ECS 171 Project

Nghi Dao
Computer Science
University of California, Davis
Davis, United States
nmdao@ucdavis.edu

Jeffrey Wang
Computer Science
University of California, Davis
Davis, United States
jfywang@ucdavis.edu

Baron Fung
Computer Science
University of California, Davis
Davis, United States
bfung@ucdavis.edu

Adisak Sangiamputtakoon
Computer Science
University of California, Davis
Davis, United States
asangiamputtakoon@ucdavis.edu

Travis Liang
Computer Science
University of California, Davis
Davis, United States
traliang@ucdavis.edu

***Abstract*—This paper will present the development of a model that can predict the life expectancy of a country given various socioeconomic factors such as GDP, healthcare access, and more. Our goal is to identify the variables that have the most significant effect on life expectancy. The results of this study can provide valuable insights for policy makers, healthcare organizations, and the general population on actions to take in order to increase the life expectancy and the overall well-being of a country. From the analysis of these variables, we hope to create a robust model that these stakeholders can utilize to enhance their decision-making abilities.**

I. Introduction

Life expectancy is influenced by a wide range of factors, including socioeconomic status, lifestyle behaviors, healthcare access, environmental conditions, and more. Therefore, it is important to understand these factors in order to design effective public health policies. Using a dataset provided by the World Health Organization, we will develop a comprehensive model to estimate life expectancy. Through various machine learning techniques, we will be able to develop a model that is both easy to interpret as well as being accurate in predicting life expectancy.

This paper contributes to a better understanding of the intricate interplay between these variables and life expectancy. It highlights the need for adaptable models that can reflect emerging trends and shifts in influencing factors. Ultimately, the predictive model will assist policymakers in crafting data-driven interventions tailored to specific population needs, thereby improving health outcomes globally.

II. Problem Statement

Life expectancy is a fundamental metric used to evaluate the health and overall quality of life of a country. Therefore, it is important to understand the various factors that can influence life expectancy. By doing so, we can aid policymakers as well as the public in making various decisions to improve the average life expectancy. This project aims to find the relationship between life expectancy and various factors, such as health and economics.

III. Background

Life expectancy is a significant indicator of the overall health and well-being of a country. Historically, the global trend has shown a steady increase in life expectancy due to medical advancements, improved living standards, and better access to healthcare. However, stark disparities remain between different regions, as well as between developed and developing countries, which highlights the influence of various socioeconomic, environmental, and lifestyle factors. Here is a list of some of the factors that we identified as potentially being important for life expectancy:

A. Socioeconomic Factors

Socioeconomic status significantly impacts life expectancy, with higher incomes often leading to more healthcare access and healthier lifestyle choices. Additionally, education, employment, and income distribution contribute to the variance in health outcomes. In low and middle income countries, rapid economic changes and income inequality can exacerbate disparities in access to healthcare and essential services. We will capture these factors in our model by using variables such as GDP and years of schooling.

B. Environmental Factors

Environmental conditions also contribute to life expectancy. For example, air and water pollution significantly contribute to respiratory and cardiovascular diseases, while urbanization often brings about better access to services and healthcare but may also lead to increased exposure to pollution. Unfortunately, the WHO dataset that we utilize for this paper does not provide any environmental attributes, so we will not be including this aspect of life expectancy analysis in our paper.

C. Healthcare Access and Quality

Access to healthcare services is crucial in managing health risks and reducing premature mortality. Countries with higher healthcare spending and well-established health systems typically see improved life expectancy outcomes. We will capture these factors in our model by using variables such as government healthcare expenditure and immunization rates.

D. Lifestyle Choices

Lifestyle choices like diet, physical activity, smoking, and alcohol consumption have direct impacts on health. Poor diet and inactivity contribute to the rise of obesity and related chronic diseases, while smoking and alcohol abuse remain significant risk factors for several health conditions. We will capture these factors in our model by using variables such as alcohol consumption and BMI.

Our paper does not break any new ground; indeed, recent advances in data science and machine learning (ML) have already led to the development of predictive analytics in healthcare. Algorithms like decision trees, random forests, and neural networks can help quantify the relative importance of various risk factors and identify high-risk groups that could benefit from targeted interventions. For example, Rajkomar et al. [1] highlighted the potential of ML models in making predictions based on electronic health records. ML models can also analyze individual health data to develop personalized treatment plans, integrating data such as genetic factors, lifestyle habits, and demographic information in order to give health recommendations that improve long-term outcomes. Finally, ML models can also assist in optimizing healthcare resource allocation by predicting patient demand and identifying service gaps. Hospitals can thus improve care delivery, reduce preventable mortality, and enhance life expectancy. All in all, by leveraging ML algorithms, we believe that healthcare professionals can better understand the factors influencing life expectancy and identify key trends to implement more effective public health interventions. We hope that our paper will be a meaningful contribution to this growing research area.

