Determinants of Life Expectancy: Analyzing Economic and Health Predictors

Jinyi Xue

2024-12-03

Introduction

This study delves into life expectancy across different countries by utilizing a comprehensive dataset provided by the World Health Organization (WHO). The dataset includes a wide array of variables, such as economic factors and health-related statistics. By analyzing these predictors, we aim to understand the intricate relationships and key determinants that influence life expectancy.

The significance of this analysis lies in its potential to inform public health policies and strategies aimed at improving life expectancy. Understanding which factors have the most substantial impact can help governments and organizations prioritize interventions and allocate resources more effectively to enhance the overall health and longevity of populations.

Exploratory Data Analysis

The descriptive statistics was summarized by life expectancy levels, divided by quantile value. Life expectancy ranged from **36.3 to 62.9** is classified as `Low`; between **62.9 to 71.8** is classified as `Below_Average`; from **71.8 to 75.4** belongs to `Average`; and life expectancy from **75.4 to 89.0** is defined as `High`.

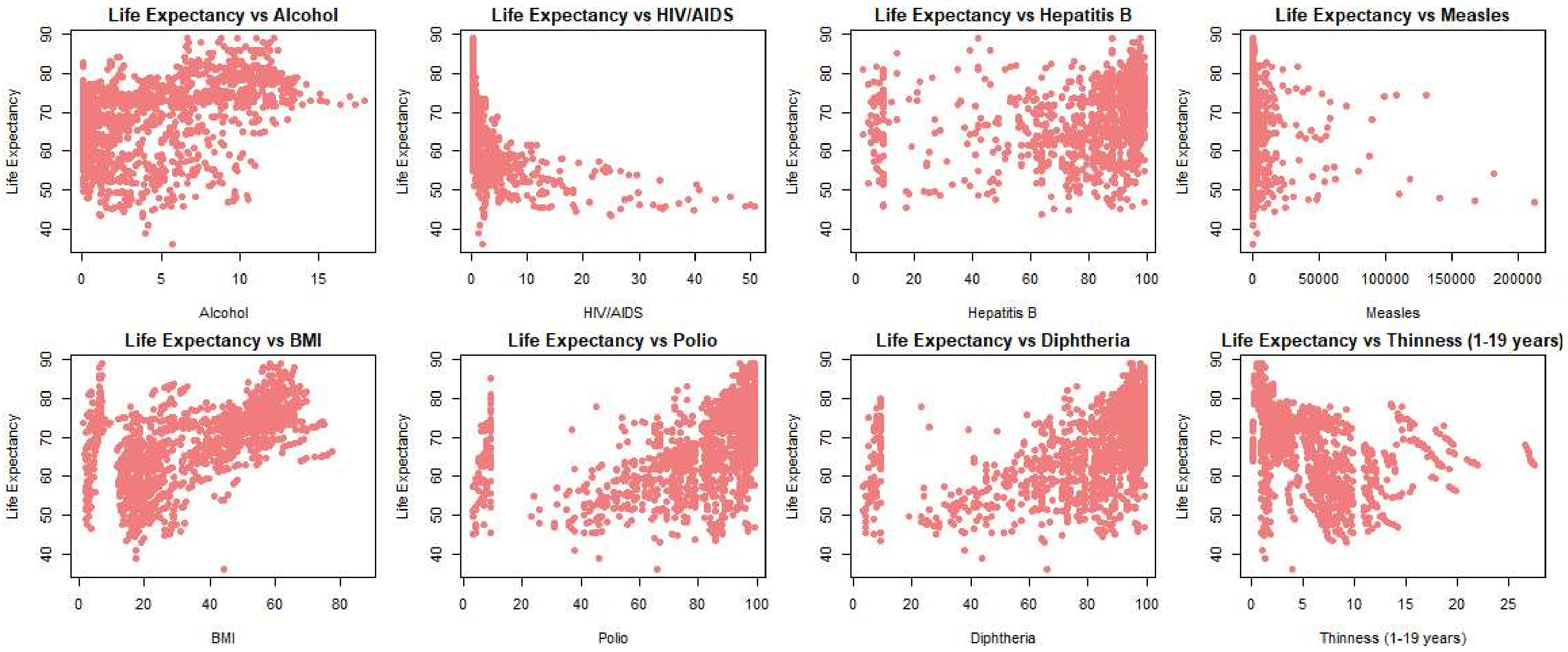
Comparison between each level were conducted using *ANOVA* for normally distributed variables and *Kruskal-Wallis* test for non-normally distributed variables.

	Low	Below_Average	Average	High	p	test
n	445	439	437	440		
Life.expectancy (mean (SD))	5.5e+01 (5.03)	6.7e+01 (2.58)	7.4e+01 (0.99)	8.0e+01 (3.04)	<0.001	
Adult.Mortality (mean (SD))	2.9e+02 (155.61)	1.8e+02 (77.07)	1.2e+02 (54.92)	7.3e+01 (40.39)	<0.001	
infant.deaths (median [IQR])	2.9e+01 [8.00, 56.00]	8.0e+00 [1.00, 27.00]	1.0e+00 [0.00, 8.00]	0.0e+00 [0.00, 2.00]	<0.001	nonnorm
Alcohol (median [IQR])	1.8e+00 [0.50, 4.58]	1.7e+00 [0.18, 4.60]	3.8e+00 [1.29, 7.53]	8.6e+00 [5.29, 10.69]	<0.001	nonnorm
percentage.expenditure (median [IQR])	2.5e+01 [2.82, 67.34]	3.4e+01 [1.99, 143.04]	1.5e+02 [15.26, 555.93]	6.8e+02 [20.30, 3 , 467.75]	<0.001	nonnorm
Hepatitis.B (median [IQR])	7.7e+01 [63.25, 89.00]	8.9e+01 [73.00, 96.00]	9.5e+01 [88.00, 98.00]	9.5e+01 [88.00, 97.00]	<0.001	nonnorm
Measles (median [IQR])	2.8e+02 [12.00, 2 ,	1.5e+01 [0.00, 820.00]	2.0e+00 [0.00,	7.0e+00 [0.00,	<0.001	nonnorm

