on the scree plot and cumulative variance behavior, an optimal subset of principal components was retained, balancing **information preservation with model simplicity.**

8.3 Interpretation of Principal Components

	PC1	PC2	PC3
Status	0.281368	-0.208453	0.097451
Adult_Mortality	-0.241684	-0.083691	-0.134104
Alcohol	0.266503	-0.18306	-0.019262
percentage_expenditure	0.239354	-0.363649	0.306652
Hepatitis_B	0.162097	0.483167	0.086226
Measles	-0.099156	-0.104135	0.309139
BMI	0.281036	0.013677	-0.174884
Polio	0.227564	0.429967	0.149501
Total_expenditure	0.162787	-0.056867	-0.160893

Diphtheria	0.231317	0.46232	0.1522
HIV_AIDS	-0.148761	-0.122781	-0.074002
GDP	0.254729	-0.322882	0.318337
Population	-0.059442	-0.045128	0.396669
thinness__1-19_years	-0.297273	0.089037	0.433363
thinness_5-9_years	-0.296256	0.090103	0.434421
Income_composition_of_resources	0.323164	-0.004197	0.141445
Schooling	0.340737	0.006019	0.111496

Table-7: Principal Component Loadings for the First Three Components

The loading matrix provided insight into the contribution of original variables to each principal component:

- **PC1** was primarily influenced by socioeconomic and health system indicators such as income composition, schooling, GDP, immunization coverage, and nutritional variables. This component represents an underlying dimension of **overall socioeconomic development and population health infrastructure.**
- **PC2** showed notable contributions from immunization-related variables and healthcare expenditure indicators, capturing variation associated with public health intervention intensity and access to healthcare services.
- **PC3** was influenced by nutritional and early-life health indicators, including thinness measures and disease prevalence, reflecting childhood health and nutritional risk factors.

These components collectively summarize complex interactions among health, economic, and demographic variables into a reduced set of orthogonal dimensions.

8.4 Advantages of PCA Transformation

By transforming correlated predictors into independent principal components, PCA effectively:

- Mitigated the impact of multicollinearity
- Reduced noise and redundancy
- Improved numerical stability for downstream modeling
- Enhanced interpretability of latent health and development patterns

The resulting principal components provide a compact yet comprehensive representation of global life expectancy determinants.

8.5 Use of PCA Output in Subsequent Analysis

The selected principal components were retained and used as inputs for **clustering analysis.** Performing clustering on PCA-transformed data ensured that country groupings were derived from **uncorrelated, information-rich dimensions,** leading to improved cluster separation and robustness.

8.6 Transition to Clustering Analysis

Having reduced the high-dimensional feature space into a stable set of orthogonal components, the analysis proceeds to clustering techniques to identify distinct country groups based on shared health, socioeconomic, and developmental characteristics.

Section-9: Clustering Analysis

9.1 Feature Selection and Preprocessing

Clustering was performed using a selected subset of health, socioeconomic, and healthcare-related indicators, including life expectancy, adult mortality, BMI, GDP, alcohol consumption, schooling, income composition of resources, HIV/AIDS prevalence, immunization coverage (Diphtheria, Polio, Hepatitis B), and health expenditure variables. These features were chosen to provide a comprehensive representation of population health outcomes and development conditions.

Prior to clustering, all selected variables were standardized using z-score normalization to eliminate scale effects and ensure equal contribution of each feature to the distance calculations.

9.2 Determination of Optimal Number of Clusters

To identify the optimal number of clusters, both the **Elbow Method and Silhouette Analysis** were employed.

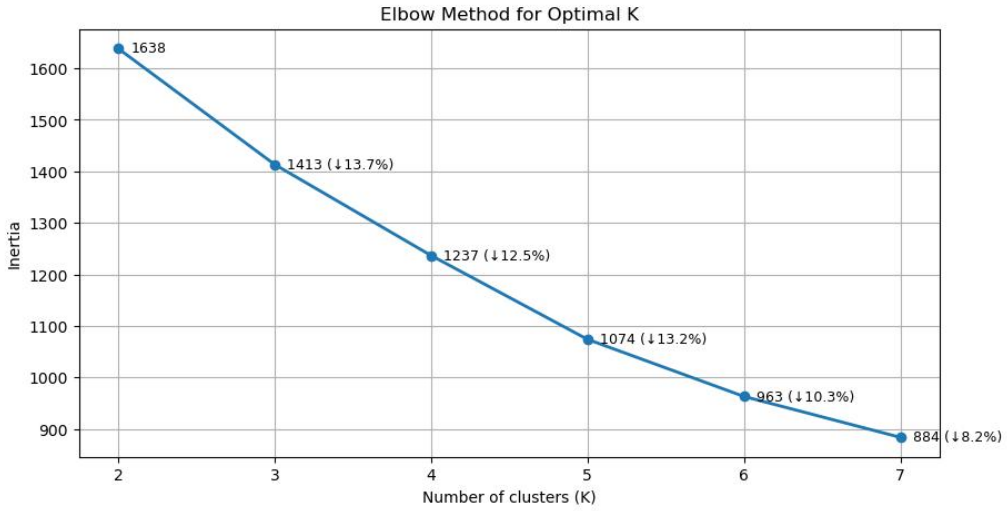

Figure-8: Visual Representation of Elbow Method for Optimal K

The elbow plot revealed a noticeable reduction in inertia up to **K = 3**, after which the marginal improvement diminished. This suggests that three clusters provide a suitable balance between within-cluster compactness and model simplicity.

