Journal of Mathematics and Computing Science, 2024,
Volume 10, No 2, 52-63

# The Impact of Socio-Health Factors Towards Life Expectancy across Countries using Machine Learning

**Mohamad Hafizuddin Roslan[1], Siti Nur Kamariah AbRasid[2], Alfi Tristan Al-Tavip[3], Nurzulaikha Abdullah[4*] and Fakhitah Ridzuan[5]**

[1]Faculty of Data Science and Computing, Universiti Malaysia Kelantan, 16100, Kota Bharu, Kelantan
[2]Fakultas Ilmu Komputeran, Mercu Buana University, Jl. Meruya Selatan No. 1 Kembangan,11650 Jakarta Barat, Indonesia

Authors' email: s21a0025@siswa.umk.edu.my, s21a0055@siswa.umk.edu.my, s23m0260@siswa.umk.edu.my, nurzulaikha.mal@umk.edu.my* and fakhitah.r@umk.edu.my

[*]Corresponding author

Received 25 October 2024; Received in revised 26 November 2024; Accepted 10 December 2024
Available online 15 December 2024

**Abstract**: Life expectancy is a critical indicator of societal well-being and quality of life. Discovering and discussing the numerous factors that contribute to variations in life expectancy is crucial. Understanding these factors is important for shaping policies and interventions aimed at improving population health. With the advancement of the technology, prediction using machine learning is one of the alternatives in discovering the predictive factors that impact life expectancy. Therefore, the objective of this study is to identify and analyse the socio-health factors influencing life expectancy across countries using machine learning techniques. The study found that age group, immunisation status, and the presence of diseases such as HIV/AIDS were significant predictors of life expectancy. These insights are important for policymakers' public health strategies and resource allocation.

**Keywords**: Contributing factor, Health, Life expectancy, Random forest, Regression

## 1 Introduction

The word "life expectancy" was defined as the length of time that a living thing, especially a human being, is likely to live based on the Cambridge Dictionary [1]. The study of the contributing factors towards life expectancy has been the subject of a variety of research in the last few decades. However, with the advancement of the technology, there are still many problems associated with these issues that need to be studied and analysed in order to find solutions [2, 3].

One of the primary interests of medical research and national public health profile indicators is the extension of life expectancy, where it exhibited patterns of continuous growth over time with high variability between countries over the years [2,3]. The changes in life expectancy can be the result of changes in many factors, including health, environmental, and economic development factors [4-6]. For adults in the United States (US), adopting a healthy lifestyle contributes to the reduction of premature mortality and life expectancy extension [7], other than occupation and wage affecting socio-economic variation in life expectancy [8]. Life expectancies were associated with changes in income [9], advanced education, experienced better health outcomes and health satisfaction [10], healthcare expenditures, healthcare resources, mortality rates, the prevalence of Human Immunodeficiency Virus (HIV), and health outcomes [11].

Besides, annual pharmaceutical expenditures, decreasing tobacco consumption, or increasing consumption of vegetables and fruits can also increase life expectancy [11]. There was a report stating that people are healthy and live longer where the average life expectancy was estimated to increase by 7 years until 2025 from 1997, with life expectancy of 80 years by 26 countries. However, the variations in life expectancy still existed between countries of high and low-income groups, which may include ASEAN and other developing countries. Furthermore, investigating related contributing factors helps governments to suggest alternatives in increasing life expectancy across the country.

A few studies on the impacts of life expectancy, which used spatial Durbin, standard error regression estimation models with least square model likelihood estimation models, bootstrap categorical models, geodetectors, correlations, and simple regression models for analysis, believed that multiple regression models provide more in-depth support for future studies and have been widely used [4]. Karacan et al. [12] suggest that decision trees using Chi-square Automatic Interaction Detector (CHAID) techniques were used as they decide the most significant predictor using the Chi-square test, focusing on several components of the decision tree, which include the root node, parent node, child nodes, and leaf nodes, involving all of the important target variables in the data. Pisal et al. [13] suggest that data mining and classification are some of the suitable analyses for predicting life expectancy. While Pisal et al. [13] used the same data as the present study, it focused on the ASEAN population only.

Investigating these contributing factors is essential for governments to propose effective strategies to enhance life expectancy. Previous studies have explored machine learning techniques for predicting life expectancy, recognising its significant impact on social and financial structures worldwide. Huang et al. [14] analysed the socio-economic and healthcare factors contributing to life expectancy in prefecture-level cities in China using classical ordinary least-squares regression and geographically weighted regression on data from the latest census. While these studies provide valuable insights, there is a need for more comprehensive analyses using advanced machine learning techniques.

Therefore, this study aims to explore the socio-health factors affecting life expectancy across countries using machine learning methods. By understanding the relationships between healthcare spending, lifestyle behaviours, education, and mortality rates, this research can offer valuable insights for policymakers and healthcare professionals. These insights can guide resource allocation, public health campaigns, and interventions to improve global health outcomes.

## 2 Methodology

### *A Data Selection*

In this research, the data chosen is the dataset available for the public at https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who [15]. The data appears to be derived from the World Health Organisation (WHO) or a comparable health-focused organisation. It covers a broad range of indicators pertaining to health, development, and socioeconomic aspects in different nations. Nevertheless, the graphics do not provide detailed information regarding the dataset or database utilised. To guarantee precision in citation and the ability to replicate the findings, it is crucial to determine the precise origin of the data.