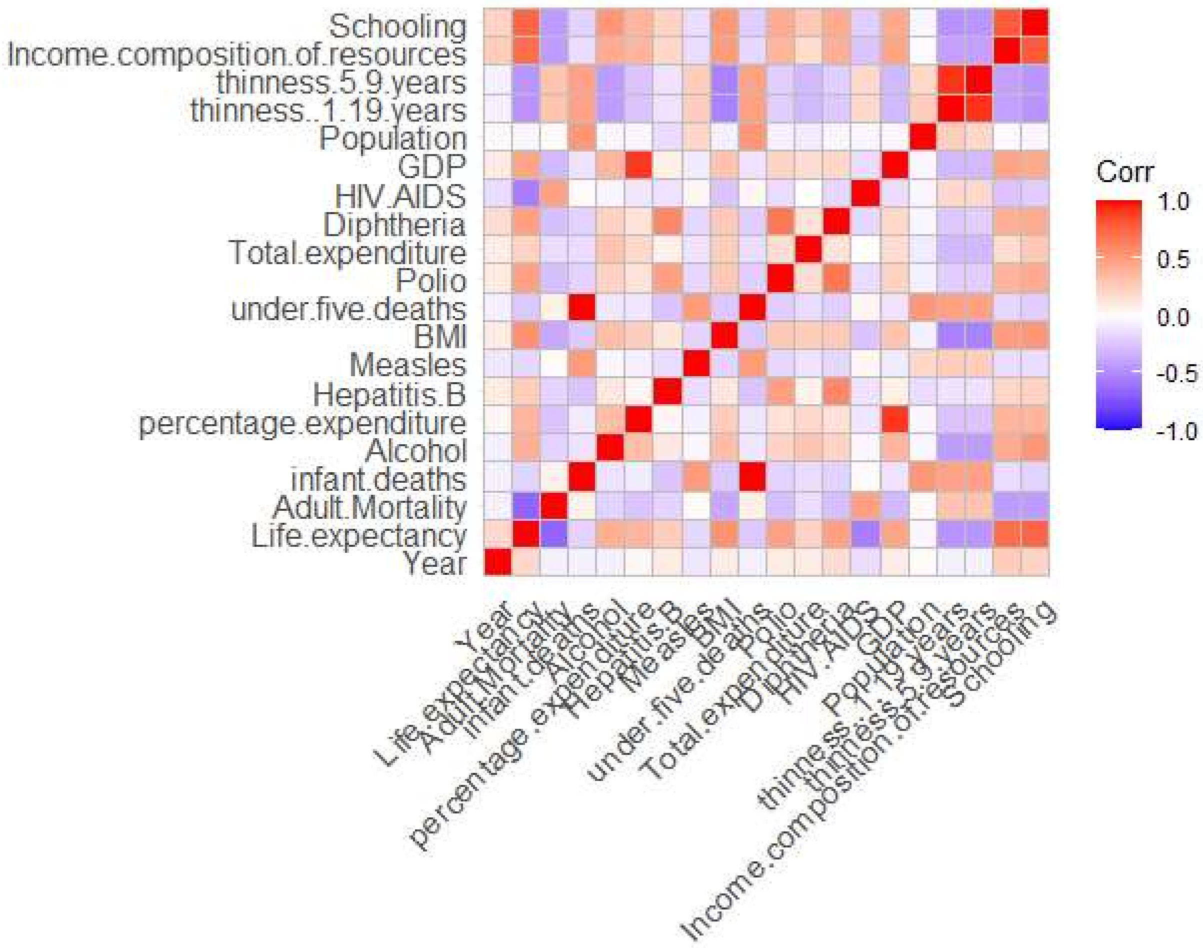
	Low	Below_Average	Average	High	p	test
	027.00]		55.00]	92.00]		
BMI (mean (SD))	2.0e+01 (9.21)	3.4e+01 (18.30)	4.6e+01 (18.07)	5.0e+01 (18.39)	<0.001	
under.five.deaths (median [IQR])	4.5e+01 [11.00, 89.00]	1.0e+01 [1.00, 34.00]	1.0e+00 [0.00, 10.00]	1.0e+00 [0.00, 2.00]	<0.001	nonnorm
Polio (median [IQR])	7.2e+01 [51.00, 85.00]	9.1e+01 [76.00, 97.00]	9.5e+01 [91.00, 98.00]	9.6e+01 [93.00, 98.00]	<0.001	nonnorm
Total.expenditure (mean (SD))	5.5e+00 (2.24)	5.2e+00 (2.17)	5.7e+00 (1.75)	7.4e+00 (3.10)	<0.001	
Diphtheria (median [IQR])	7.2e+01 [47.00, 85.00]	8.9e+01 [77.00, 96.00]	9.5e+01 [91.00, 98.00]	9.6e+01 [93.00, 98.00]	<0.001	nonnorm
HIV.AIDS (median [IQR])	3.4e+00 [1.50, 7.30]	1.0e-01 [0.10, 0.40]	1.0e-01 [0.10, 0.10]	1.0e-01 [0.10, 0.10]	<0.001	nonnorm
GDP (median [IQR])	4.5e+02 [232.58, 835.50]	9.2e+02 [337.36, 2 , 358.94]	3.5e+03 [1 , 158.55, 7 , 316.98]	1.2e+04 [3 , 472.38, 35 , 396.67]	<0.001	nonnorm
Population (median [IQR])	1.8e+06 [573 , 416.00, 9e+06]	1.5e+06 [223 , 143.00, 1.2e+07]	1.1e+06 [95 , 574.00, 6 , 178 , 908.50]	1.1e+06 [209 , 107.50, 5 , 622 , 067.75]	<0.001	nonnorm
thinness..1.19.years (median [IQR])	7.8e+00 [6.30, 9.30]	3.7e+00 [2.40, 7.60]	3.1e+00 [2.10, 5.70]	1.1e+00 [0.80, 2.00]	<0.001	nonnorm
thinness.5.9.years (median [IQR])	7.6e+00 [6.00, 9.10]	3.8e+00 [2.60, 8.10]	3.1e+00 [2.10, 5.80]	1.0e+00 [0.70, 2.00]	<0.001	nonnorm
Income.composition.of.resources (median [IQR])	4.4e-01 [0.38, 0.49]	6.0e-01 [0.50, 0.67]	7.2e-01 [0.68, 0.77]	8.5e-01 [0.79, 0.89]	<0.001	nonnorm
Schooling (median [IQR])	8.8e+00 [6.80, 10.20]	1.1e+01 [9.95, 12.50]	1.3e+01 [12.10, 14.20]	1.5e+01 [13.80, 16.65]	<0.001	nonnorm
Status = Developing (%)	445 (100.0)	431 (98.2)	387 (88.6)	194 (44.1)	<0.001	
Continent (%)					<0.001	
Africa	395 (88.8)	88 (20.0)	51 (11.7)	4 (0.9)		
Americas	9 (2.0)	64 (14.6)	139 (31.8)	81 (18.4)		
Asia	31 (7.0)	197 (44.9)	136 (31.1)	94 (21.4)		