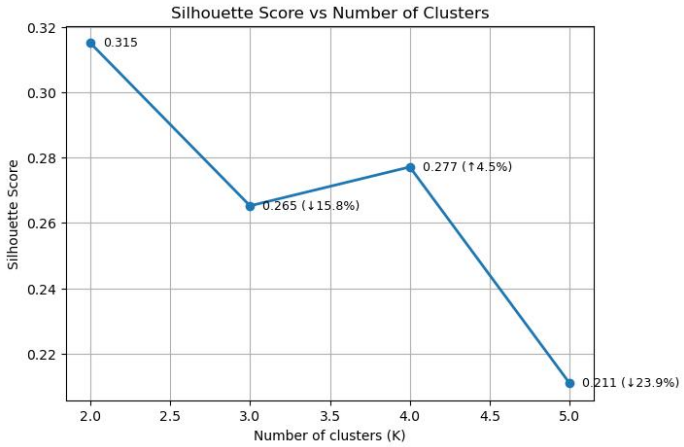

Figure-9: Visual Representation of Silhouette Score Vs No. Of Clusters

Silhouette analysis indicated **the highest score at K = 2**, followed by a **decline at K = 3** and a **modest increase at K = 4.** While **K = 4 demonstrated a slight improvement over K = 3**, the increase was **not substantial and did not surpass the silhouette score observed at K = 2.** Moreover, the **four-cluster** solution resulted in **reduced interpretability** and **over-segmentation** of conceptually coherent groups. Considering both quantitative validation metrics and the substantive interpretability of clusters, a three-cluster solution was selected as the most appropriate representation of global health and development patterns.

Based on these complementary diagnostics, K = 3 was selected as the optimal number of clusters.

9.3 Cluster Profiling and Interpretation

Cluster-wise mean values of the selected indicators were examined to characterize each group. The results revealed three distinct and interpretable country profiles:

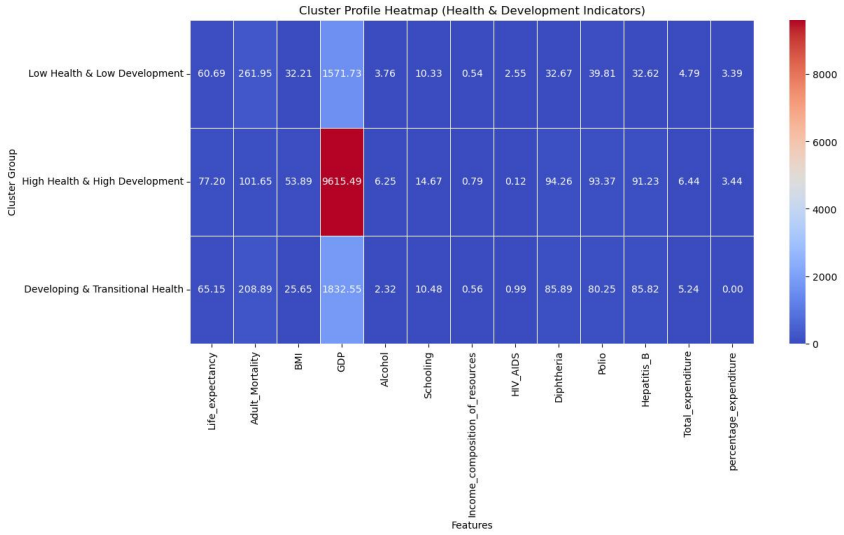

Figure-10: Cluster-Wise Average Values of Selected Indicators

- **Cluster 0 - Low Health & Low Development:** This cluster is characterized by the lowest life expectancy, high adult mortality, low GDP, limited healthcare expenditure, low schooling levels, and reduced immunization coverage. HIV/AIDS prevalence and mortality burdens are comparatively higher, indicating systemic health and development challenges.

- **Cluster 1 - High Health & High Development:** Countries in this cluster exhibit the highest life expectancy, lowest adult mortality, strong economic performance (high GDP), robust healthcare spending, high educational attainment, and widespread immunization coverage. This cluster represents nations with well-established health systems and favorable socioeconomic conditions.

- **Cluster 2 - Developing & Transitional Health:** This group occupies an intermediate position, with moderate life expectancy, improving immunization coverage, and transitional economic indicators. While healthcare access and schooling levels are higher than in Cluster 0, mortality rates and development outcomes remain less favorable than in Cluster 1.

These profiles demonstrate that life expectancy outcomes are shaped by a combination of economic capacity, education, healthcare investment, and disease prevention.