The dataset consists of 22 columns and 2938 rows, examining many categories like immunisation, mortality, economic, and social aspects from 2000 to 2015. The categorical variables consist of Country (textual), Year (categorical), and Status (developing/developed, likely textual). Conversely, the numerical variables encompass a wide array of health and socioeconomic measures, including life expectancy (in years), adult mortality (number of deaths per 1000 individuals), infant mortality (number of deaths per 1000 live births), alcohol consumption (in litres per capita), percentage of thinness among individuals aged 1-19 years, percentage of thinness among individuals aged 5-9 years, and schooling duration (in years) [16].

Resource allocation was based on income composition. Health expenditure as a percentage of Gross Domestic Product (GDP). The indicators of interest were Hepatitis B immunisation coverage for one-year-old children, measles, immunisation coverage for one-year-old children, polio immunisation coverage for one-year-old children, diphtheria immunisation coverage for one-year-old children, HIV/AIDS prevalence among the population aged 15-49, Gross Domestic Product (GDP) per capita in US dollars, and total population. This vast array of variables offers a comprehensive overview of health and development indicators, enabling a thorough analysis and providing valuable insights.

### *B Data Cleansing*

Data were cleansed for missing data (impute 0) and influential outliers [17,18]. The outlier in a probability distribution function is a number that is more than 1.5 times the length of the data set away from either the lower or upper quartiles. Specifically, if a number is less than Q1−1.5×IQR or greater than Q3+1.5×IQR, then it is an outlier. From the box plot, only data on 'Adult mortality', 'Hepatitis B', and 'Infant Deaths' were problematic and needed to be removed. The first and third quartiles were used to calculate the upper and lower boundaries for outliers. Following that, a new boxplot was constructed for the cleaned data, demonstrating the distribution of 'Adult Mortality', 'Hepatitis B', and 'Infant Deaths' in the absence of outliers. Figure 1 shows the boxplot of adult mortality where the yellow line inside the box indicates the median value.

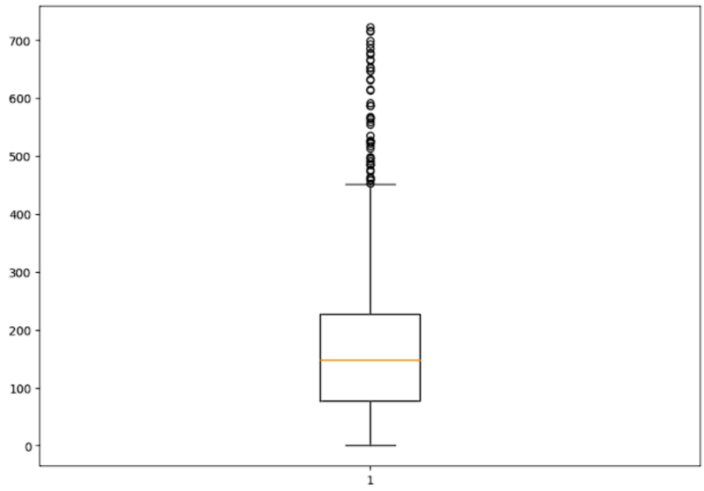


Figure 1: Boxplot of Adult Mortality

### *C Data Exploration*

The analysis demonstrates a direct association between life expectancy and BMI, alcohol consumption, and schooling. This indicates that as these factors grow, life expectancy likewise tends to increase. Countries that have higher average Body Mass Index (BMI) rates, alcohol consumption rates, and education levels tend to have longer life expectancies [19,20]. The correlation coefficients for these associations are remarkably close to 1, suggesting a robust positive correlation and showing a statistically significant linear link between the variables. This suggests that alterations in one variable are linked to proportional alterations in the other. Table 1 shows the correlation between variables towards life expectancy.

Table 1: Correlation between variables towards life expectancy.

| Variables | Correlation coefficient, r |
|---|---|
| **Year** | 0.170 |
| **Adult Mortality** | -0.696 |
| **Infant deaths** | -0.197 |
| **Alcohol** | 0.405 |
| **Percentage expenditure** | 0.382 |
| **Hepatitis B** | 0.257 |
| **Measles** | -0.158 |
| **BMI** | 0.568 |
| **Under-five deaths** | -0.223 |

| | |
|---|---|
| **Polio** | 0.466 |
| **Total expenditure** | 0.218 |
| **Diphtheria** | 0.479 |
| **HIV/Aids** | -0.557 |
| **GDP** | 0.461 |
| **Population** | -0.022 |
| **Thinness 1-19 years** | -0.477 |
| **Thinness 5-9 years** | -0.472 |
| **Income composition of resources** | 0.725 |
| **Schooling** | 0.752 |

There were direct associations between life expectancy and BMI, alcohol consumption, and schooling. This indicates that as these factors grow, life expectancy likewise tends to increase. Countries that have higher average Body Mass Index (BMI) rates, alcohol consumption rates, and education levels tend to have longer life expectancies. The correlation coefficients for these associations are remarkably close to 1, suggesting a robust positive correlation and showing a statistically significant linear link between the variables [21]. This suggests that alterations in one variable are linked to proportional alterations in the other.