IV. Literature Review

Chen et al. [2] provides valuable insights into the factors influencing life expectancy in developed and developing countries. We summarize some key findings below:

A. Socioeconomic Factors

- **GDP Per Capita**: GDP per capita was found to positively affect life expectancy in developed countries by enhancing healthcare and social services. However, in developing nations, it may negatively impact life expectancy due to uneven wealth distribution.
- **Healthcare Spending**: Increased healthcare expenditure correlates with longer life expectancy, but the efficiency of healthcare systems also plays a role.
- **Income Inequality**: High income inequality leads to shorter life expectancy in developing countries where social disparities exacerbate health issues.

B. Environmental Factors

- **PM2.5 Air Pollution**: Exposure to PM2.5 particulate matter significantly reduces life expectancy in developing countries due to cardiovascular and respiratory health effects.
- **Carbon Dioxide Emissions**: CO2 emissions negatively impact life expectancy due to the many health risks arising from pollution.
- **Urbanization Rate**: Higher urbanization rates positively correlate with life expectancy by improving access to healthcare and living standards.

There have been many advancements in predictive models for life expectancy, especially with the integration of machine learning techniques to handle longitudinal data (i.e. data that repeatedly observes the same variables over time). Yang et al. [3] introduces a dynamic indicator-restricted mean survival time (RMST) model that uses time-dependent covariates to more accurately predict the average survival time of patients at different time points. This method provides a more personalized prediction and adapts to changes in a patient's health status over time, and during testing, the model was found to outperform existing predictive models. This improvement demonstrates the potential of carefully designed ML algorithms in handling critical healthcare challenges.

Machine learning techniques have also been used to predict various diseases, including cancer. These techniques use sophisticated algorithms to uncover patterns in complex healthcare data that traditional statistical methods might miss. For example, Das et al. [4] introduces the Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP), which uses advanced ML techniques such as K-means oversampling in order to predict breast cancer with high accuracy.

Our paper differs from these past works in that we will seek to obtain comparatively simpler models from our data. This simplicity comes with the advantage of increased interpretability of our model, which is beneficial for stakeholders who must understand our model's results before enacting major policies. We will also be using a wider variety of variables in our dataset, ranging from socioeconomic factors like years of schooling to health factors like immunization rates. We hope that this diversity of variables results in a more robust model.

V. Dataset Overview

We will be using a life expectancy dataset provided by the WHO on Kaggle. The dataset contains 18 attributes that can be used to predict life expectancy, along with life expectancy itself. Each attribute then contains data points for 193 different countries over a period of 15 years. In total, there are over 2000 individual datapoints that the dataset provides, which should be large enough to allow us to draw reasonably strong conclusions.

Before we delve into the analysis of the dataset, we will first provide a description of all 19 attributes that the dataset provides. Table I on the next page contains a short description of each variable, along with some statistics that describe each variable's distribution.

VI. Exploratory Data Analysis

Highly Correlated Variables

To ensure our model is robust and less sensitive to noise, we will remove highly correlated variables from the dataset. Highly correlated variables could lead to inaccurate predictions, as the model could overfit to extraneous variables whose contributions to the model can be adequately explained by other, more impactful variables. The variables' correlation matrix is shown in Figure 1. From it, we identified the following pairs of highly correlated variables:

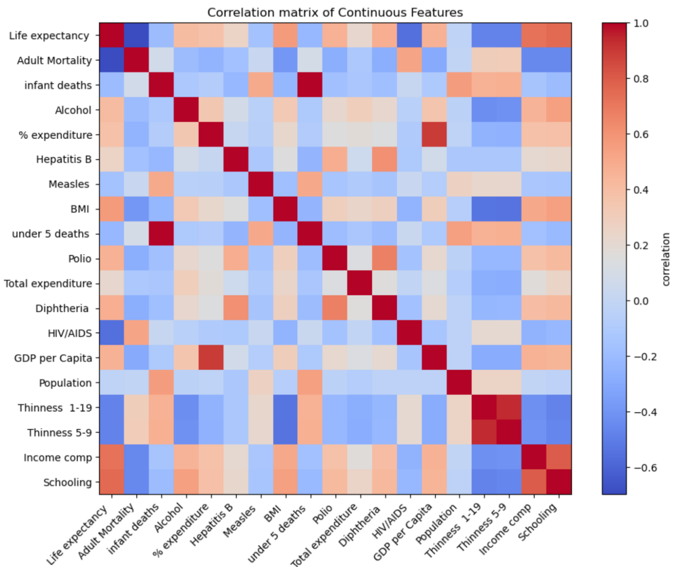

Fig. 1. Correlation matrix of the variables

- **GDP and Percentage Expenditure**: GDP and percentage expenditure have a very high positive correlation, which makes sense since countries who have a stronger economy have more money to spend towards healthcare. We decided not to use percentage expenditure in our model, since GDP appears to correlate better with life expectancy.
- **Infant Deaths and Under-Five Deaths**: The correlation of these two variables makes sense. However, despite these variables being related to mortality rate, they do not appear to have a strong correlation with life expectancy. Nevertheless, we will only include infant deaths in our model.
- **Thinness of Children**: Thinness of people ages 1-19 and ages 5-9 is highly correlated, which makes sense. In fact, as can be seen by the correlation matrix and by their distributions, these two variables are nearly identical. We will only use thinness 1-19 in the model.
- **Schooling and Income Composition**: Schooling and income composition index seem to be highly correlated, and, like the two thinness variables, these two variables also seem to be almost identical. We will remove income composition from the model, as the number of schooling years is more easily interpretable.

Variable	**Mean**	**Median**	**Std**	**25th Q**	**75th Q**	**Description**
Life Expectancy	69.22	72.1	9.52	63.1	75.7	Average lifespan of people in a country; target variable.
Adult Mortality	164.8	144.0	124.3	74.0	228.0	Adult mortality rate per 1,000 people.
Infant Deaths	30.3	3.0	117.9	0.0	22.0	Infant deaths per 1,000 people.
Alcohol	4.6	3.76	4.05	0.88	7.70	Average alcohol consumption per capita, in liters.
Percentage Expenditure	738.3	64.91	1987.9	4.69	441.53	Government health expenditure as a percentage of GDP per capita.
Hepatitis B	80.9	92.0	25.1	77.0	97.0	Immunization coverage (%).
Measles	2419.6	17.0	11467.3	0.0	360.25	Cases per 1000 people.
Under-5 Deaths	42.0	4.0	160.0	0.0	28.0	Deaths of people under 5 years of age per 1,000 people.
Polio	82.6	93.0	23.4	78.0	97.0	Immunization coverage (%).
Total Expenditure	5.94	5.76	2.5	4.26	7.49	Total health expenditure (% of GDP).
Diphtheria	82.3	93.0	23.7	78.0	97.0	Immunization coverage (%).
HIV/AIDS	1.74	0.1	5.08	0.1	0.8	Deaths per 1,000 individuals.
GDP	7.48	1.77	14.27	0.46	5.91	GDP per capita in USD.
Population (millions)	12.75	1.39	61.01	0.20	7.42	Total population in millions.
Thinness 1-19 Years	4.84	3.3	4.42	1.6	7.2	Percentage of thinness in this age group.
Thinness 5-9 Years	4.87	3.3	4.51	1.5	7.2	Percentage of thinness in this age group.
Income Composition of Resources	0.63	0.68	0.21	0.49	0.78	Human development index.
BMI	38.52	38.3	7.8	33.7	43.5	Mean body mass index.
Schooling	11.99	12.3	3.36	10.1	14.3	Average years of schooling.

TABLE I
DESCRIPTIVE STATISTICS OF THE VARIABLES

Skewedness of Variables

The skewedness of the variables has to be taken into account when building a model. Since many machine learning methods rely on the data to not have any extreme distributions, having very skewed data could lead to biased parameters or inaccurate predictions.

As seen from the histograms in Figure 2 (and also by comparing the means and medians in Table I), we can see that most of the variables are skewed. Variables such as thinness and adult mortality have a moderate right skew, while income composition has a moderate left skew. As more extreme examples, variables such as infant deaths, GDP, population, and measles cases have a very strong right skew, while the vaccination variables have a very strong left skew. The remaining variables, such as schooling and total expenditure, appear to be fairly symmetric. Before we can create our model, we will need to find transformations of the skewed variables to make their distributions more symmetric.

Outlier Detection

To ensure that the model does not overfit to any unusual extreme value in the data, we will also have to remove outliers from the data. To detect an outlier, we will use the interquartile range (IQR). A data point is marked as an outlier if it lies above the Upper Quartile + 1.5IQR threshold or below the Lower Quartile - 1.5IQR