	Low	Below_Average	Average	High	p	test
Europe	0 (0.0)	39 (8.9)	92 (21.1)	239 (54.3)		
Oceania	10 (2.2)	51 (11.6)	19 (4.3)	22 (5.0)		

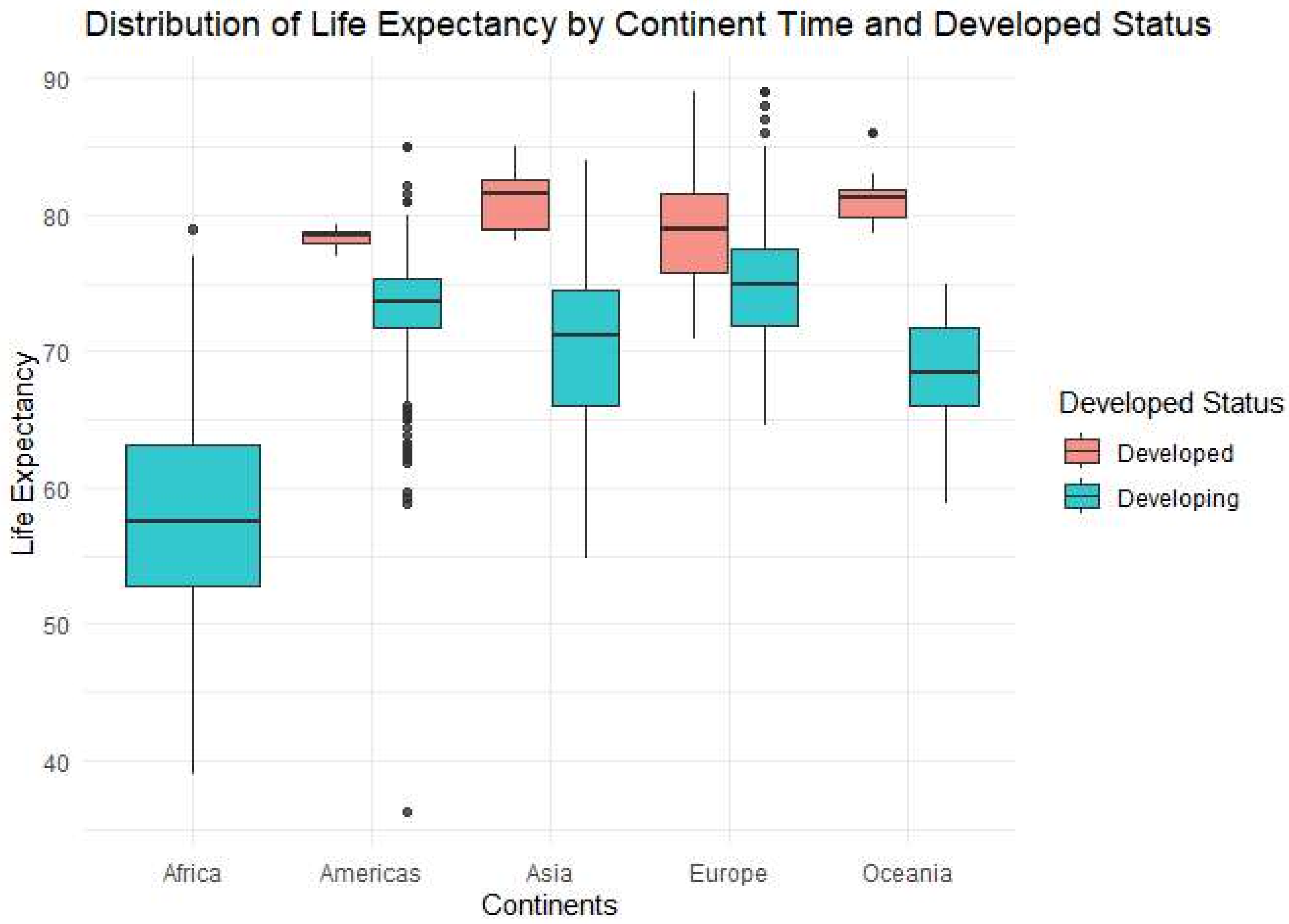
From the table, each level contain similar amount of observations, minimized bias from different sample size. The p-value implies significant differences **(p < .001)** between each life expectancy level are present across all variables.



The scatter plots of different predictors on the x-axis and life expectancy on the y-axis indicate clear relationships between them. However, the plots also reveal that these relationships are mostly **non-linear**, as evidenced by *clusters* and *curved trends*.



The correlation matrix revealed that several predictors are strongly correlated with life expectancy, as well as with each other. For example, `life expectancy` is strongly correlated with `adult mortality`, `BMI`, the `five-year death rate`, `HIV/AIDS` prevalence, `schooling`, and the `income composition of resources`. Additionally, `schooling` and the `income composition of resources` are positively and strongly correlated with `GDP`. Conversely, `BMI` is strongly and negatively correlated with `thinness`.



