

Analysis and Prediction of Life Expectancy at the Global Level and in Selected Countries

Dimitrios Kazolis ^{1,*}, Ioanna Kranioti ¹ and Theodosia Kazoli ²

¹ Department of Physics, Democritus University of Thrace, Kavala, Greece; okrani@physics.duth.gr

² Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece; ics26195@uom.edu.gr

*E-mail of corresponding author: dkazolis@physics.duth.gr

Abstract: The relationship between life expectancy and healthcare costs has been the subject of numerous studies by various organizations. This is because it serves as a key tool for evaluating the effectiveness of the healthcare sector, as it affects both the national budget and the quality of life and life expectancy of ordinary citizens. The purpose of this study is to contribute to the advancement of this research by incorporating additional parameters and applying combinations of data analysis techniques. Initially, the differences between selected and representative countries were presented, but basically this analysis focused on predicting life expectancy using a set of economic, health, and environmental indicators. To achieve this goal, two data sets from the World Bank's open data platform were used for the period 2000–2023. The study involved handling of missing values, performing correlation analysis, applying a linear regression model, conducting principal component analysis, and creating a regression model using the random forest method. The process was implemented using the Python programming language. Finally, the summary results were presented and recommendations for future research were formulated. This comprehensive effort is intended to serve as an advisory tool for relevant decision-making and provides the basis for further analysis of the topic.

Received 14 April 2026

Accepted 15 May 2026

Published 1 September 2026

Keywords: Life Expectancy; Health Expenditure; k-Nearest Neighbors; Correlation Analysis; Regression Model; Principal Component Analysis; Random Forest.

1. Introduction

Various projects have been conducted on life expectancy and the factors that influence it [1,2]. This paper seeks to delve deeper into this important topic by using additional data and combining different methodologies. Drawing on a variety of databases from the World Bank [3], an effort is made to identify correlations between these databases, even if they do not necessarily reflect a causal relationship. Thus, databases used that include life expectancy, per capita health expenditure, GDP per capita, the number of hospital beds per 1,000 people, liters of pure alcohol per person aged 15 and older, smoking prevalence including all types of tobacco, and infant mortality for children up to five years of age per 100 births.

In addition, data sets related to education have been collected, such as the percentage of enrollments in higher education relative to the population of the corresponding "standard" age group. Additionally, a separate dataset is provided showing the average number of years of education completed by adults (25+), which is considered to provide greater accuracy, as it further filters out noise

from the previous list. Another dataset is also provided that focuses more on environmental factors, as it examines air pollution, specifically the average concentration of PM2.5 particulate matter in the air in micrograms per cubic meter.

Moreover, records have been included regarding the percentage of the population that has access to drinking water, the percentage of the labor force that is unemployed but actively seeking work, and the percentages of the adult population that is obese, i.e., has a body mass index (BMI) of 30 or higher.

The purpose of this study, therefore, is to examine all the factors that may influence life expectancy. This is achieved through various procedures, as described in the following chapter.

2. Materials and Methods

As mentioned above, As mentioned above, the data were retrieved from the World Bank's open data platform. Initially, they were processed to clean and standardize them. The main challenge was filling in the missing values. The columns with relatively small percentages of missing data (such as 'Obesity', 'Pure Alcohol', 'PM2.5', and 'Education') were imputed with the mean or median value. However, for 'Smoking prevalence %' with over 63% missing data, more advanced imputation techniques were considered. This task was accomplished using the k-Nearest Neighbors (KNN) method [4], which, however, will not be analyzed for reasons of brevity, and thus, the



© The Author(s). Published by CIEES Association Press. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Generic License.

missing values were imputed by considering patterns in the remaining features.

2.1. Visualization of data

This study examines the correlation not only at the global level but also analyzes data for 22 selected countries. These are: Belgium, Brazil, Bulgaria, Canada, Cyprus, Germany, France, Greece, Italy, Japan, Luxembourg, Norway, Poland, Romania, Portugal, South Africa, Sweden, Switzerland, Spain, Turkey, the United States, and the United Kingdom. Fig. 1 illustrates the relationship between life expectancy and the healthcare costs per citizen.