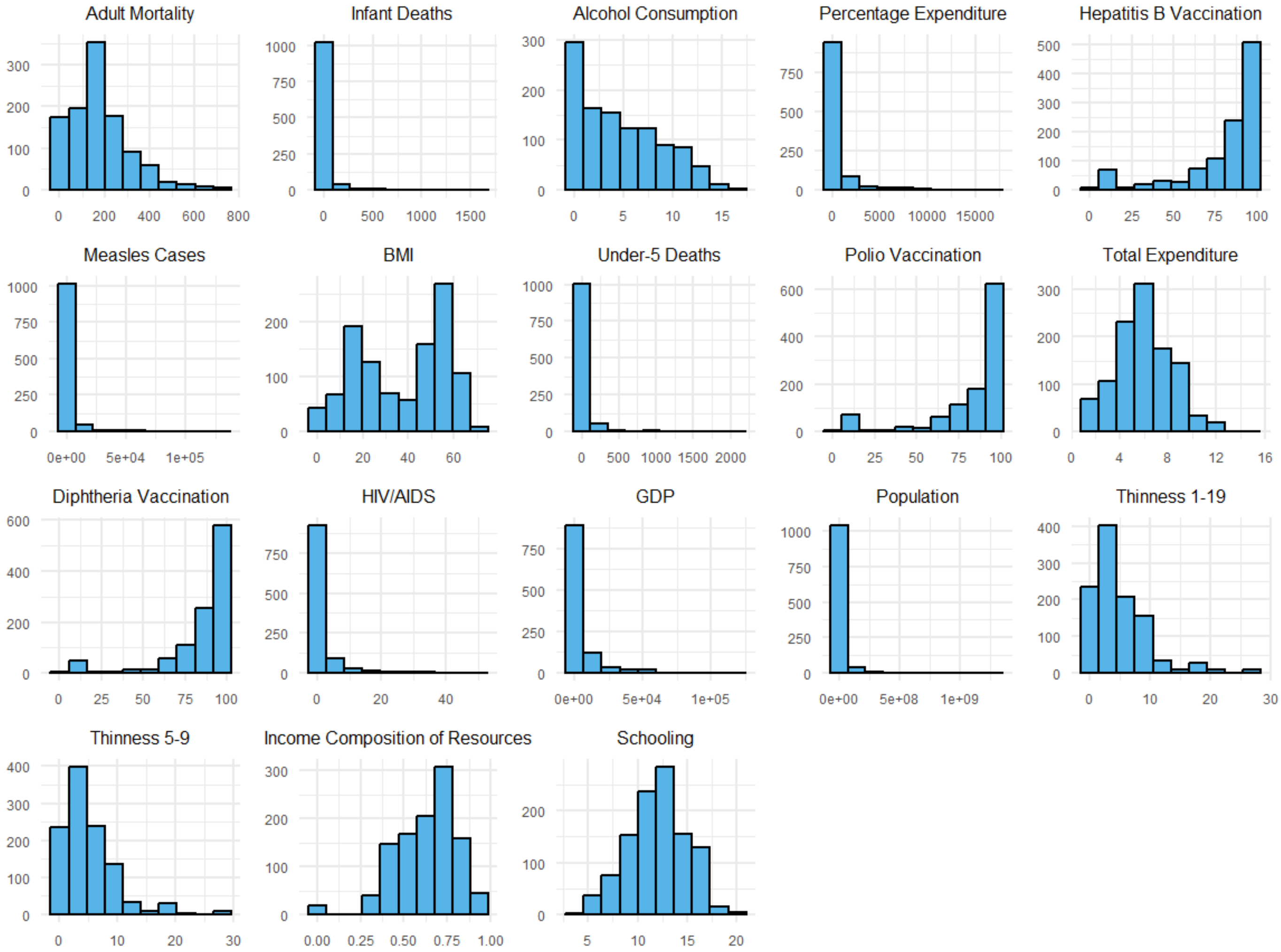

Fig. 2. Histograms of all the variables

threshold. These outliers can be identified visually using boxplots, which are shown in Figure 3.

Unfortunately, since the data is so heavily skewed, the IQR of many variables is very small compared to the standard deviation. As a result, a large proportion of the data points are considered as outliers by this method. Variables such as measles cases and HIV/AIDS deaths show an overwhelming number of outliers. Therefore, in order to avoid throwing out a large portion of our dataset, we will first find transformations of the data so that extreme points become less extreme. After performing such transformations, we can then use the criteria stated earlier for detecting outliers.

VII. Proposed Methodology

Predictors

As we mentioned during exploratory data analysis, we can use the correlation matrix to select variables that have high correlation with life expectancy but low correlation with each other. Using this criteria, we narrowed the list of variables to use down to the following 13 variables:

- Alcohol
- Adult Mortality
- Hepatitis B
- Measles
- BMI
- Polio
- Total Expenditure
- Diphtheria
- HIV/AIDS