The correlation coefficient between life expectancy and BMI is roughly 0.568, suggesting a moderate positive association. Countries with higher average BMIs often exhibit longer life expectancies, while additional factors such as healthcare accessibility and environmental conditions also exert an influence on life expectancy [22]. There is a moderate positive link between life expectancy and alcohol, with a correlation value of approximately 0.405. Increased alcohol consumption may correlate with extended life expectancy under specific circumstances [23].

The correlation coefficient between life expectancy and schooling is approximately 0.752, indicating a more robust positive association than that observed between BMI and alcohol. These findings indicate a positive correlation between greater levels of education and longer life expectancies [24,24]. This relationship can be attributed to various variables like better healthcare accessibility, healthier habits, and higher socioeconomic standing. It is imperative to underscore that correlation does not necessarily indicate causality [25,26].

Based on the data distribution shown in Figure 2, it was observed that the data set has its pattern skewed either skewed to the left, skewed to the right, or not skewed. Data distribution that is skewed to the left is Year, BMI, Polio, Diphtheria, and other income composition of resources, while the data that is skewed to the right is infant death, alcohol, percentage expenditure, and measles. Under five deaths, total expenditure thinness is 1-19 years, thinness 5-9 years. From the data distributions also, the data are not skewed, such as schooling. A left-skewed distribution, also known as negatively skewed, has most data points clustered towards the higher end with an elongated left tail. Natural limits on variable values, as well as the presence of positive outliers, can both influence this asymmetry. Data is skewed to the right, or positively skewed, when most values are concentrated on the left with a longer tail stretching to the right. This skewness might be attributable to natural lower bounds, floor effects, or the impact of negative outliers on the data.