This graph can effectively be obtained by removing the time component from the two corresponding time series. The Python program was used in the Thonny environment to create this graph. The results shown, are very important for drawing initial conclusions. In general, both variables revile an upward trend over time, with a slight apparent decline during the COVID-19 pandemic, which appears to have had the greatest impact in 2021, the year in which the effects of the global pandemic became apparent.

The general trend suggests that countries with higher expenditures may have better prospects for a longer life expectancy. It can be noted that the US has the highest expenditures; nevertheless, its life expectancy is significantly lower compared to other countries. Japan, on the other hand, achieves the highest life expectancy among the selected countries with very low expenditures, a fact that may be attributed to various socio-economic factors. South Africa appears to follow the pattern of low spending and low life expectancy and has the lowest figures.

Furthermore, Greece has a relatively high life expectancy relative to its expenditures. As for Turkey, Bulgaria, and the global trend, the low life expectancy is justified by the correspondingly low expenditures. Norway and Switzerland have high life expectancy, but also high expenditures.

In addition, the following chart illustrates the variables of life expectancy and healthcare costs for the selected countries, as well as the global trend. These figures are for the year 2023. Fig. 2 better illustrates the strong correlation between life expectancy and health expenditure per capita.

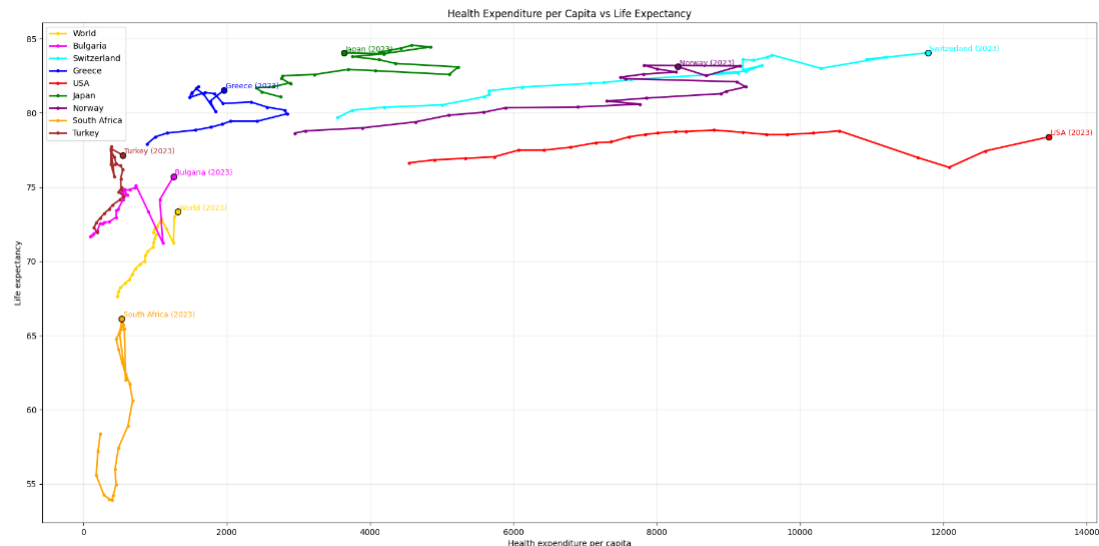


Fig. 1. Health Expenditure per Capita vs Life Expectancy.