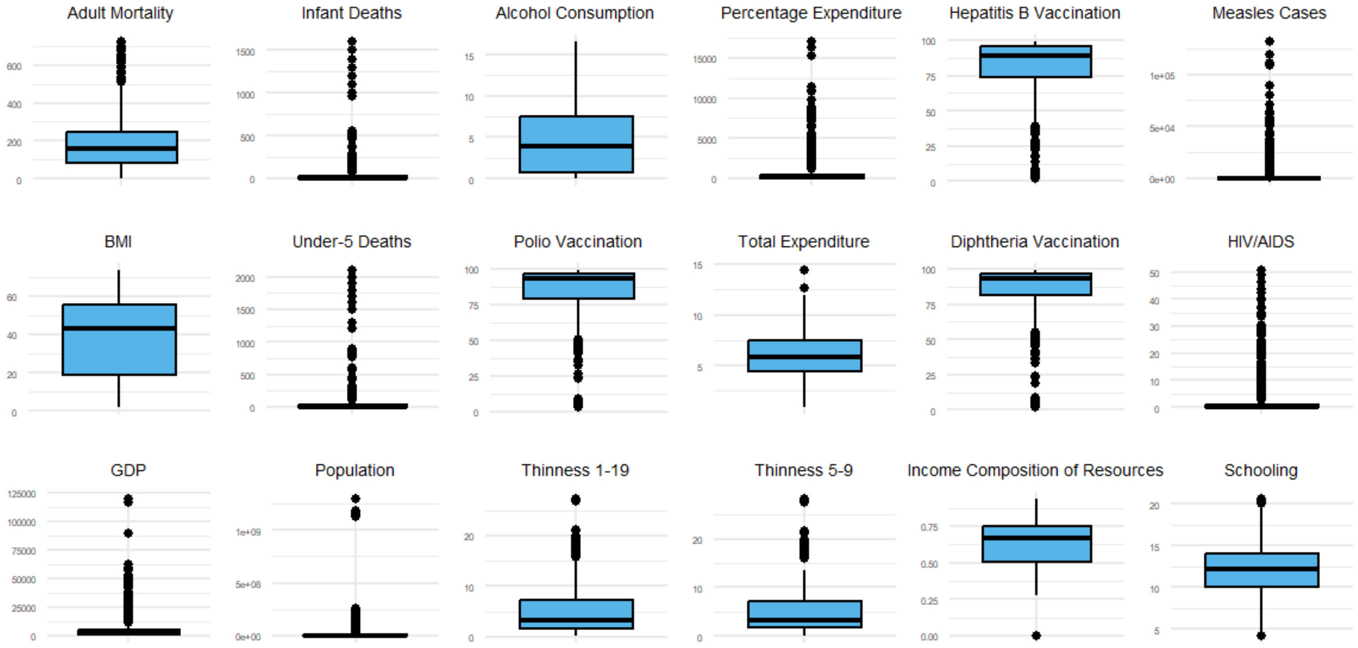

Fig. 3. Boxplots of all the variables

- GDP per Capita
- Thinness 1-19
- Schooling
- Infant Deaths

Linear Regression Model

Since we are striving to create a simpler model, we decided to use multiple linear regression (MLR) for this task. MLR attempts to create a “line of best fit” of form

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where $x_1, \cdots, x_k$ denote the values of the k predictor variables, $\hat{y}$ is the final prediction, and $\beta_1, \cdots, \beta_n$ are coefficients determined by minimizing the quantity

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

assuming that there are n data points in the dataset and that y_i is the actual value of the response variable for data point i. This last quantity is called the residual sum of squares (RSS). The final model generated by MLR is relatively easy to interpret, as one can often simply glance at the values of the estimated coefficients β_i to get a sense of how each predictor variable affects the model.

Data Cleaning

We first cleaned the data by removing all outliers according to the procedure we discussed previously, with any data points falling below the Lower Quartile - 1.5IQR threshold or above the Upper Quartile + 1.5IQR threshold being taken out of the dataset. Next, we noticed that there were quite a few missing values in the dataset. Further inspection showed that these missing values were often coming from smaller, less populated countries, for which measuring and tracking these statistics may perhaps be more difficult. We decided to fill in these missing values with the mean of each statistic, and by removing the outliers first, we were able to ensure that these means were not too affected by any extreme values.

Variable Transformation

In order to give our data properly symmetric distributions, we tried three different transformation methods: log transformation, square root transformation, and inverse transformation. For most of the variables, we were able to transform them to achieve a much more even distribution; the transformations are shown in Table II, and the resulting distributions are shown in Figure 4. However, for some of the variables, like HIV/AIDS and infant deaths, we were not able to find an appropriate transformation. This is because most of the countries have very few infant deaths and HIV/AIDS cases, while