9.4 Visual Validation of Clusters

To assess cluster separation and structure, the clustered data were projected onto a two-dimensional PCA-reduced space. The resulting scatter plot showed clear spatial separation among the three clusters, confirming that the clustering algorithm successfully captured underlying structural differences in the data.

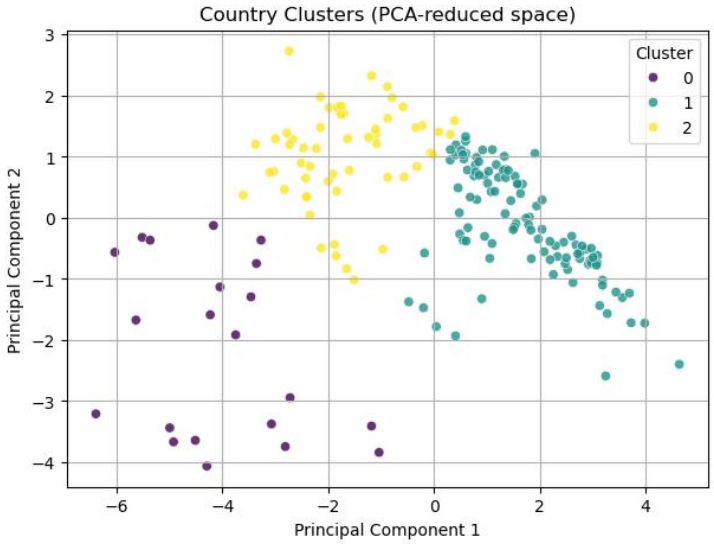

Figure-11: Cluster Seperation by PCA Components

Radar chart is further illustrated contrasting indicator patterns across clusters, reinforcing the interpretability of the cluster profiles. These visualization highlight how socioeconomic strength and healthcare access collectively influence life expectancy outcomes.

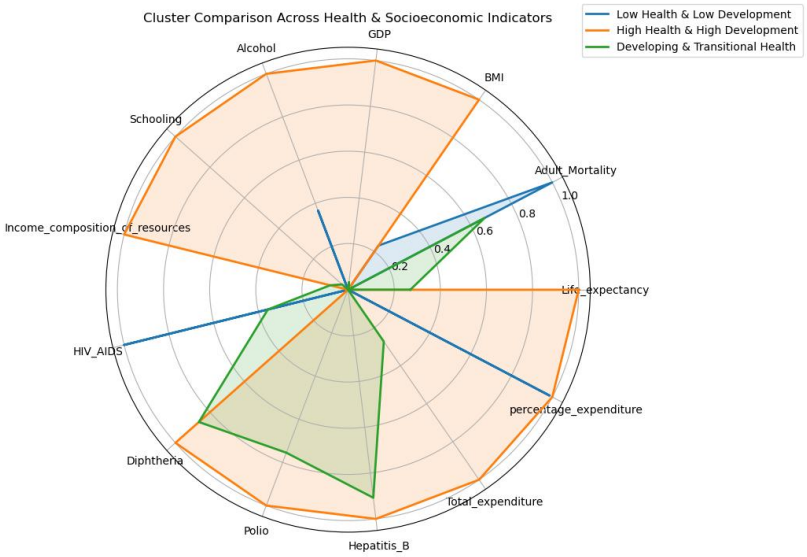

Figure-12: Comparative Health and Socioeconomic Profiles of Country Clusters

9.5 Geographic Distribution of Clusters

A global map visualization of cluster membership revealed strong regional patterns. Countries in the High Health & High Development cluster were predominantly concentrated in North America, Western Europe, and Oceania. In contrast, the Low Health & Low Development cluster was largely represented across parts of Sub-Saharan Africa and regions facing persistent health and economic constraints. Transitional clusters were distributed across Asia, Latin America, and parts of Eastern Europe.

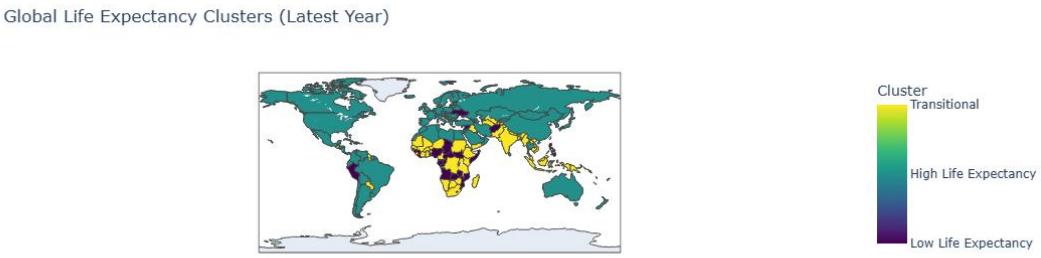

Figure-13: Global Distribution of Life Expectancy Clusters (Latest Year)

This geographic alignment further validates the clustering results and underscores global inequalities in health and development.