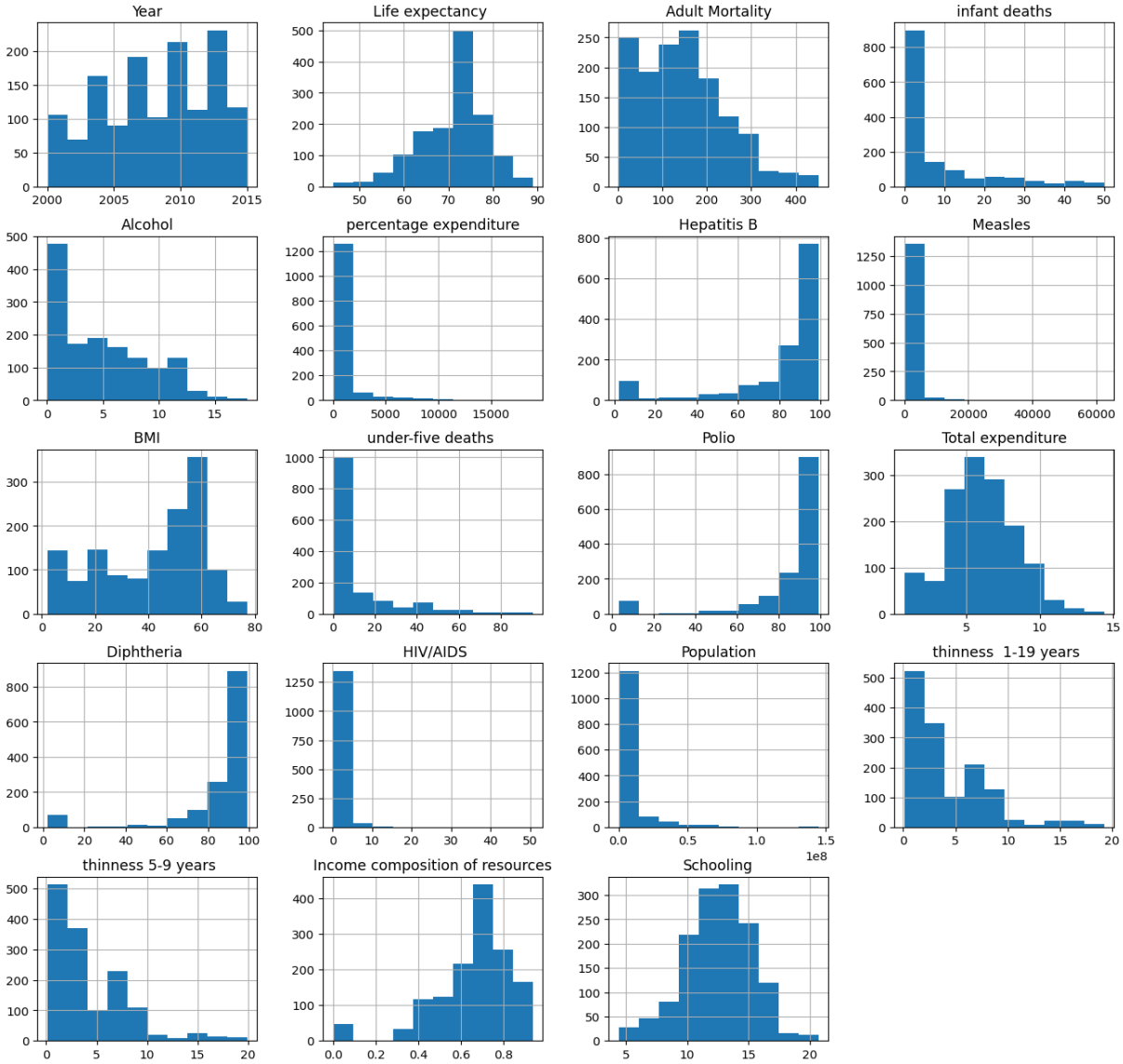


Figure 2 : Data Distribution for all variables

Figure 3 shows the heatmap of the correlation analysis performed between the variables. Each box in the heatmap shows their correlation between the variable on the axis. Correlation in the heatmap also shows their range value from -1 until 1. The value that is close to zero means there are no correlations among two variables. The values that are close to 1 show the positive correlation among the two variables. There are several variables that are close to 1, such as ‘under five deaths and infant deaths’, ‘population and thinness 1-19 years’, ‘population and 5-9 years’, and ‘income composition of resources and schooling’. The heat map shows that the variable which is close to 1 is lighter than the others. Besides that, the correlation is close to -1, which is already represented in a darker colour. In this heat, we already have a diagonal box that shows 1 because of the correlation itself. From the heatmap, the plotting number is not symmetrical among them.

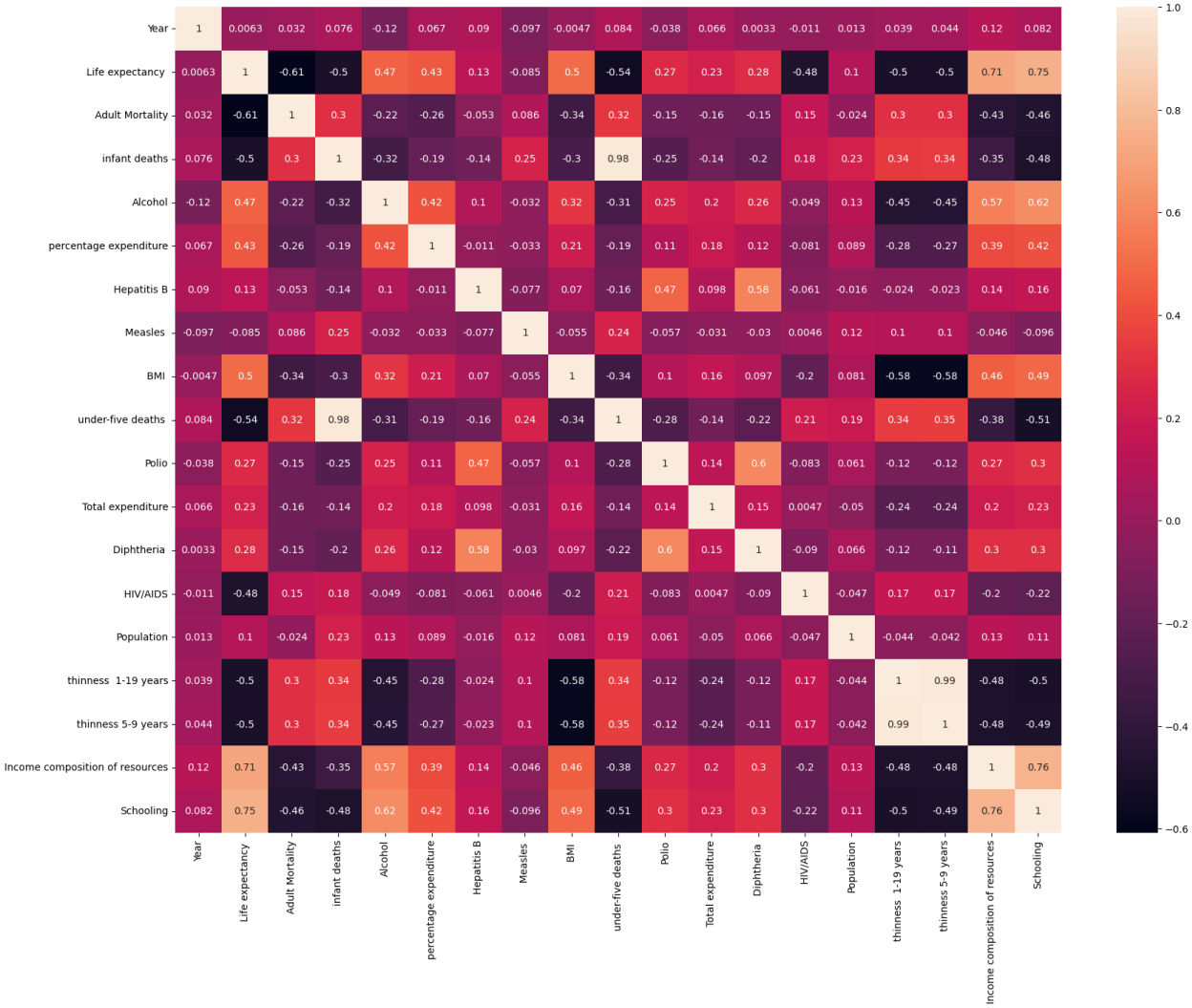


Figure 3: Heatmap of correlation between variables

## *D Data Analysis*

Regression and classification were used in the study to examine the relationship between related variables and life expectancy [26-31]. Both regression and classification are two fundamental methods in statistical and machine learning modelling that serve different purposes and are used in various applications based on the nature of the problem. Classification techniques offer significant insights when it is necessary to categorise life expectancy into separate groups, such as high, medium, and low. Methods such as random forests can detect distinctive characteristics that separate countries according to their projected life expectancy values. This is crucial for implementing customised interventions that concentrate on populations encountering distinct challenges. Adopting both perspectives provides a thorough comprehension of the factors that shape the terrain. Regression analysis quantifies the magnitude of influence of each component, whereas classification allows for the categorisation of countries based on specific policies. This synergistic method optimises the value of the data, resulting in practical insights for enhancing life expectancy in various countries.