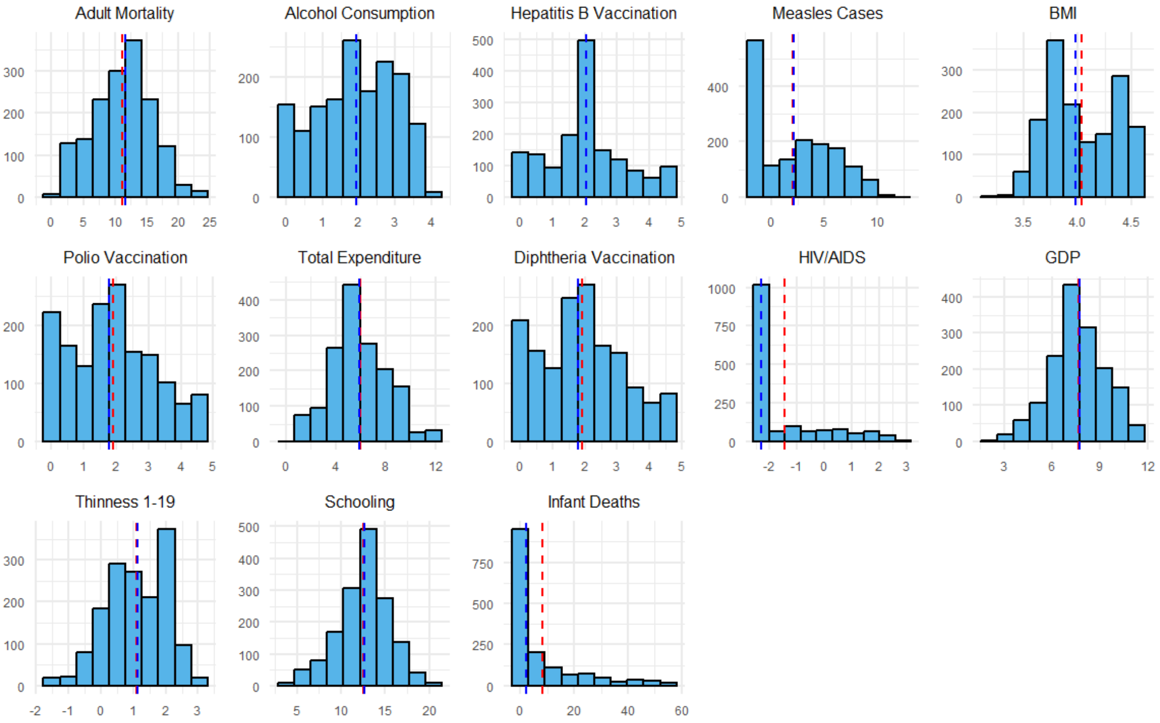

Fig. 4. Transformations of the variables. Red and blue dotted lines are the mean and median, respectively.

Variable	Transformation
Adult Mortality	$\sqrt{X}$
Alcohol	$\sqrt{X}$
Hepatitis B	$\log(100 - X)$
Measles	$\log(X + 0.1)$
BMI	$\log(100 - X)$
Polio	$\log(100 - X)$
Total Expenditure	X
Diphtheria	$\log(100 - X)$
HIV/AIDS	$\log(X)$
GDP per Capita	$\log(X)$
Thinness 1-19	$\log(X)$
Schooling	X
Infant Deaths	X

TABLE II
TRANSFORMATION USED FOR EACH VARIABLE

a few countries have a very high amount, which leads to these variables still being heavily skewed to the right. However, because these variables are highly correlated with life expectancy, we will still use them.

VIII. Experiment and Evaluation

Variable Selection

Although we narrowed down the variables that we will use for the model, we still need to identify which variables are the strongest predictors. This is because simply using all the predictors can lead to overfitting. Since we have only 13 features, it is feasible to train a model against all possible subsets of predictors and then select the model that has relatively high accuracy while also having a minimal number of features. To do this, we first perform an 80%/20% split of the dataset into a training and testing set, and then further split the training set by 80%/20% in order to create a new training set and a validation set. Each model will have a different subset of the parameters and will be trained on the training set, and the best model will be selected as the one with the lowest mean squared error (MSE) on the validation set.

The equation for MSE is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

which is essentially the average of the RSS.

We will perform this selection for the developed countries and developing countries separately, since earlier in the literature review, we found some sources say that life expectancy is influenced by different factors depending on whether a country is developing or developed.

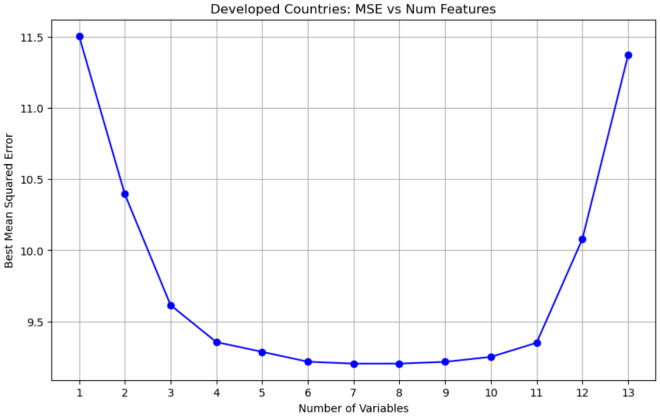

Fig. 5. Variable selection for developed countries

Evaluation

Figure 5 shows the MSE on the validation set of the best model for developed countries for different numbers of variables. We can see that after more than 5 variables are used, there begins to be less decrease in the MSE. In fact, after about 10 variables are used, the model actually begins to overfit to the training data, resulting in a significant increase in the MSE. We also noticed that thinness was the only predictor selected by all the best models, which suggests that it may be among the best predictors of life expectancy for developed countries. Alcohol, on the other hand, was the least used predictor, only being selected once. In the end, based on the graph, we decided to use the model with 3 variables, since there wasn't a significant difference between the MSE of 3 variables and the overall best MSE of 7 variables, and using a minimal amount of variables would make the model easy to interpret and prevent overfitting. Our final model is