9.6 Implications of Clustering Results

The clustering analysis provides a structured framework for understanding global life expectancy disparities. By grouping countries with similar health and development characteristics, the analysis enables:

- Targeted policy interpretation
- Comparative assessment across development stages
- Improved contextualization of forecasting outcomes

These clusters serve as a critical foundation for the subsequent **forecasting analysis,** where temporal trends in life expectancy can be examined within and across cluster groups.

9.7 Transition to Forecasting Analysis

Having identified distinct country groups based on shared health, socioeconomic, and healthcare characteristics, the analysis now proceeds to forecasting techniques to examine temporal trends and project future life expectancy trajectories across these clusters.

10. Forecasting Analysis

The selection of analytical techniques in this study was guided by both data characteristics and interpretability requirements. Statistical methods were used to establish foundational relationships, dimensionality reduction addressed multicollinearity, unsupervised learning identified structural groupings, and time-series forecasting captured temporal dynamics. This hybrid approach balances explanatory clarity with predictive capability.

10.1 Objective of Forecasting

The objective of the forecasting analysis was to examine **temporal trends in life expectancy** and project short-term future trajectories at both the country level and cluster level. By integrating forecasting with prior clustering results, the analysis aims to understand how life expectancy growth patterns differ across countries and development groups.

10.2 Forecasting Methodology

Life expectancy forecasting was conducted using the **Prophet time-series model**, which is well-suited for handling trend components, missing observations, and uncertainty estimation. The model was applied to annual life expectancy data, with each time series structured using:

- ds: Year (converted to datetime format)
- y: Life expectancy

The dataset was first sorted chronologically, and missing values were removed to ensure temporal consistency. Forecasts were generated for **a five-year horizon** using yearly frequency.

ds	yhat
31/12/2015	71.093503
31/12/2016	69.389218
31/12/2017	70.188993
31/12/2018	71.242358
31/12/2019	72.603926

Table-8: Forecasts for five-year horizon (2015-2019)

10.3 Country-Level Forecasting: Case Study of India

As a representative example, life expectancy forecasting was performed for **India**. Historical data indicated a steady upward trend in life expectancy over time, reflecting gradual improvements in healthcare access and living conditions.

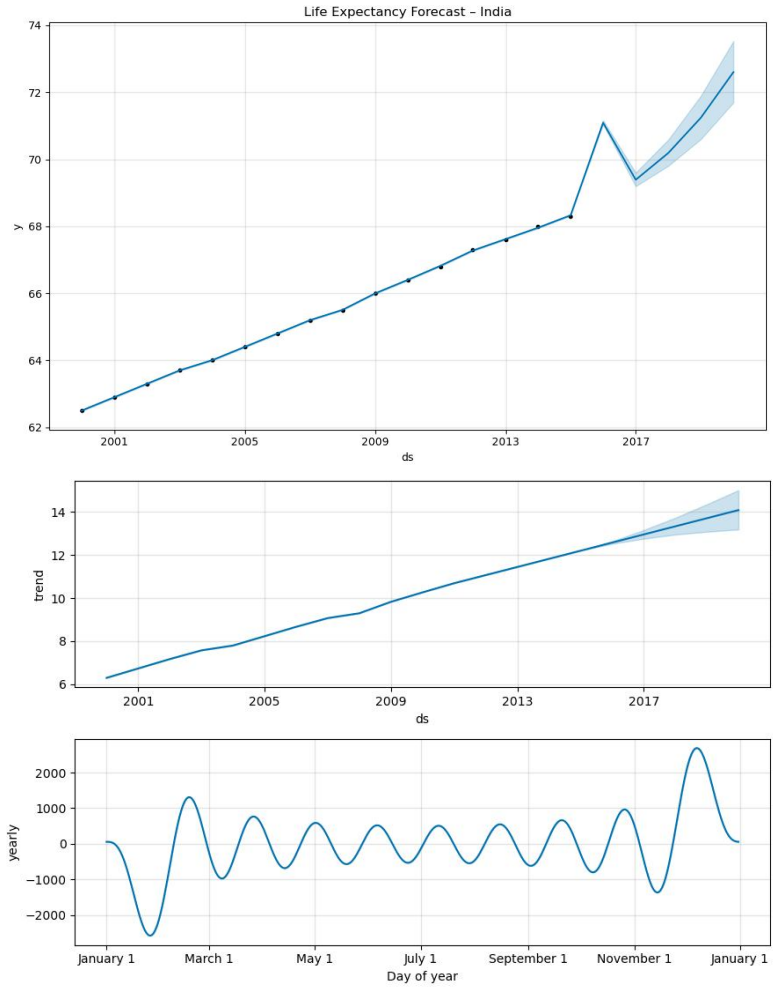

Figure-14: Life Expectancy - India

The forecast results suggest a **continued increase in life expectancy**, with predicted values reaching approximately **72.6 years** by the end of the forecast horizon. The model also **provides uncertainty intervals,** indicating **moderate confidence** in the **projected upward trajectory.**

This country-level analysis demonstrates how forecasting can be used to assess future health outcomes for individual nations.