Regression is used to model the relationship between using the Python code that utilises essential metrics from the scikit-learn module to evaluate the model's regression performance. The analysis involves the calculation and printing of three specific metrics: the mean squared error (MSE), the mean absolute error (MAE), and the R-squared (coefficient of determination). When evaluating

regression, it is preferable to have lower MSE and MAE values. Conversely, a higher R-squared number indicates a better fit for the model.

A confusion matrix is a table that is often used to evaluate the performance of a classification algorithm on a set of data for which the true values (ground truth) are known. It is a tool that allows the visualisation of the performance of a classification model by displaying the counts of true positive, true negative, false positive, and false negative outcomes. True Positive (TP) is the value of when the instances are positive and were correctly predicted as positive by the model. While True Negative (TN) was when the instances are negative and are correctly predicted as negative by the model. The false positive (FP) is the instance that is negative but was incorrectly predicted as positive by the model. Also known as a Type I error. Then, False Negative (FN) is the instance that is positive but was incorrectly predicted as negative by the model. Also known as Type II error.

## 3 Results and Discussion

### *A Linear Regression*

The computed MSE of 2.53 suggests a rather minor average squared deviation between the observed and projected values. The MAE of 10.50 indicates that, on average, there is an absolute difference of around 10.50 units. The R-squared value of 0.83 indicates that there was 83% of the variability in the data. The reported metric values indicate that the model exhibits great performance, with a high predictive capacity and the ability to explain a significant percentage of the data variability. Figure 4 shows the graph of actual vs predicted values in linear regression.

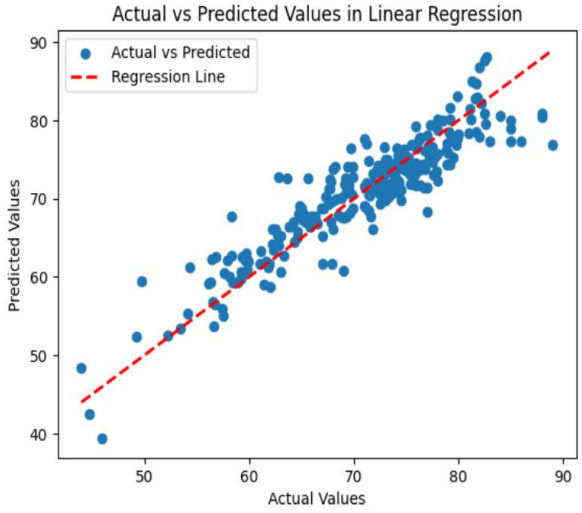


Figure 4: Actual vs Predicted Linear Regression.

Along with metrics, an MSE of 26.19, an MAE of 3.93, and an R-squared value of 0.58 present a thorough evaluation of the effectiveness of a linear regression model. This scatter plot graphically depicts the association between observed and forecasted values, demonstrating a clear positive linear correlation. It suggests that the model may not be adequately capturing the complexity of the data, which could lead to underfitting.

The graph displays a linear regression model represented by a blue line, which aims to forecast values based on the independent variable (x-axis), in contrast to the actual values represented by red dots. The constant occurrence of the red dots below the regression line indicates that the model consistently underestimates the true values. This emphasises the necessity for improvements to boost the accuracy of the model in representing actual relationship between variables. The linear regression line, obtained by the method of least squares regression, aims to minimise the discrepancy between

predicted and observed values. To mitigate underfitting, viable approaches include acquiring supplementary data to enhance the overall comprehension of the connection or utilising a more advanced model. However, the second option presents difficulties in terms of training and interpretation. It is important to acknowledge that even a highly accurate linear regression model may not be able to make perfect predictions because of intrinsic random variation. Nevertheless, a properly calibrated model should provide mistakes that are both random and uncorrelated. However, the presence of correlation in the data depicted in this figure indicates that the model is insufficient in accurately representing the data. Therefore, additional enhancements are required.

The linear regression model may have limitations in capturing the full complexity of the relationships between variables. The presence of underfitting, as indicated by the scatter plot and the lower R-squared value in one of the models, suggests that a more sophisticated model might be required to adequately represent the data. To address this, more advanced machine learning techniques, such as random forests, were adopted, which can better handle nonlinear relationships and complex interactions among the variables.