$$\hat{y}_{developed} = 77.70 - 0.27x_{AM} + 0.44x_{GDP} - 2.91x_T$$

where

$$\begin{aligned} x_{AM} &= \text{adult mortality rate per 1000 people} \\ x_{GDP} &= \text{GDP per capita in USD} \\ x_T &= \text{percentage of thinness in 1-19 year olds} \end{aligned}$$

Note that each of these variables must be transformed according to the transformation in Table II before they can be fed into the model. Interpreting this model is fairly straightforward: for example, we can see from the equation that an increase of one death per 1000 people for the transformed mortality rate would correspond to a drop of 0.27 years in predicted life expectancy. When we tested this model on the testing set, we obtained an MSE of approximately 9.42, which is even lower than the MSE we got on the validation set. This suggests that our model is fairly robust and relatively accurate on new data.

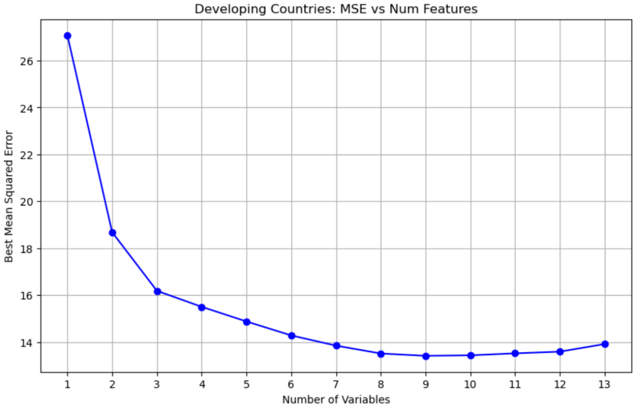

Fig. 6. Variable selection for developing countries

Figure 6 shows the same thing as Figure 5, but for developing countries instead. The first thing we noticed is that the overall MSE of the developing countries' model is higher than that of the developed countries. Additionally, the MSE of the models only starts to stagnate at around 5-7 variables, and only starts to overfit at around 11-12 variables. This suggests that the life expectancy of the developing countries are much more prone to fluctuation and are influenced by a wider range of factors as compared to the developed countries. Additionally, while the most significant variable for the developed countries was thinness, we found that thinness was not as strong of a predictor of life expectancy for developing countries. Instead, the number of HIV/AIDS deaths was the most significant predictor of life expectancy for developing countries, as it was selected by all the best models. Another significant variable was number of schooling years, while the least significant predictor was total expenditure on healthcare. Overall, we used similar criteria to the developed countries model in selecting the 7 variable model as our final model for the developing countries, as we felt it struck the right balance between lowering MSE and maintaining simplicity. Our final model is

$$\begin{aligned}\hat{y}_{developing} &= 56.34 - 0.26x_{AM} + 0.93x_{HB} - 0.16x_M \\ &\quad -1.20x_D - 2.33x_{HA} + 0.46x_{GDP} + 0.88x_S\end{aligned}$$

where

$$\begin{aligned}x_{AM} &= \text{adult mortality rate per 1000 people} \\ x_{HB} &= \text{hepatitis B immunization coverage \%} \\ x_M &= \text{number of measles cases per 1000 people} \\ x_D &= \text{diphtheria immunization coverage \%} \\ x_{HA} &= \text{number of HIV/AIDS deaths per 1000 births} \\ x_{GDP} &= \text{GDP per capita in USD} \\ x_S &= \text{number of years in school}\end{aligned}$$

Similar to the developed model, we again note that these variables must first be transformed according to the transformations in Table II. We can also interpret this model similarly to the developed countries model—an increase of one death per 1000 people for the transformed mortality rate would result in a similar decrease in predicted life expectancy of about 0.26 years. Finally, testing our model on the test set achieves similarly great results, with an MSE of 11.88 that is even lower than the MSE on the validation set.

IX. Conclusion and Discussion

One thing that was notable about our results is that, besides adult mortality, there are no other variables that the best model for developed countries and the best model for developing countries have in common. This suggests that developing and developed countries' life expectancy are influenced by significantly different factors. Many of the explanatory variables for the developing countries' model are related to disease—hepatitis B vaccination, measles cases, etc.—while the same such variables are absent in the developed countries' model. Conversely, the developed countries' model contains variables such as "thinness 1-19 years" that are absent from the developing countries' model. These results seem to confirm that developed countries and developing countries should seek to address different issues in order to improve life expectancy, which is in line with both our initial suspicions as well as the prevailing literature [2].