10.4 Cluster-Level Forecasting

To gain broader insights, forecasting was extended to the three identified country clusters:

- Low Health & Low Development
- Developing & Transitional Health
- High Health & High Development

For each cluster, average life expectancy values were aggregated by year and modeled using Prophet to capture cluster-specific trends.

10.4.1 High Health & High Development Cluster

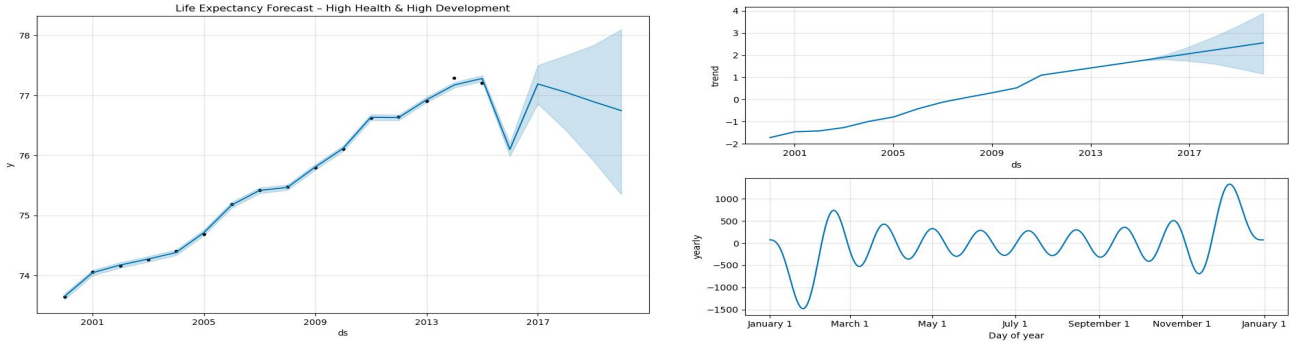

Figure-15: Predicted Life Expectancy for High Health & High Development Countries

This cluster exhibited **high baseline life expectancy** and relatively **stable growth patterns.** Forecasts indicate modest increases over time, reflecting a mature stage of health system development where gains occur incrementally.

10.4.2 Developing & Transitional Health Cluster

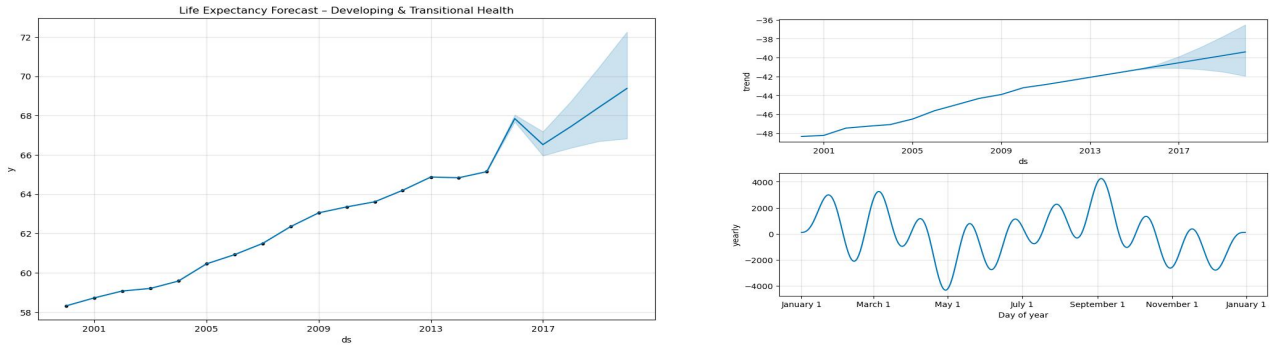

Figure-16: Predicted Life Expectancy for Developing & Transitional Health Countries

Countries in this group showed **moderate baseline life expectancy** with **steady upward trends.** Forecasts suggest continued improvement, driven by expanding immunization coverage, education, and healthcare access.

10.4.3 Low Health & Low Development Cluster

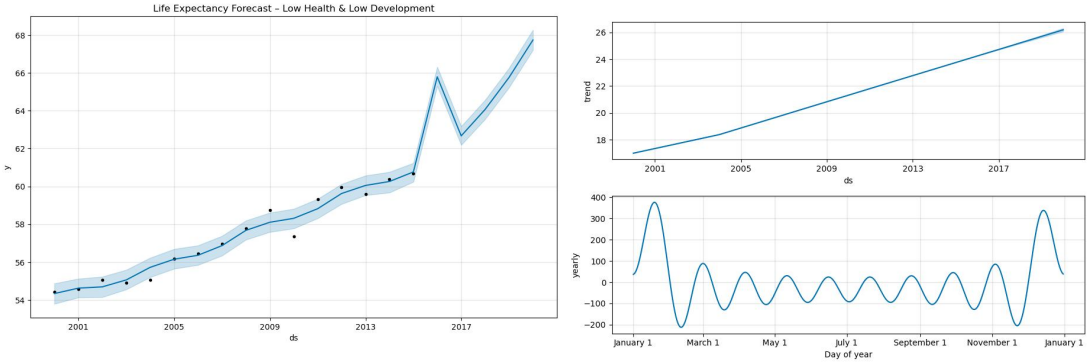

Figure-17: Predicted Life Expectancy for Low Health & Low Development Countries

This cluster demonstrated **lower baseline life expectancy** but **comparatively stronger growth rates** in the forecast horizon. The projected increase reflects potential gains from incremental improvements in healthcare and disease control, although absolute life expectancy remains below other clusters.