### *B Classification*

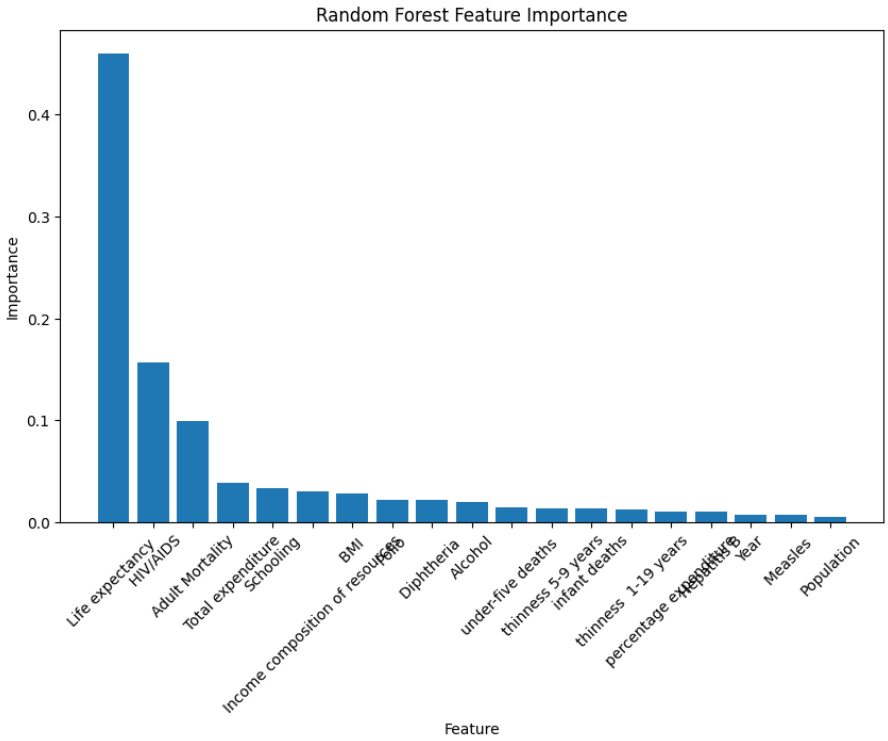


Figure 5: Random Forest Feature importance.

Figure 5 shows the Random Forest feature importance. It is an attribute that shows how important each feature (independent variable) is to the prediction results made by the model. Thus, the higher the importance score, the greater the contribution of the feature to the prediction results. Random Forest, an ensemble learning method that combines forecasts from numerous decision trees, measures the significance of features by their contribution to impurity or error reduction. Average life expectancy, which serves as a complete indication of a population's health and well-being, incorporating aspects such as healthcare quality and lifestyle, is one of three important features that heavily impact projections. AIDS/HIV prevalence is critical, affecting life expectancy directly, showing the model's recognition of its significance in influencing population health. Adult mortality rates, which are indicative of healthcare system efficacy and general population health, also serve as important indicators due to their association with lower life expectancy and poorer health outcomes.

Measles, an infectious disease whose impact on total life expectancy is minimised by successful vaccination programmes, is one of three less influential aspects. Other factors, such as healthcare facilities and socioeconomic conditions, have a more major influence in predicting population size. The variable "year", which captures temporal patterns, is regarded as less important for directly forecasting life expectancy because it may lack the explanatory power of other health-related factors. It is critical to recognise that feature importance is context-dependent, emphasising the importance of taking domain expertise and dataset-specific aspects into account when assessing feature importance.