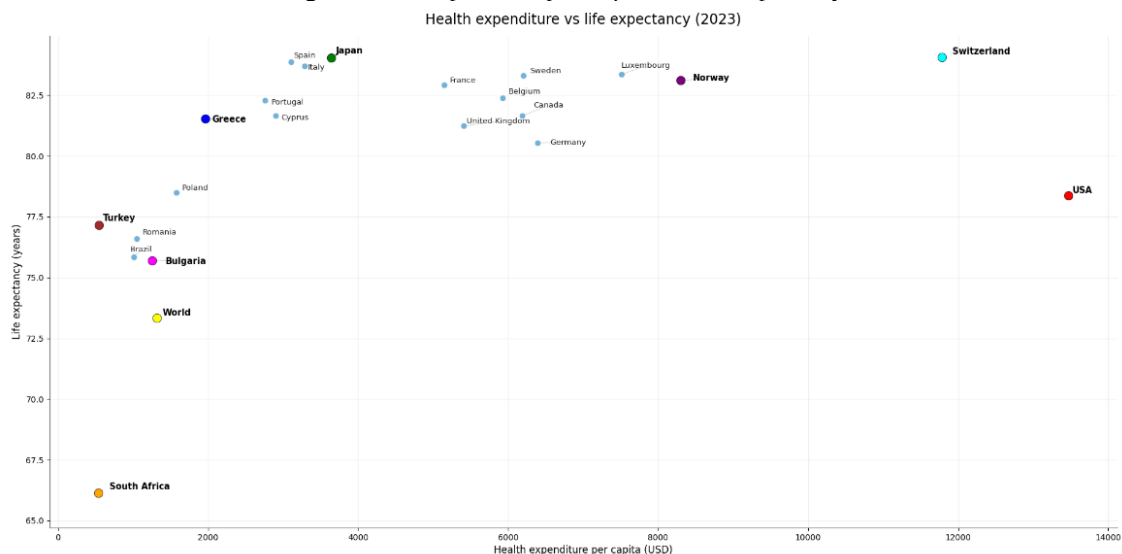


Fig. 2. Health Expenditure vs Life Expectancy (2023).

These graphs reflect that while there is generally a strong correlation between the two variables, it is not absolute, and furthermore, it is noted that there is not necessarily a causal relationship between them. Some of the reasons why life expectancy is not directly linked to costs are that it also depends on various parameters, which will be examined below.

2.2. Correlation analysis

Since all missing values in the specified columns have been handled, a correlation heatmap was generated for the numerical variables in the DataFrame. This will help visualize the relationships between the various features. This visualization presents the Pearson correlation coefficients between all numerical features, as shown in Fig. 3 below.

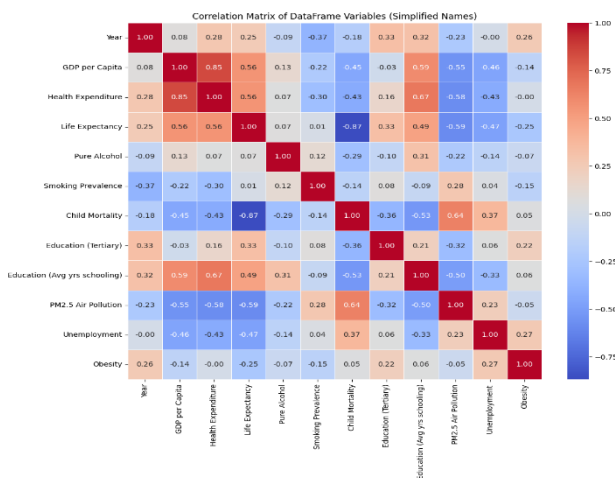


Fig. 3. The Correlation Matrix.

Based on the above correlation matrix, the strongest positive and negative correlations can be observed:

The strongest positive correlation observed is between 'GDP per capita' and 'Health expenditure', with a correlation coefficient of 0.85. This indicates a very strong direct relationship, suggesting that countries with higher GDP per capita tend to also have higher health expenditure.

On the other hand, the strongest negative correlation is between 'GDP per capita' and 'Hospital beds per 1000 people' with a correlation coefficient of -0.20. This suggests a weak inverse relationship, where countries with higher GDP per capita tend to have slightly fewer hospital beds per 1000 people. However, it is important to note that this is a relatively weak correlation. To provide a clearer overview, a bar chart was created showing the five strongest correlations, clearly illustrating the most significant positive and negative relationships within the dataset.