The boxplot illustrates that life expectancy varies significantly both across continents and within continents (between developed and developing countries). Except for Africa, which has no developed countries, within each continent, developed countries tend to have higher life expectancies compared to developing countries. The shape and length of the boxplots suggest that life expectancy differs greatly under various conditions and that its variance changes accordingly.

Methods Motivation

In our study, we employ **tree classification** and **Partial Least Squares (PLS)** modeling to predict life expectancy.

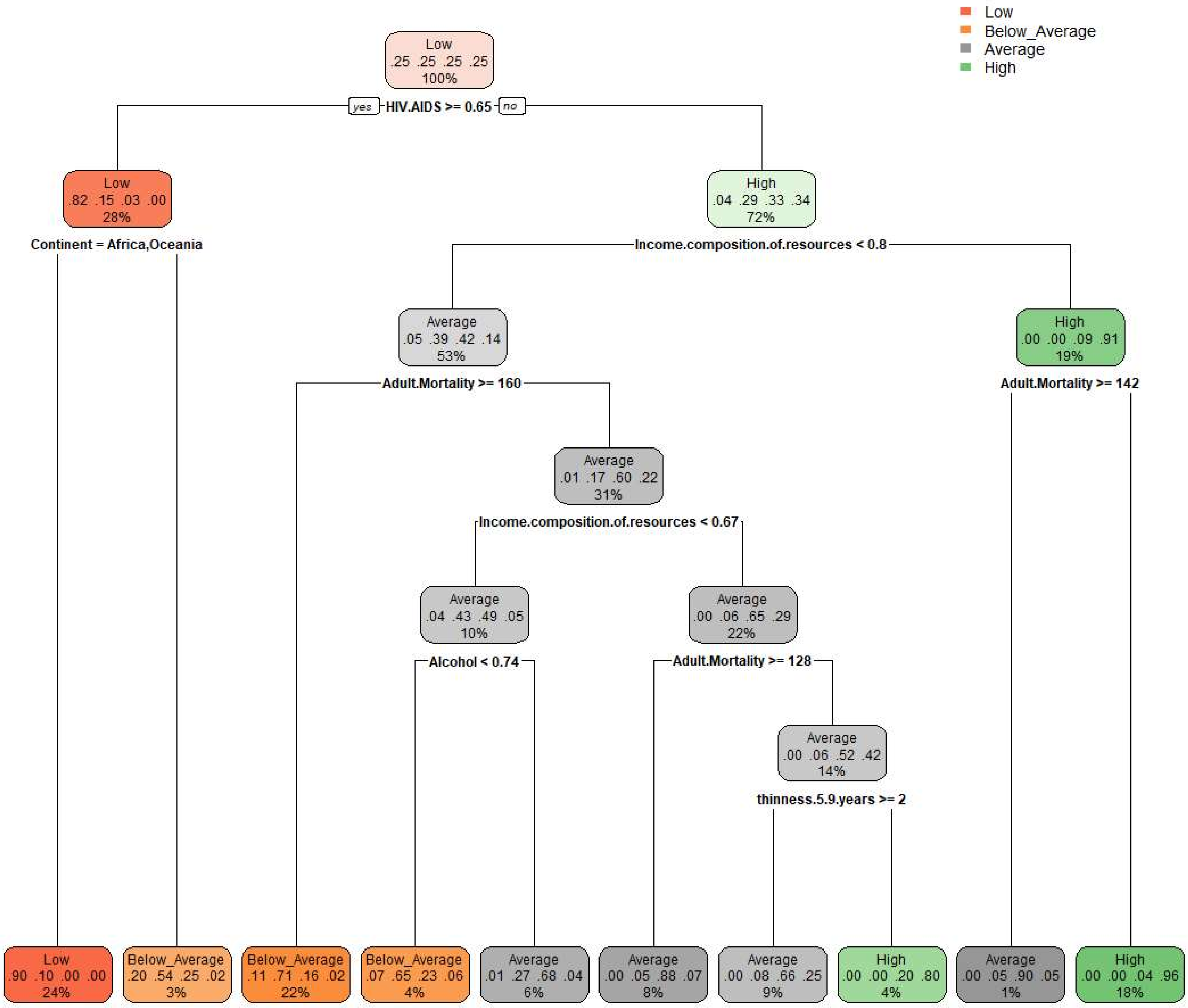
The tree classification method was chosen due to the observed *clusters* in the scatterplots. Besides its intuitive *interpretability*, making it ideal for identifying key determinants of life expectancy in a diverse dataset. Additionally, **PLS modeling** is particularly useful for dealing with *multicollinearity* among predictors, which is common in datasets with numerous interrelated variables.

By leveraging both methods, we aim to capture the intricate relationships within the data and enhance the robustness and accuracy of our predictions.

Tree Classification

Model Assumption

The decision tree model assumes that the training data is **representative** of the overall population and that the observed **patterns or clusters** in the data are consistent. It also assumes that the predictors are accurately measured and that any missing data has been properly handled, ensuring that the relationships between variables remain reliable for making predictions.


🟥 Low
🟧 Below_Average
▨ Average
🟩 High
Low .25 .25 .25 .25 100%
yes HIV.AIDS >= 0.65 no
Low .82 .15 .03 .00 28%
Continent = Africa,Oceania
High .04 .29 .33 .34 72%
Income.composition.of.resources < 0.8
Average .05 .39 .42 .14 53%
Adult.Mortality >= 160
High .00 .00 .09 .91 19%
Adult.Mortality >= 142
Average .01 .17 .60 .22 31%
Income.composition.of.resources < 0.67
Average .04 .43 .49 .05 10%
Alcohol < 0.74
Average .00 .06 .65 .29 22%
Adult.Mortality >= 128
Average .00 .06 .52 .42 14%
thinness.5.9.years >= 2
Low .90 .10 .00 .00 24%
Below_Average .20 .54 .25 .02 3%
Below_Average .11 .71 .16 .02 22%
Below_Average .07 .65 .23 .06 4%
Average .01 .27 .68 .04 6%
Average .00 .05 .88 .07 8%
Average .00 .08 .66 .25 9%
High .00 .00 .20 .80 4%
Average .00 .05 .90 .05 1%
High .00 .00 .04 .96 18%