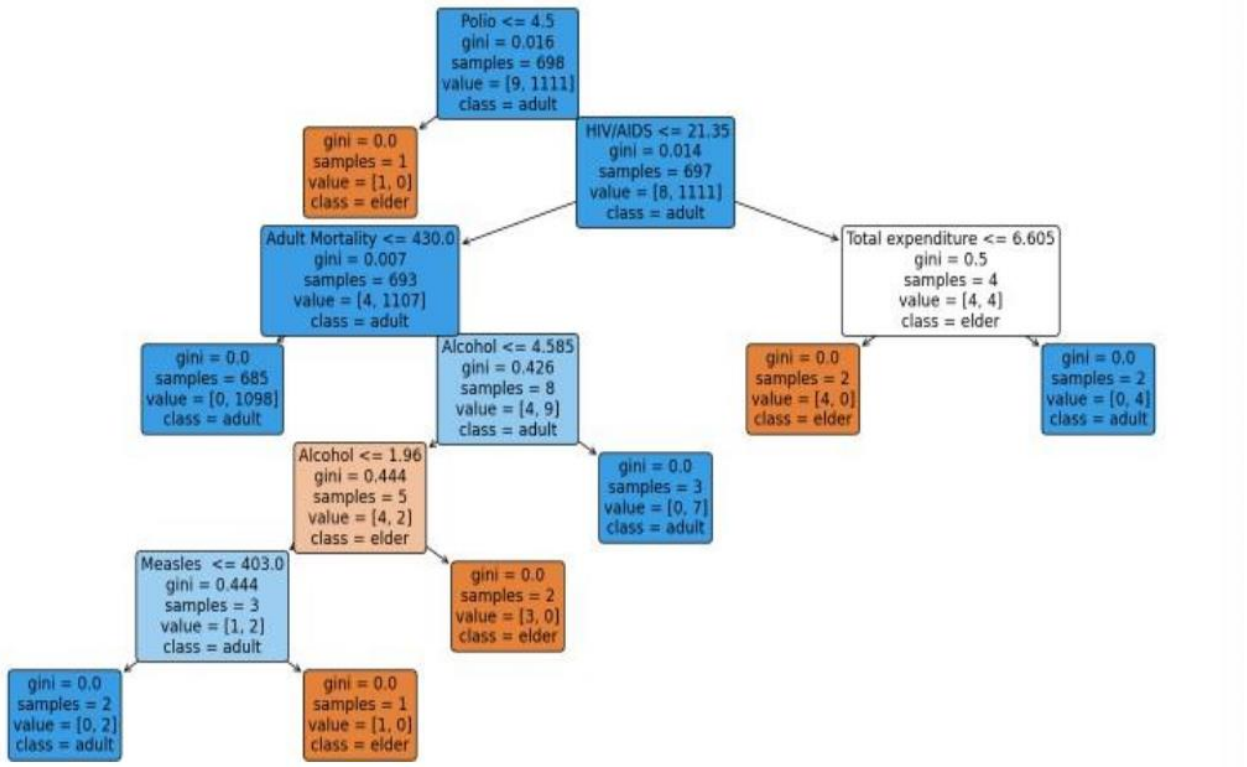


Figure 6: Random Forest Tree Distribution

The provided tree shown in Figure 6 is specifically created to categorise persons into two groups: adults or elders, by considering several characteristics. The tree evaluates the polio vaccination status at the root node. An individual who has contracted polio or whose immunisation status is uncertain is categorised as an adult. Alternatively, the tree evaluates the HIV/AIDS status at the subsequent level. Individuals diagnosed with HIV/AIDS are categorised as elderly, whereas those without the condition are included in the calculation of adult mortality rate.

Additional division takes place depending on the rate of death among adults. If the rate is 430.0 or lower, the tree considers the total spending at the following node. An individual is classed as an adult if their total spending is less than or equal to 6.605; otherwise, they are classified as older. These criteria are applicable to individuals who are polio-free, HIV/AIDS-free, and have an adult mortality rate exceeding 430.0. For individuals belonging to this group, the tree continues to evaluate their alcohol use. If the level of alcohol intake is 4.585 or less, the tree proceeds to evaluate the possibility of measles. An individual is considered an adult if their measles level is less than or equal to 403.0; otherwise, they are classified as an elder.

The last level of branching refers to individuals who are unaffected by polio, do not have HIV/AIDS, have an adult mortality rate exceeding 430.0, and use alcohol at a level higher than 4.585. An individual is categorised as an adult if their alcohol intake is less than or equal to 1.96. Otherwise, they are categorised as older. Decision trees are preferred in the field of machine learning due to their

simplicity and ability to be easily understood. However, they are susceptible to overfitting, which occurs when the model becomes excessively customised to the training data and fails to perform well on new data. The decision tree provided is a reliable technique for categorising persons into certain groups of adults and seniors. It uses several factors such as polio vaccine status, HIV/AIDS status, adult mortality rate, total expenditure, alcohol intake, and measles. The tree's hierarchical structure is composed of distinct levels, each signifying a pivotal juncture in the classification process.

In Level 1, the root node is established based on the individual's polio vaccination status. Individuals who have previously contracted polio or whose immunisation status is unknown are promptly categorised as adults. For individuals who do not have polio, the tree advances to Level 2, where it assesses their HIV/AIDS status. If an individual is diagnosed with HIV/AIDS, they are categorised as an elder. Otherwise, the analysis moves on to Level 3, which examines the adult death rate.

The decision-making process progresses by moving to subsequent levels, which diverge based on certain thresholds or circumstances associated with each feature. Level 3 determines if the adult mortality rate is 430.0 or below, which helps with the categorisation. This hierarchical division persists until the last tiers, where the categorisation is ultimately established according to the values of attributes such as overall spending, alcohol intake, and measles cases.

Random forests are highly regarded in the field of machine learning due to their simplicity and interpretability, which enables users to quickly understand the process of decision-making. Nevertheless, a significant obstacle is the risk of overfitting, in which the model becomes excessively customised to the training data, thereby impeding its capacity to generalise proficiently to unfamiliar data. The comprehensive analysis of the random forest by levels offers a clear understanding of the sequential classification procedure, facilitating the grasp of how adults and elders are differentiated based on the provided characteristics.

## 4 Conclusion

This research has conducted a thorough investigation into the complex correlation between life expectancy and important factors such as BMI, alcohol consumption, and education. Although BMI has a modest significance score, indicating its relevance, the wealth composition of the population is considered to have less influence. For this study, the evaluation metrics for the classification model demonstrate outstanding performance, with an accuracy of 1.00. These models, when combined with comprehensive evaluation and analysis of feature importance, offer useful insights into the factors that affect life expectancy, and the classification of individuals based on certain attributes. The findings might provide valuable insights for specific interventions and policy decisions, highlighting the significance of continuous improvement for achieving the best model performance.

## References

[1] "Life expectancy", *Cambridge Dictionary*. [Accessed: Feb. 28, 2024].

[2] D. Vagero, "Health inequalities across the globe demand new global policies," *Scand. J. Public Health*, vol. 35, no. 2, pp. 113–115, 2007.

[3] K. Moser, V. Shkolnikov, and D. A. Leon, "World mortality 1950–2000: Divergence replaces convergence from the late 1980s," *Bull. World Health Organ.*, vol. 83, no. 3, pp. 202–209, 2005.

[4] Z. Chen, Y. Ma, J. Hua, Y. Wang, and H. Guo, "Impacts from economic development and environmental factors on life expectancy: A comparative study based on data from both developed and developing countries from 2004 to 2016," *Int. J. Environ. Res. Public Health*, vol. 18, p. 8559, 2021.