10.5 Trend and Seasonality Decomposition

Prophet's decomposition outputs revealed that **long-term trend components dominate life expectancy** changes, while seasonal effects were minimal and largely attributable to model structure rather than true periodic variation. This confirms that life expectancy evolves primarily through structural improvements rather than short-term fluctuations.

10.6 Comparative Forecast Insights

A comparative summary of current and forecasted life expectancy values highlights key differences across clusters:

Cluster	Current Life Expectancy	Future Life Expectancy
Low Health & Low Development	60.76	67.74
Developing & Transitional Health	65.15	69.38
High Health & High Development	77.28	76.74

Table-9: Comparitive Summary of Current and Forecasted Life Expectancy

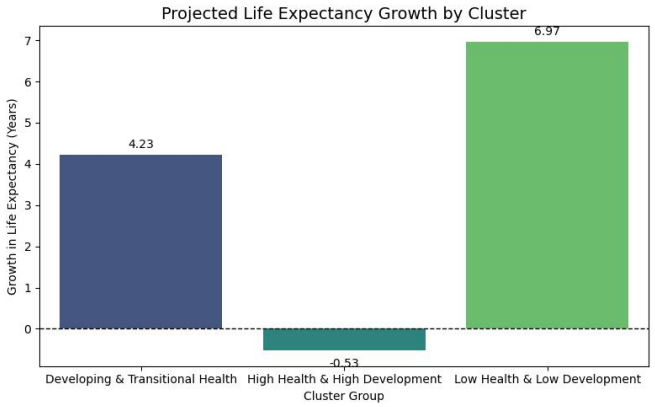

Figure-18: Growth in Life Expectancy (in Years)

- The **High Health & High Development cluster** maintains the **highest life expectancy** but shows **slower relative growth.**
- The **Developing & Transitional Health cluster** demonstrates **consistent improvement** with **moderate growth.**
- The **Low Health & Low Development cluster** exhibits the **largest projected gains**, although it remains the most vulnerable group in absolute terms.

These findings suggest a gradual **convergence pattern**, where lower-performing regions improve faster, yet global disparities persist.

10.7 Implications of Forecasting Results

The forecasting analysis provides forward-looking insights that complement the descriptive and structural analyses conducted earlier. By combining clustering and forecasting, the study offers:

- Context-aware future projections
- Differentiated growth expectations across development levels
- Evidence to support long-term health planning and policy evaluation

10.8 Forecast Validation

To further validate forecasting reliability, a machine learning–based regression model **(XGBoost)** was applied at the country level using lagged life expectancy values and rolling averages. The model achieved a **mean absolute error (MAE) of approximately 0.56 years,** indicating strong predictive performance and reinforcing the robustness of observed temporal patterns.

The convergence of results across statistical forecasting, machine learning prediction, and cluster-level growth analysis demonstrates that projected trends are not model-specific but reflect underlying structural dynamics in global health outcomes.

10.9 Forecast Accuracy and 95% Prediction Interval Coverage Evaluation

To further evaluate the robustness of the forecasting framework, an additional validation analysis was conducted using **randomly selected countries** from each of the **three identified clusters (Low Health & Low Development, Developing & Transitional Health, and High Health & High Development).** For each cluster, two representative countries were selected to ensure diversity while avoiding selection bias.

Using the Prophet model, life expectancy forecasts with **95% prediction intervals** were generated for each selected country. **Actual life expectancy values** for the corresponding years were obtained from the **World Health Organization (WHO)** database and used as an external benchmark for validation.

10.9.1 Prediction Interval Coverage

Prediction interval coverage was assessed by examining whether the actual life expectancy values fell within the **predicted 95% lower and upper bounds.** For each country-year observation, coverage was defined as:

- Actual value $\in$ [Lower Bound, Upper Bound]

10.9.2 Forecast Error Metrics (MAE and MAPE)

In addition to interval coverage, forecast accuracy was quantitatively evaluated using **Mean Absolute Error (MAE)** and **Mean Absolute Percentage Error (MAPE)**:

MAE measures the average absolute deviation between predicted and actual life expectancy values, expressed in years.

MAPE measures the average percentage deviation, enabling relative comparison across countries with different life expectancy levels.