In the chart in Fig. 4, each bar represents a pair of variables. It is easy to be noticed the high negative correlation between life expectancy and child mortality, and conversely, the high positive correlation between GDP, health expenditure and education. This suggests that economic prosperity and educational attainment are often linked to higher health spending.

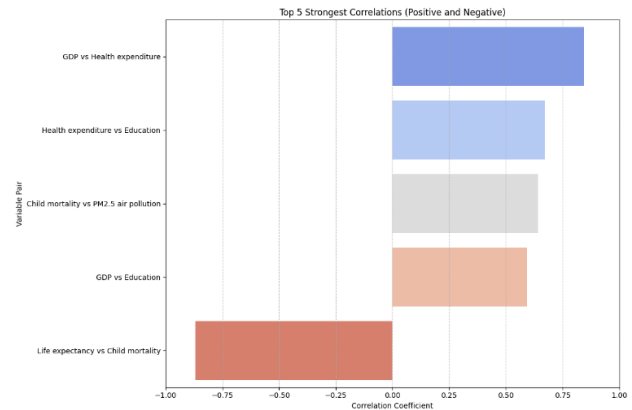


Fig. 4. The Correlation Bar Chart.

Subsequently, the correlation coefficients between 'Life Expectancy' and all other variables in the dataset were identified (Table 1).

Table 1. Correlations with 'Life Expectancy'.

Variable	Correlation
Life Expectancy	1.000000
Health Expenditure	0.561414
GDP per Capita	0.558794
Education (Avg yrs schooling)	0.493118
Education (Tertiary)	0.333733
Year	0.249680
Pure Alcohol	0.070023
Smoking Prevalence	0.011482
Obesity	-0.246621
Unemployment	-0.474428
Air Pollution	-0.589279
Child Mortality	-0.868321

The main conclusions that can be drawn from the Table 1 above are as follows:

- Firstly, the strongest positive correlations of Health Expenditure (0.561) and GDP per Capita (0.559), indicates that higher health expenditure is moderately associated with higher life expectancy and respectively, countries with higher GDP per capita, tend to have higher life expectancy. In addition, the education, more years of schooling, is linked to increased life expectancy (0.493).

- The strongest negative correlations are between child mortality (-0.868), indicating that a higher child mortality rate is strongly associated with lower life expectancy. This is a very significant relationship. Furthermore, PM2.5 air pollution (-0.589) indicates that higher levels of air pollution are moderately associated with lower life expectancy. The same principle applies to unemployment, (-0.474), and higher unemployment rates are linked to lower life expectancy.

- In summary, factors like child mortality, air pollution, and unemployment have a significant negative impact on life expectancy, while health expenditure, GDP per capita, and education show a strong positive relationship.

- Finally, weak or negligible correlations can be observed in pure alcohol (0.070) and smoking prevalence (0.011). These weak positive correlations, may be due to confounding factors or the specific nature of the dataset. The ten strongest correlations with life expectancy, are presented in the following Fig. 5.

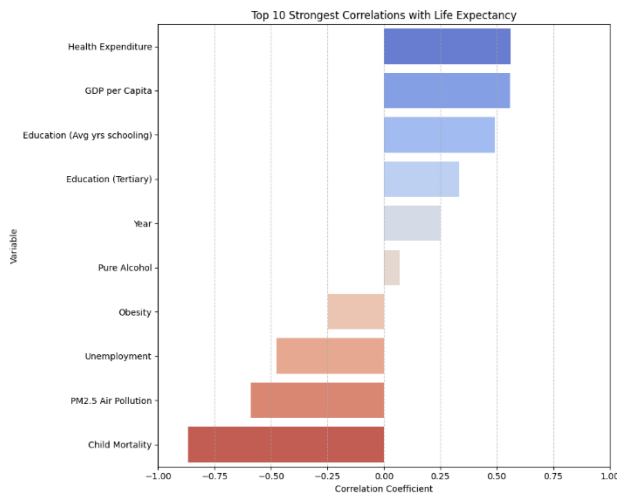


Fig. 5. The 10 strongest correlations with life expectancy.