[5] C. E. Shen and J. B. Williamson, “Child mortality, women’s status, economic dependency, and state strength: A cross-national study of less developed countries,” *Soc. Forces*, vol. 76, no. 2, pp. 667–694, 1997.
[6] K. Mahfuz, "Determinants of life expectancy in developing countries," *J. Dev. Areas*, vol. 41, no. 2, pp. 185–204, 2008.
[7] Y. Li et al., "Impact of healthy lifestyle factors on life expectancies in the US population," *Circulation*, vol. 138, pp. 345–355, 2018.
[8] F. C. Ingleby et al., "Describing socio-economic variation in life expectancy according to an individual's education, occupation and wage in England and Wales: An analysis of the ONS longitudinal study," *SSM - Population Health*, vol. 14, p. 100815, 2021.
[9] Y. H. Khang et al., "Decomposition of socioeconomic differences in life expectancy at birth by age and cause of death among 4 million South Korean public servants and their dependents," *Int. J. Epidemiol.*, vol. 39, pp. 1656–1666, 2010.
[10] Z. Zimmer and P. Amornsirisomboon, "Socioeconomic status and health among older adults in Thailand: An examination using multiple indicators," *Soc. Sci. Med.*, vol. 52, pp. 1297–1311, 2021.
[11] J. W. Shaw, W. C. Horrace, and R. J. Vogel, "The determinants of life expectancy: An analysis of the OECD health data," *South. Econ. J.*, vol. 71, pp. 768–783, 2005.
[12] I. Karacan, B. Sennaroglu, and O. Vayvay, "Analysis of life expectancy across countries using a decision tree," *East. Mediterr. Health J.*, vol. 26, no. 2, 2020.
[13] N. S. Pisal, S. Abdul-Rahman, M. Hanafiah, and S. I. Kamarudin, "Prediction of life expectancy for Asian population using machine learning algorithm," *Malays. J. Comput.*, vol. 7, no. 2, pp. 1150–1161, 2022.
[14] D. Huang, S. Yang, and T. Liu, "Life expectancy in Chinese cities: Spatially varied role of socioeconomic development, population structure, and natural conditions," *Int. J. Environ. Res. Public Health*, vol. 17, p. 6597, 2020.
[15] "Life expectancy (WHO)," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. [Accessed: Feb. 10, 2018].
[16] "Statistics | Definition, types, & importance," *Encyclopedia Britannica*, Nov. 29, 2023. [Online]. Available: https://www.britannica.com/science/statistics/Random-variables-and-probability-distributions.
[17] "Pandas DataFrame dropna() method"
[18] "Statistics - Outlier function," *TutorialsPoint*”
[19] S. H. Preston, Y. C. Vierboom, and A. Stokes, "The role of obesity in exceptionally slow US mortality improvement," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 115, no. 5, pp. 957–961, 2018.
[20] S. Stenholm et al., "Body mass index as a predictor of healthy and disease-free life expectancy between ages 50 and 75: A multicohort study," *Int. J. Obes.*, vol. 41, no. 7, pp. 975–981, 2017.
[21] R. Schmelzer, "15 common data science techniques to know and use," *Business Analytics*, 2020. [Online]. Available: https://www.techtarget.com/searchbusinessanalytics/feature/15-common-data-science-techniques-to-know-and-use.
[22] C. L. Ranabhat, M. B. Park, and C. B. Kim, "Influence of alcohol and red meat consumption on life expectancy: Results of 164 countries from 1992 to 2013," *Nutrients*, vol. 12, no. 2, p. 459, 2020.
[23] E. Nova et al., "Potential health benefits of moderate alcohol consumption: Current perspectives in research," *Proc. Nutr. Soc.*, vol. 71, no. 2, pp. 307–315, 2012.
[24] H. J. Kim, Y. Kweon, and H. J. Hong, "Characteristics of Korean students advised to seek psychiatric treatment before death by suicide," *Front. Psychiatry*, vol. 13, p. 950514, 2022.
[25] S. Arranz et al., "Wine, beer, alcohol and polyphenols on cardiovascular disease and cancer," *Nutrients*, vol. 4, no. 7, pp. 759–781, 2012.
[26] M. Roser, E. Ortiz-Ospina, and H. Ritchie, "Life expectancy," *Our World in Data*, 2019.
[27] G. Miladinov, "Socioeconomic development and life expectancy relationship: Evidence from the EU accession candidate countries," *Genus*, vol. 76, no. 1, 2020.
[28] M. Luy et al., "The impact of increasing education levels on rising life expectancy: A decomposition analysis for Italy, Denmark, and the USA," *Genus*, vol. 75, no. 1, 2019.

[29] S. S. Meshram, "Comparative analysis of life expectancy between developed and developing countries using machine learning," *2020 IEEE Bombay Section Signature Conference (IBSSC)*, pp. 6–10, 2020.
[30] T. Sharma, A. Sharma, and V. Mansotra, "Performance analysis of data mining classification techniques on public health care data," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 11381–11386, 2016.
[31] A. K. Verma, S. Pal, and S. Kumar, "Prediction of skin disease using ensemble data mining techniques and feature selection method—a comparative study," *Appl. Biochem. Biotechnol.*, vol. 190, no. 2, pp. 341–359, 2020.