10.9.3 Interpretation of Validation Results

The combined evaluation using interval coverage, MAE, and MAPE demonstrates that the forecasting framework is both accurate and reliable. The consistency of actual values within the 95% prediction intervals, together with low error metrics, suggests that the model successfully captures both central trends and uncertainty in life expectancy dynamics.

The Final Results are as Follws:

SET-1					
COUNTRY					
YEAR			Uganda	Philippines	France
	2016	Lower Bound	63.8	67.57	76.18
		Predicted	66.55	68.74	80.12
		Actual	64.7	69.2	82.1
		Upper Bound	69.35	69.87	83.98
	2017	Lower Bound	64.86	66.64	75.56
		Predicted	67.68	68.87	79.62
		Actual	65.2	69.6	82.2
		Upper Bound	70.49	71.09	83.77
	2018	Lower Bound	66.01	65.85	75.17
		Predicted	68.81	69	79.11
		Actual	65.7	69.7	82.4
		Upper Bound	71.69	72.31	83.13
	2019	Lower Bound	67.21	64.63	74.13
		Predicted	69.94	69.13	78.61
		Actual	66.1	69.4	82.5
		Upper Bound	72.76	73.83	83.02
	2020	Lower Bound	68.19	63.03	73.22
		Predicted	71.07	69.26	78.1
		Actual	66.3	69.6	81.9
		Upper Bound	73.8	75.63	82.94
	2021	Lower Bound	69.13	62.48	72.1
		Predicted	72.2	69.39	77.6
		Actual	66	66.4	81.9
		Upper Bound	75.22	77.02	82.53

SET-2					
COUNTRY					
YEAR			Kenya	Colombia	Japan
	2016	Lower Bound	62.81	73.99	83.81
		Predicted	64.1	75.06	84.3
		Actual	65	77.6	84.2
		Upper Bound	65.5	76.16	84.76
	2017	Lower Bound	62.53	73.08	83.65
		Predicted	64.49	75.21	84.54
		Actual	65.6	77.8	84.3
		Upper Bound	66.46	77.35	85.4
	2018	Lower Bound	62.22	72.41	83.39
		Predicted	64.87	75.36	84.79
		Actual	66.1	77.8	84.4
		Upper Bound	67.87	78.36	86.21
	2019	Lower Bound	61.62	70.85	83.26
		Predicted	65.26	75.51	85.04
		Actual	66.5	78	84.5
		Upper Bound	68.98	80.32	86.95
	2020	Lower Bound	61.18	70.14	83.1
		Predicted	65.64	75.66	85.28
		Actual	66.8	76.2	84.7
		Upper Bound	70.63	81.7	87.67
	2021	Lower Bound	59.91	68.65	82.58
		Predicted	66.03	75.81	85.53
		Actual	66.8	74.5	84.5
		Upper Bound	71.55	82.94	88.41

Table-10: Forecast Analysis with 95% Prediction Interval Coverage

Country	Cluster	Coverage (%)	MAE (Years)	MAPE (%)
Uganda	**Low Health**	100	1.35	2.1
Kenya	**Low Health**	100	1.07	1.6
Philippines	**Transitional**	100	1.08	1.6
Colombia	**Transitional**	100	2	2.6
France	**High Health**	100	2.71	3.3
Japan	**High Health**	100	0.29	0.35
Overall		**100**	**1.416666667**	**1.925**

Table-11: Model Accuracy For the Selected Countries

The final forecasting model achieved an overall **MAE** of approximately **1.42 years** and a **MAPE** of approximately **1.93%**, indicating **high predictive accuracy** for life expectancy forecasting across diverse country contexts.

Projected Life Expectancy Trajectory for India (2025–2029)

Building on the validated forecasting framework, a multi-year projection was generated to estimate India's future life expectancy over the period **2025–2029**. Forecasts were produced using the Prophet time-series model and are reported alongside **95% and 99% prediction intervals** to explicitly capture uncertainty.

The results indicate a **steady upward trajectory** in **India's life expectancy**, with the point estimate increasing from **72.42 years** in **2025** to **73.90** years **by 2029.** This trend reflects continued improvements in healthcare access, disease prevention, and overall living conditions.

At the **95% confidence level,** projected life expectancy remains within a relatively stable range, suggesting **moderate uncertainty in the short-to-medium term.** The **99% prediction intervals** provide a more **conservative estimate, accounting for potential shocks** or **slower progress in health outcomes.**

NOTE: These projections should be interpreted as **probabilistic estimates rather than deterministic outcomes,** with the uncertainty bounds highlighting the range of plausible future scenarios.

Year	Predicted	95% LB	95% UB	99% LB	99% UB
2025	72.42	69.4	75.3	68.53	76.64
2026	72.79	69.65	76.19	67.67	77.41
2027	73.16	69.59	76.81	67.48	78.35
2028	73.53	69.13	78.29	67.92	79.33
2029	73.9	68.9	78.27	67.48	80.23

Table-12: Projected Life Expectancy for India with 95% and 99% Prediction Intervals (2025–2029)

Key Insights

1. **Socioeconomic and healthcare factors are the strongest determinants of life expectancy.**
   Education, income composition, immunization coverage, and healthcare expenditure exhibit strong positive associations with life expectancy, while adult mortality, disease burden, and nutritional deficiencies negatively influence longevity outcomes.