The above heatmap provides a comprehensive overview of the linear relationships within the specific dataset. As a continuation of this effort, a regression analysis will be conducted, to examine how various variables quantitatively predict life expectancy.

2.3. Regression analysis

As mentioned above, regression analysis was implemented, to understand how different variables quantitatively predict 'Life Expectancy'. For this reason, Multiple Linear Regression, was performed. The model was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared, and the coefficients were analyzed. Upon completion of the process, the main results are as follows:

- **Model Evaluation:** The MSE was found 0.02. This low MSE indicates that the model's predictions are close to the actual values. Moreover, the RMSE, which represents the typical error magnitude, was found 0.16. Finally, the R-squared (R^2) was found 0.82. This means that approximately 82% of the variance in 'Life Expectancy' can be explained by the features included in the model, which is a relatively strong R^2 -squared, suggesting a good fit.

- **Model Coefficients:** The Intercept, the expected 'Life Expectancy' when all other predictor variables are zero, was found 0.05. Furthermore, the Child Mortality of -0.82, is the strongest negative predictor. For every one-unit increase in 'Child Mortality', 'Life Expectancy' is predicted to decrease by 0.82 units, holding other variables constant. This aligns with our earlier correlation findings. The Unemployment of -0.53, suggests that one-unit increase in 'Unemployment' is associated with a 0.53-unit decrease in 'Life Expectancy'.

The variables Pure Alcohol (-0.28), Obesity (-0.17), PM2.5 Air Pollution (-0.12), Smoking Prevalence (-0.12), and GDP per Capita (-0.18), also show negative coefficients, meaning an increase in these factors is associated with a decrease in life expectancy. The negative coefficient for GDP per Capita might seem counter-intuitive at first, given its strong positive correlation. This often happens in multiple regression due to multicollinearity or when the effect of a variable is conditional on other variables in the model. It suggests that

once the effects of other variables like Health Expenditure and Education are accounted for, GDP per Capita might have a different marginal effect.

The Education (Avg years schooling) also has a negative coefficient (-0.06), which is again counter-intuitive. As a result these relationships should need further investigation.

Finally, the Health Expenditure of 0.05 is the only variable with a positive coefficient. A one-unit increase in 'Health Expenditure' is predicted to increase 'Life Expectancy' by 0.05 units. The overall results are presented in the following Table 2.

Table 2. Regression Analysis-Model Coefficients.

Feature	Coefficient
0 GDP per Capita	-0.184001
1 Health Expenditure	0.045824
2 Child Mortality	-0.823087
3 Education (Avg yrs schooling)	-0.061094
4 Air Pollution	-0.120481
5 Unemployment	-0.531425
7 Pure Alcohol	-0.275931
6 Obesity	-0.171990
8 Smoking Prevalence	-0.118452

Moreover, the Variance Inflation Factor (VIF) values have been calculated to assess multicollinearity among the predictor variables:

- Health Expenditure (122.96), GDP per Capita (51.81), PM2.5 Air Pollution (36.32), Child Mortality (22.11), and Education (Avg yrs schooling) (19.41) all show very high VIF values (significantly above the common threshold of 5 or 10). This indicates a strong presence of multicollinearity among these variables.

- Unemployment (8.64) is also above the common threshold of 5, suggesting a moderate level of multicollinearity.

- Smoking Prevalence (4.24), Pure Alcohol (3.40), and Obesity (2.18) have VIF values below 5, indicating they are less affected by multicollinearity in this model.

The impact on negative coefficients can be a high multicollinearity which can make the interpretation of individual regression coefficients unstable and potentially misleading. When predictor variables are highly correlated, the model struggles to isolate the unique effect of each variable on the dependent variable. This can lead to unexpected positive or negative signs, inflated standard errors, and the overall statistical power of the model might decrease [5,6]. In the current model, the high VIF values for GDP per Capita, Health Expenditure, Child Mortality, Education (Avg yrs schooling), and PM2.5 Air Pollution suggest that their coefficients (some of which are negative) should be interpreted with caution. So, it's challenging to precisely determine the independent effect of each of these highly correlated variables on 'Life Expectancy'. To address this, Principal Component Analysis (PCA) will be used.