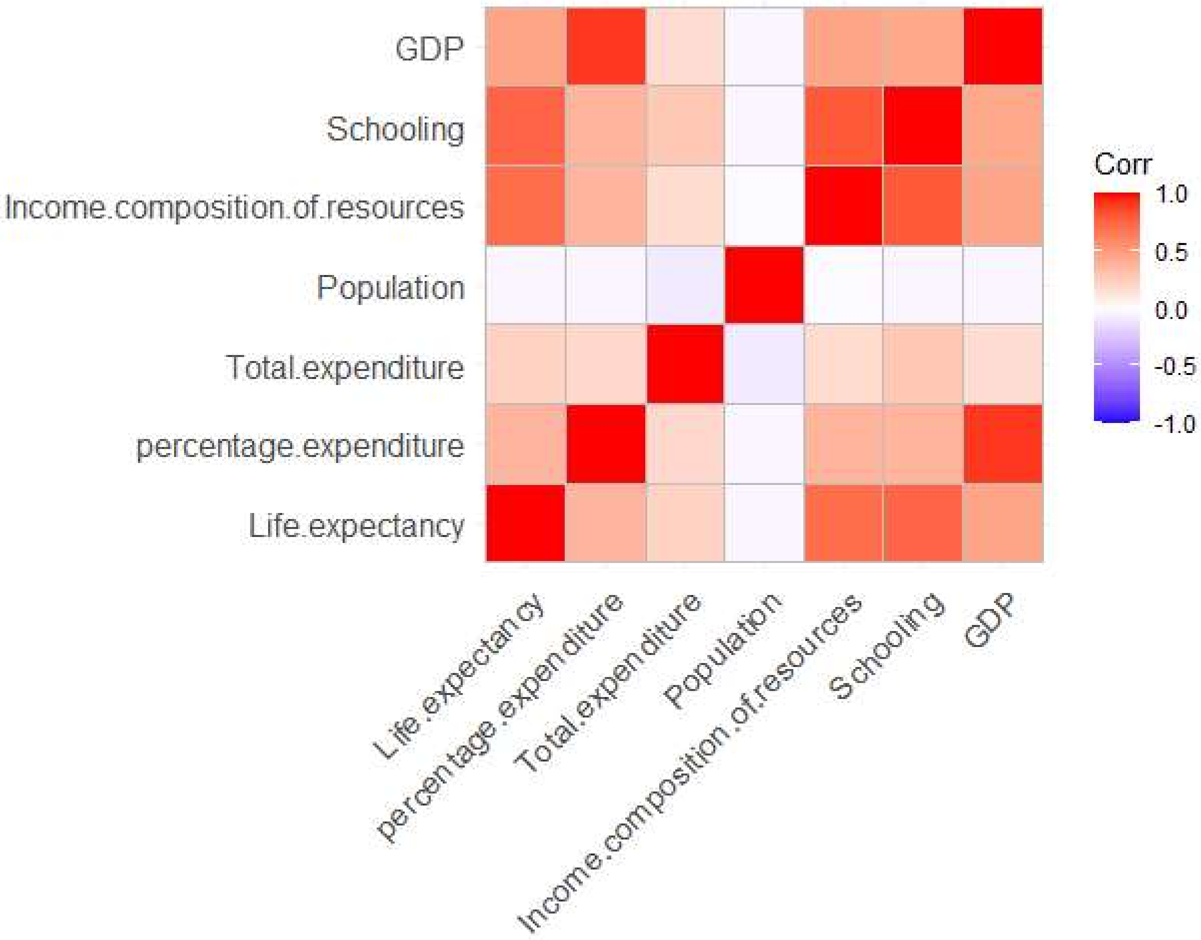
Model Description

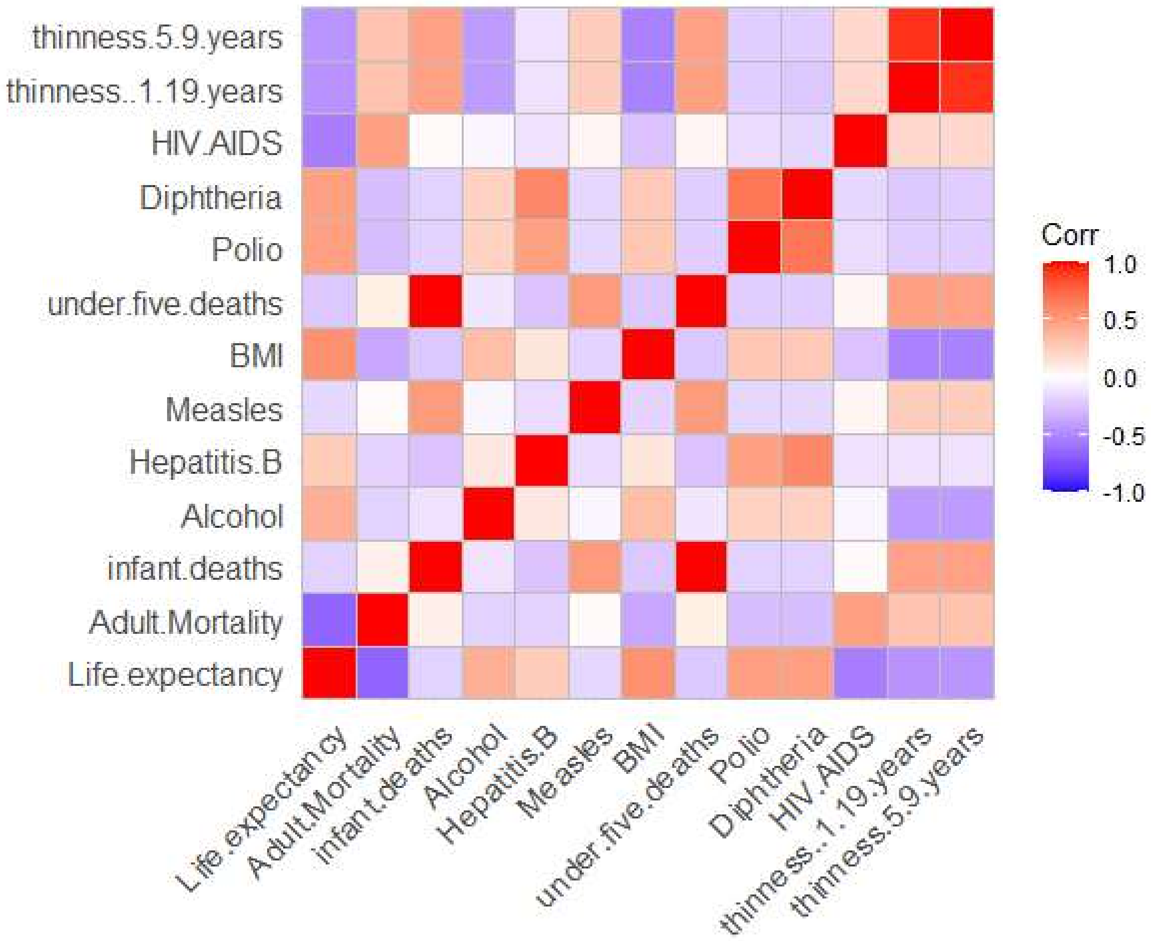
The decision tree model was trained to predict life expectancy levels using various predictors. It splits data based on key factors like HIV/AIDS prevalence, income composition of resources, and adult mortality. Higher HIV/AIDS prevalence, for example, indicates lower life expectancy, especially in Africa and Oceania. The model identifies significant predictors and their interactions, highlighting the intricate relationships that influence life expectancy across regions.