One explanation for this difference in explanatory variables is that the diseases listed here are far more prevalent in developing countries than in developed countries. Hepatitis B, measles, and diphtheria are all preventable with vaccines, and HIV/AIDS can be managed with treatment, but the high costs associated with these preventative measures means that developing countries often cannot afford to implement them. Developing countries are thus more susceptible to these diseases, and so it makes sense that these diseases have a stronger impact on life expectancy for them. It also makes sense that these same diseases are a non-factor for developed countries that actually have the resources to prevent and treat them, and so life expectancy in these countries is determined by other variables such as child thinness.

Most of our final estimates for the coefficients seem to make sense. For example, in the developed model, a one unit increase in transformed GDP would result in a higher predicted life expectancy, which is in line with what we would expect. Countries with better economies should, intuitively, be able to spend more money towards improving the health of their citizens. As another example, in the developing model, a one unit increase in transformed measles cases would result in a lower predicted life expectancy, which is again an expected result. The one exception is diphtheria immunization rate for the developing model, which leads to a decrease in life expectancy as it increases. We would expect higher immunization rates to prevent disease and increase life expectancy, so this might be an indication that our model is overfitting to a trend in the training data that is not indicative of the actual trend.

Our project has only scratched the surface of life expectancy analysis, and there are many possible avenues of future research in this area. For example, it may be worth examining other non-infectious diseases such as diabetes and cancer to see if they have the same effect as infectious diseases. Alternatively, one could focus on exploring variables not directly related to health, such as environmental factors, to see if they produce as strong of a model. Finally, it may be interesting to divide the countries into more than two categories, such as by GDP brackets, and see if stronger conclusions can be drawn then.

X. Project Roadmap

Wk 2 **Describing the Problem**: Perform basic background research to gain a more solid understanding of life expectancy and the factors that may affect it.

Wk 2 **Background Study**: Read and review already-existing papers to find methodologies that can be used for our project, as well as to find ways in which our project can be unique.

Wk 3 **Data Acquisition**: Decide to use the "Life Expectancy (WHO)" dataset, which contains the life expectancy and 21 other attributes of 193 different countries over the span of about 15 years (from 2000 to 2015).

Wk 3 **One-Page Project Report**: Complete preliminary report about our project.

Wk 4 **Exploratory Data Analysis (EDA)**: Take a deep look at the dataset for missing values, variable distributions, etc. Analyze trends and correlations between life expectancy and other variables in order to get a sense of which variables to include in model.

Wk 5 **Data Preprocessing**: Address missing data, normalize distributions, and throw out any outliers.

Wk 5 **Mid-Quarter Project Report**: Complete mid-quarter report about our progress on the project.

Wks 5-7 **Model Development/Testing**: Examine various algorithms, such as regression models and other methods, to predict life expectancy. Test their performance on validation and test datasets.

Wk 8 **Write Research Paper**: Document the research process, methodologies, model development, results, and implications in the research paper.

Wk 9 **Develop Web-Based Demo**: Use Streamlit to create a simple web-based demo of the final model.

Wk 10 **Record Video**: Record a short 5-minute video demonstrating our web demo and summarizing what we have done in this project.

XI. Assignment of Tasks

- **Nghi Dao**: Background study, data acquisition, model development/testing, project writeup.
- **Jeffrey Wang**: Data preprocessing, model development/testing, project writeup.
- **Baron Fung**: Exploratory data analysis, data preprocessing, recorded video.
- **Adisak Sangiamputtakoon**: Exploratory data analysis, data preprocessing.
- **Travis Liang**: Background study, data acquisition, model development/testing, web-based demo.

XII. Source Code

The source code for this project can be found in this Github repository.

References

[1] A. Rajkomar, J. Dean, and I. Kohane, ”Machine Learning in Medicine,” New England Journal of Medicine, vol. 380, no. 14, pp. 1347-1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[2] 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, no. 16, pp. 8559, Aug. 2021, doi: https://doi.org/10.3390/ijerph18168559

[3] Z. Yang, Z. Li, et al., ”A Dynamic Prediction Model Supporting Individual Life Expectancy Prediction Based on Longitudinal Time-Dependent Covariates,” IEEE Journal of Biomedical and Health Informatics, vol. A247, pp. 529–551, Sep. 2023, doi: 10.1109/JBHI.2023.3292475

[4] A. K. Das, P. Jena, A. K. Sahoo, S. Dehuri, and S. C. Satapathy, ”Machine Learning Based Intelligent System for Breast Cancer Prediction (MLISBCP),” *Expert Systems with Applications*, vol. 224, 120070, 2023. https://doi.org/10.1016/j.eswa.2023.122673