2.4. Principal component analysis

Based on the high VIF scores of the previous chapter, PCA was applied to the following highly correlated

features: 'Health Expenditure', 'GDP per Capita', 'PM2.5 Air Pollution', 'Child Mortality', and 'Education (Avg yrs schooling)'. After these features were scaled, the Principal Component Analysis has been successfully applied to the selected highly correlated features. As a result they were transformed into a smaller set of principal components. The output gave the following variance ratio:

- The PC1 (Principal Component 1) explains 65.63% of the variance in the original five features.
- The PC2 (Principal Component 2) explains an additional 15.35% of the variance.

Together, PC1 and PC2 explain approximately 81% of the total variance in the original PCA features. This means that by using just these two components, we retain a significant amount of information from the original five features. The loadings for PC1 and PC2, are presented to the following Table 3.

Table 3. The loadings for PC_1 and PC_2.

Feature	PC1	PC2
Health Expenditure	0.489190	0.403190
GDP per Capita	0.474301	0.380021
Air Pollution	-0.429355	0.439111
Child Mortality	-0.394383	0.683411
Education (Avg yrs schooling)	0.442547	0.182077

The interpretations from the above are as follows:

- For PC1, it appears to be a general 'Socio-Economic Development and Health Access' factor. Higher values of PC1 would correspond to higher GDP, more health expenditure, better education, and lower air pollution and child mortality. This component captures the general well-being and development aspects of a country.
- For PC2, it seems to be largely driven by 'Child Health and Environmental Quality'. Higher values of PC2 would primarily indicate higher child mortality and higher air pollution, even if other factors like health expenditure and GDP also contribute positively but less strongly. This component might distinguish countries facing significant challenges in child health and environmental quality, possibly independent of the overall development captured by PC1.

The variance ratio of the rest are PC3: 0.1001, PC4: 0.0633, and PC5: 0.0268. The number of chosen_components was 2, meaning the five original highly correlated features have been replaced by two new principal components: PC1 and PC2.

The X train transformed and X test transformed dataframes now have a reduced number of columns, and include PC1 and PC2 alongside the features that were not part of the PCA (e.g., 'Unemployment', 'Obesity', 'Pure Alcohol', 'Smoking Prevalence'). This transformation has effectively addressed the multicollinearity issue by creating uncorrelated principal components.

The next step would be to re-run the Linear Regression model using these transformed datasets to see how the model performance and coefficients change.

2.5. Linear Regression model (using the PC1 and PC2 variables)

The Since PCA has created a cleaner environment for the other variables to express their individual relationships

with 'Life Expectancy', and this provided a more stable and interpretable regression model. The results of this new model are:

For Model Evaluation: MSE 7.02 and for RMSE 2.65. On average, the model's predictions for 'Life Expectancy' are off by about 2.65 years, which is a relatively small error given the scale of life expectancy values. Finally the R-squared is 0.82, and this indicates that approximately 82% of the variance in 'Life Expectancy' can be explained by the features included in the model (the two principal components and the remaining original features). This is a strong fit, suggesting the model is good at predicting life expectancy.

The Model Coefficients are shown in Table 4.

Table 4. New Regression Analysis-Model Coefficients.