	x
Accuracy	0.7849186
Kappa	0.7131931
AccuracyLower	0.7602131
AccuracyUpper	0.8081789
AccuracyNull	0.2759212
AccuracyPValue	0.0000000
McnemarPValue	NaN

The **confusion matrix** shows an accuracy of approximately *78.5%*, with a Kappa statistic of 0.713, indicating substantial agreement between observed and predicted classifications. The accuracy lower and upper bounds range from 76.02% to 80.82%. The baseline accuracy, without any model, would be 27.59%, demonstrating that the model’s performance is significantly better.

PLS Model





The factors are separated based on diseased factors and economic factors. Both correlation matrix confirmed that each either diseased factors and economic factors have strong correlation between predictors and response (Life Expectancy).

Model Assumption

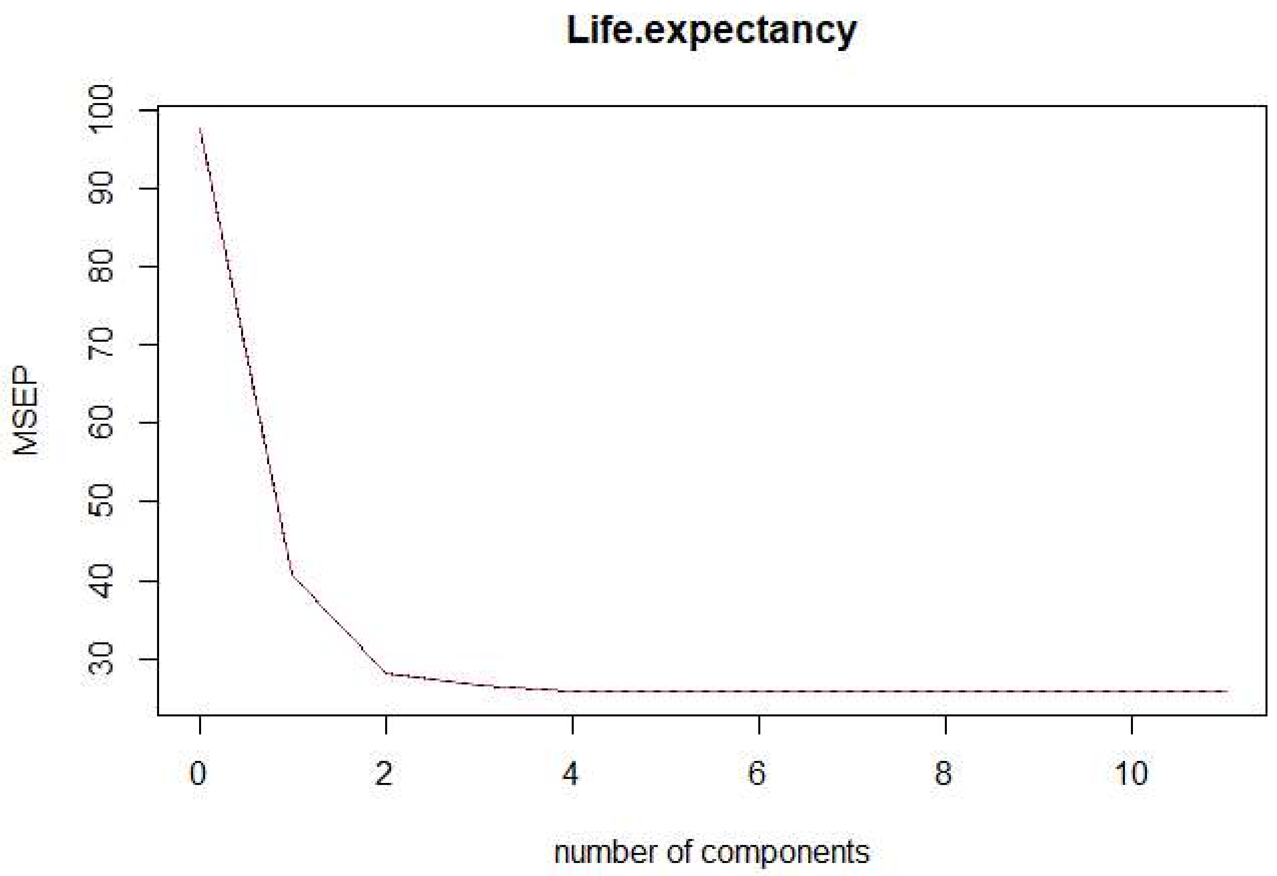
The PLS model assumes that the predictors are **highly correlated** with life expectancy and among themselves, as confirmed by the correlation matrices. It also assumes that the data has been appropriately **scaled and standardized**, ensuring that the relationships between economic and disease-related factors and life expectancy are accurately captured in the model.