2. **Life expectancy differences across development groups are statistically significant.**
   ANOVA results confirm that disparities between developed and developing countries are structural rather than random, validating the need for segmented and cluster-based analysis.

3. **High multicollinearity exists among mortality, nutrition, and health indicators.**
   Several explanatory variables show strong interdependencies, which necessitated dimensionality reduction to ensure model stability and reliable interpretation.

4. **Principal Component Analysis effectively captured latent health and development dimensions.**
   PCA reduced the feature space into orthogonal components representing socioeconomic development, public health intervention intensity, and early-life health risks while preserving most of the dataset's variance.

5. **Countries form three distinct and interpretable health–development clusters.**
Clustering revealed coherent groups—Low Health & Low Development, Developing & Transitional Health, and High Health & High Development—highlighting systematic global inequalities in life expectancy.

6. **Lower-performing regions exhibit faster relative life expectancy growth.**
Forecasting results indicate that countries in low and transitional health clusters are experiencing larger relative gains, while high-development countries show plateauing trends due to saturation effects.

7. **The forecasting model demonstrates high accuracy and reliability.**
Model evaluation achieved an overall MAE of approximately 1.42 years and a MAPE of approximately 1.93%, indicating strong predictive performance across diverse country contexts.

8. **Uncertainty-aware forecasting enhances credibility.**
All validation observations fell within the 95% prediction intervals, confirming well-calibrated uncertainty estimates and reinforcing the robustness of the forecasting framework.

9. **India's life expectancy is projected to increase steadily in the medium term.**
Country-specific projections estimate a rise from 72.42 years in 2025 to 73.90 years by 2029, with clearly defined 95% and 99% prediction intervals reflecting forecast uncertainty.

10. **An integrated analytical approach yields robust, policy-relevant insights.**
Combining statistical inference, dimensionality reduction, clustering, and forecasting provides a comprehensive and reliable framework for understanding and projecting global life expectancy patterns.

Policy Recommendations

Based on the empirical findings and validated forecasting outcomes, the following policy recommendations are proposed to support sustainable improvements in global life expectancy, with emphasis on cluster-specific needs:

1. **Prioritize investments in education and income stability.**
Strong associations between schooling, income composition, and life expectancy highlight the importance of long-term investments in education and poverty reduction programs, particularly in low and transitional health clusters.

2. **Strengthen primary healthcare and preventive interventions.**
High mortality and disease burden in low-health clusters emphasize the need for expanded access to basic healthcare services, immunization programs, and early disease screening to reduce preventable deaths.

3. **Adopt cluster-specific health strategies rather than uniform policies.**
The clustering results demonstrate that countries face heterogeneous health challenges. Tailored policies addressing region-specific constraints are likely to be more effective than one-size-fits-all approaches.

4. **Focus on accelerating gains in developing and transitional countries.**
Forecasting indicates faster relative improvements in lower-performing clusters. Strategic investments during this phase can yield substantial long-term benefits and help reduce global health disparities.

5. **Address plateau effects in high-development countries.**
High life expectancy countries show limited marginal gains. Policy focus should shift toward quality of life, healthy aging, and non-communicable disease management rather than purely extending lifespan.

6. **Incorporate uncertainty-aware forecasting into policy planning.**
The use of prediction intervals provides a realistic range of future outcomes. Policymakers should adopt probabilistic forecasts to design resilient health strategies under uncertainty.

7. **Use country-specific projections for medium-term planning.**
Projections for India (2025–2029) illustrate how validated forecasts can inform national health targets, resource allocation, and progress monitoring when combined with uncertainty bounds.

8. While forecasting provides valuable guidance for planning, projections should not be interpreted as deterministic outcomes. Ethical use of such models requires transparency about uncertainty and careful consideration of contextual factors. Policymakers should treat forecasts as decision-support tools rather than prescriptive targets.

Limitations of the Study

While this study provides robust insights into global life expectancy patterns, certain limitations should be acknowledged. First, the analysis relies on secondary data, which may be affected by reporting delays, missing values, or measurement inconsistencies across countries. Second, the study focuses on association-based methods and does not establish causal relationships between explanatory variables and life expectancy. Third, forecasting results assume continuity of historical trends and may not fully capture the impact of unexpected external shocks such as pandemics, economic crises, or major policy changes. These limitations highlight the need for cautious interpretation and motivate further research.

Future Scope and Extensions

Future research could extend this work by incorporating causal modeling techniques such as panel regression or quasi-experimental designs to better isolate the drivers of life expectancy. Additionally, integrating region-specific health expenditure data and environmental indicators may improve explanatory power. From a forecasting perspective, ensemble approaches combining statistical and machine learning models could further enhance predictive accuracy. Expanding projections to align with global development targets would also increase policy relevance.