Feature	Coefficient	Feature	Coefficient
PC1	2.04	Unemployment	-0.10
PC2	-2.27	Pure Alcohol	-0.42
Smoking Prevalence	0.39	Obesity	-0.16

It follows the interpretation of coefficients:

- For PC1 (Coefficient: 2.04): A one-unit increase in PC1 is associated with an increase of approximately 2.04 years in 'Life Expectancy', holding other factors constant. This is a strong positive relationship.
- PC2 (Coefficient: -2.27): A one-unit increase in PC2, representing 'Child Health and Environmental Quality', is associated with a decrease of approximately 2.27 years in 'Life Expectancy', holding other factors constant. This is a strong negative relationship. A higher PC2 score actually represents worse child health and higher air pollution, which would logically lead to lower life expectancy, thus explaining the negative coefficient.
- Smoking Prevalence (Coefficient: 0.39): This still shows a positive coefficient, which remains counter-intuitive. It suggests that a one-unit increase in smoking prevalence is associated with a 0.39-year increase in life expectancy, which contradicts general health knowledge. This could indicate a complex interaction not fully captured by the model or other confounding factors.
- Unemployment (Coefficient: -0.10): A one-unit increase in 'Unemployment' is associated with a decrease of 0.10 years in 'Life Expectancy', which aligns with expectations.

- Obesity (Coefficient: -0.16): A one-unit increase in 'Obesity' is associated with a decrease of 0.16 years in 'Life Expectancy', which is also expected.

Pure Alcohol (Coefficient: -0.42): A one-unit increase in 'Pure Alcohol' consumption is associated with a decrease of 0.42 years in 'Life Expectancy', which is consistent with known health impacts.

- Intercept (86.44): This is the predicted 'Life Expectancy' when all features are zero. In this context, with scaled and transformed features, it doesn't have a direct interpretable meaning on its own but is part of the regression equation.

Finally, a scatter plot was generated to visualize the model's predictions against the actual values of Life Expectancy.

In the Fig. 6, the red dashed line represents a perfect prediction scenario, where predicted values exactly match

actual values. It can be observed that most of the points cluster closely around the red dashed line, especially in the middle range of life expectancy. This indicates that the model generally makes accurate predictions for a significant portion of the data.

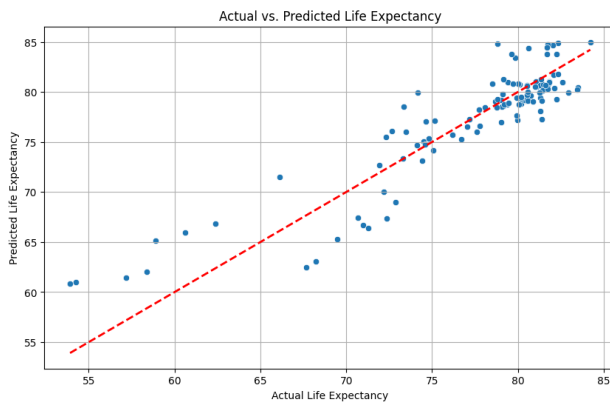


Fig. 6. Actual vs. Predicted Life Expectancy.

On the other hand, there's some spread around the line, particularly at the lower and higher ends, suggesting that the model might have slightly less accuracy for extreme life expectancy values. However, considering the R-squared of 0.82, the model provides a strong fit overall.

In order to improve forecasts, especially for extreme values, a Random Forest regressor model we trained.

2.6. Random Forest model

Random Forest is a powerful model capable of capturing nonlinear relationships and interactions in the data [7]. The model was successfully trained, and its evaluation shows a significant improvement over the linear regression model.

The Random Forest achieved an R-squared of 0.98, which is much higher than the 0.82 of the linear regression. This means that 98% of the variability in 'Life Expectancy' is explained by the Random Forest model.

Furthermore, the RMSE decreased significantly from 2.65 to 0.87. This indicates that, on average, the Random Forest model's predictions are less than 1 year off from the actual values, which is an excellent improvement.

These results confirm that the Random Forest model is much more accurate in predicting 'Life Expectancy,' especially for outliers, due to its ability to capture non-linear relationships and interactions.

Also, the feature importances are PC1=0.65, PC2=0.02, Unemployment=0.20, Obesity=0.07, Pure Alcohol =0.04, and Smoking Prevalence=0.02.

The PC1, representing socioeconomic development and access to healthcare, is by far the most important feature for predicting 'Life Expectancy' in the Random Forest model, as well as in linear regression.