Model Setup

The economic and diseased factor models were fitted using Partial Least Squares (PLS) regression with cross-validation. The analysis aimed to predict life expectancy based on multiple economic and disease predictors, utilizing a kernel PLS algorithm with scaling of predictors.

Economic Factor Model

```
## Data:    X dimension: 1271 11
##  Y dimension: 1271 1
## Fit method: kernelpls
## Number of components considered: 11
## 
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           9.876    6.372    5.311    5.169    5.081    5.077    5.076
## adjCV        9.876    6.370    5.310    5.164    5.078    5.075    5.074
##        7 comps  8 comps  9 comps  10 comps  11 comps
## CV       5.077    5.078    5.078     5.078     5.078
## adjCV    5.074    5.075    5.075     5.075     5.075
## 
## TRAINING: % variance explained
##                  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                  36.06    46.66    52.68    58.95    67.56    73.79    81.65
## Life.expectancy    58.67    71.39    73.27    74.04    74.07    74.09    74.09
##                  8 comps  9 comps  10 comps  11 comps
## X                  83.50    87.79     91.00    100.00
## Life.expectancy    74.09    74.09     74.09     74.09
```



Model Description

The model stabilizes around **4-5 components**, with minimal improvements in prediction accuracy beyond this point. At 4 components, the cross-validated Root Mean Squared Error of Prediction (RMSEP) is 5.078, indicating the average prediction error in life expectancy years.

Regarding variance explanation, at 4 components, the X variables (predictors) explain 58.95% of the variance, while the life expectancy model explains 74.04% of the variance, suggesting a strong predictive capability.

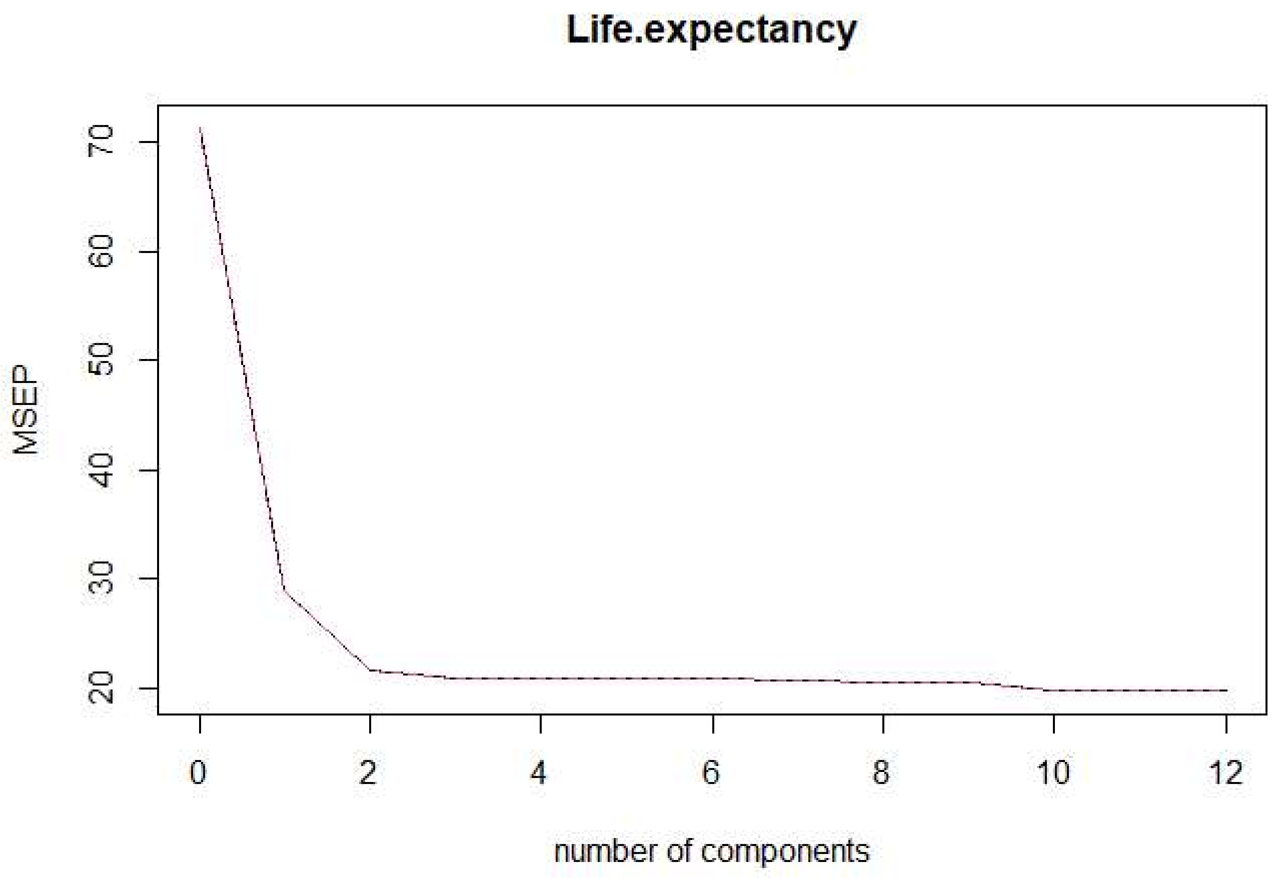
```
## [1] 23.71213
```

```
## [1] 4.86951
```

Using 4 components, the model was applied to the test dataset, resulting in a Mean Squared Error (MSE) of **23.712** and a Root Mean Squared Error (RMSE) of **4.870** years. The RMSE indicates that, on average, the model’s predictions deviate by approximately **±4.870** years from the actual life expectancy. This suggests a reasonable level of predictive accuracy, though there is still notable variability in the predictions.