Unemployment and obesity follow in importance, while PC2, alcohol consumption, and smoking have a smaller effect in the Random Forest model. Fig. 7 scatter plot visualize the model's predictions.

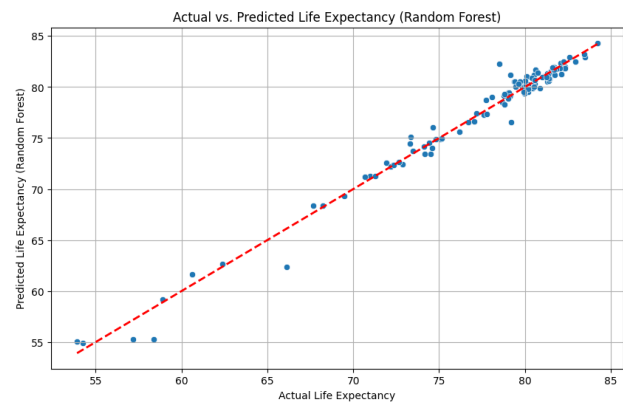


Fig. 7. Actual vs. Predicted Life Expectancy(Random Forest).

The final 'Actual vs. Predicted Life Expectancy (Random Forest)' chart shows that the data points are now much closer to the red dashed line across the entire range of values, including the outliers.

3. Conclusion

This study has completed an extensive analysis of the data, from data cleaning to the training and evaluation of two different regression models for predicting life expectancy.

Initially, the investigation of correlations revealed strong relationships between variables, with the strongest positive correlation to be observed between 'GDP per capita' and 'Health expenditure' (0.85), indicating that countries with higher GDP tend to have higher health expenditures.

With regard to correlations with Life Expectancy, strong negative was observed with 'Child Mortality' (-0.87) and 'PM2.5 Air Pollution' (-0.59). This is to be expected, as high child mortality and poor air quality are linked to reduced life expectancy. In contrast, Life Expectancy showed a strong positive correlation with 'Health Expenditure' (0.56) and 'GDP per Capita' (0.56), highlighting the importance of economic prosperity and health investments.

The revised linear model with PCA had an R-squared of 0.82 and RMSE of 2.65, suggesting a good level of prediction, though with some weakness in extreme values.

Random Forest Regressor was implemented to further improve predictions, using the same variables, including PC1 and PC2. The Random Forest model demonstrated excellent improvement, with R-squared 0.98, and RMSE 0.87.

The overall analysis highlights the strong impact of socio-economic development, health investments, child mortality, and environmental quality on life expectancy. While linear regression provided a good initial understanding, the complexity of the relationships required a more robust model. The Random Forest Regressor proved to be highly effective, providing a very accurate and reliable model for predicting life expectancy, successfully addressing both general trends and extreme values in our data.

Some approaches for further investigation of the issue could involve the use of polynomial regression, residual analysis, weighted regression, and the enrichment of the data, in terms of the number of variables and their quality.

References

- [1] Dattani, S. Rodés-Guirao, L. Ritchie, H. Ortiz-Ospina, E. & Roser, M. (2023) Life Expectancy. Published online at OurWorldinData.org., <https://ourworldindata.org/life-expectancy>.
- [2] Gazilas, E. T. (2024) Factors Influencing Life Expectancy in Low- Income Countries: A Panel Data Analysis. *Journal of Applied Economic Research*.
- [3] World Bank's open data platform. Available online:<https://www.worldbank.org/ext/en/home> (accessed on 10 April 2026).
- [4] Malarvizhi, R. Thanamani, A. (2012) K-Nearest Neighbor in Missing Data Imputation. *International Journal of Engineering Research and Development* . Volume 5, Issue 1.
- [5] Tsagris, M. Pandis, N. (2021) Multicollinearity. *Statistics and research design*. Volume 159, Issue 5, 2021.
- [6] Shrestha, N. (2020) Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*, Vol. 8, No. 2, 39-42.
- [7] Auret, L. Aldrich, C. (2012) Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, Volume 35, Pages 27-42.