Diseased Factor Model

```
## Data:    X dimension: 1305 12
##  Y dimension: 1305 1
## Fit method: kernelpls
## Number of components considered: 12
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           8.441    5.385    4.650    4.577    4.567    4.568    4.567
## adjCV        8.441    5.380    4.649    4.575    4.565    4.566    4.564
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## CV       4.553    4.532    4.518     4.454     4.442     4.442
## adjCV    4.551    4.531    4.519     4.450     4.440     4.439
##
## TRAINING: % variance explained
##                  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                  30.24    46.72    55.85    63.67    75.73    81.75    84.24
## Life.expectancy    60.20    69.93    71.13    71.33    71.35    71.39    71.55
##                  8 comps  9 comps  10 comps  11 comps  12 comps
## X                  87.27    91.31      92.3     96.16    100.00
## Life.expectancy    71.70    71.92      72.8     72.86     72.86
```



Model Description

The Partial Least Squares (PLS) regression model for disease-related factors demonstrates a robust approach to predicting life expectancy. The model shows a gradual stabilization of predictive performance across components, with the most significant improvements occurring in the first few components. At **3 components**, the cross-validated Root Mean Squared Error of Prediction (RMSEP) is **4.597** years, representing the average prediction error.

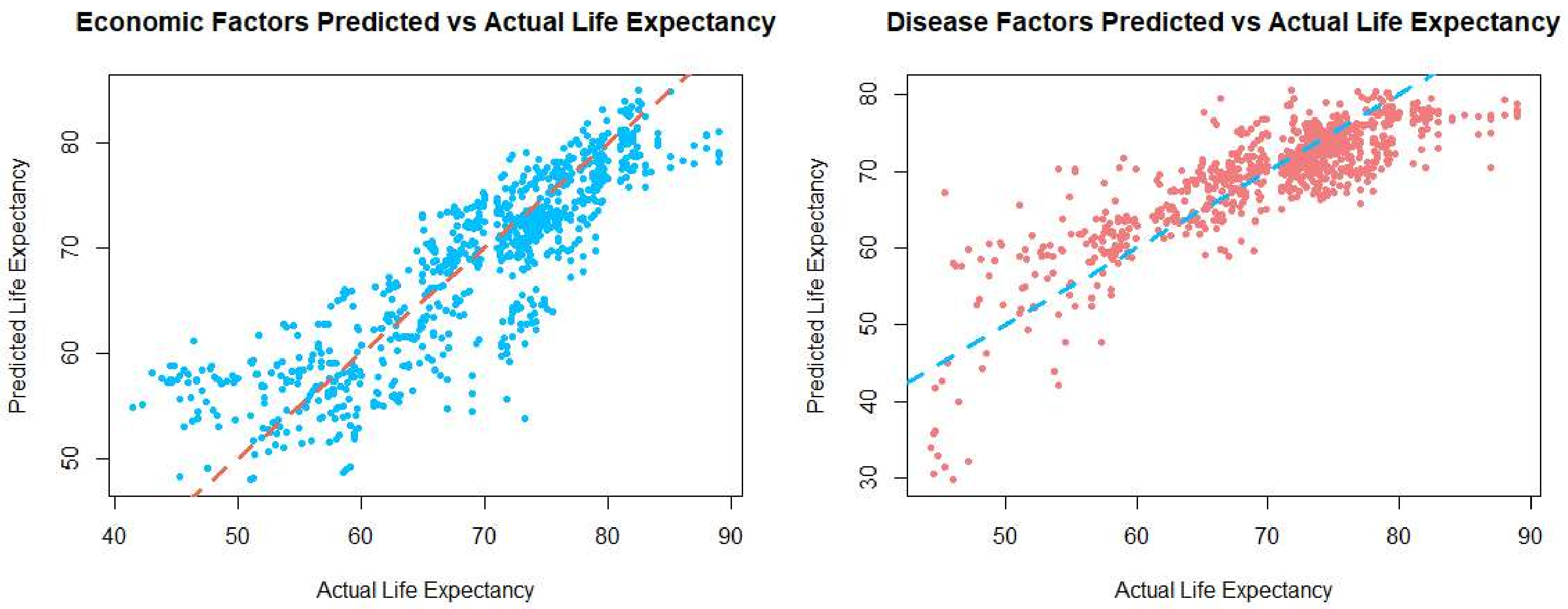
At 3 components, the disease predictors (X variables) explain 55.85% of the variance. The life expectancy model explains 71.13% of the variance, suggesting a strong predictive capability and a good model fit. This indicates that the selected disease-related variables capture a substantial proportion of the factors influencing life expectancy.

```
## [1] 21.13231
```

```
## [1] 4.59699
```

Using 3 components, the model was applied to the test dataset, resulting in a Mean Squared Error (MSE) of **21.132** and a Root Mean Squared Error (RMSE) of **4.597** years. The RMSE indicates that, on average, the model’s predictions deviate by approximately **±4.597** years from the actual life expectancy. This suggests a reasonable level of predictive accuracy, with performance comparable to the economic factors model.

Models Comparison



For both figures, data points are mostly clustered around the red/blue dashed line, which represents a perfect 1:1 relationship between predicted and actual values. This suggests that either the economic or diseased factors model is generally able to predict life expectancy fairly accurately, with the majority of predictions falling close to the actual values.

However, the disease factors model shows slightly tighter clustering around the line compared to the economic factors model, suggesting lower overall prediction error.Additionally, the disease factors model appears to have particularly strong predictive accuracy for higher life expectancy values around 60~80 years.

Discussion

In summary, the analysis explored the key determinants of life expectancy using economic and disease factors. The decision tree model identified important predictors like HIV/AIDS prevalence, income composition, and adult mortality as influential factors in determining life expectancy levels across regions.

The Partial Least Squares (PLS) regression models showed that both economic and disease-related factors are significant predictors of life expectancy, with the disease factors model exhibiting slightly stronger performance.

These findings underscore the need for holistic, evidence-based policies that address the interconnected economic, social, and health-related determinants of longevity. Future research in this area could explore additional predictors, such as environmental factors, access to healthcare, and sociocultural norms, to develop a more comprehensive understanding of the drivers of life expectancy.