Key Determinants of Global Life Expectancy Trends

By: Shane Gao, Kian Ghaffari, Tristan Kao, Amir Yaacoobi

Introduction

In this exploration, we plan on finding what the main determinants are in global life expectancy trends, starting from the year 2000 up until 2016. We focus on the specific health behaviors and socioeconomic factors that play a significant role in how life expectancy is determined worldwide. We plan on creating visualizations for exploratory data analysis and several different models that we have learned about throughout the course. Our goal is to understand what variable plays the most important role in determining the life expectancy of a population of people. We plan on utilizing several different models that we learned about in this class, starting with multiple linear regression, single decision tree, cross-validation, and random forests.

Data

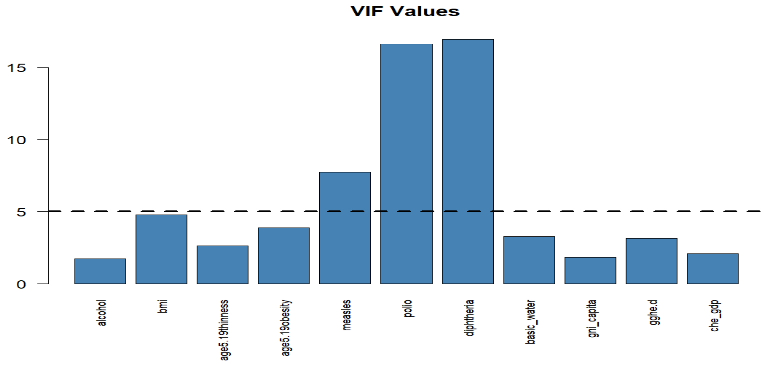
The dataset we use in this exploration is the WHO National Life Expectancy Dataset on the Kaggle database. This dataset contains data from every country, starting in 2000 and ending in 2016. The information collected in this dataset stems from two global organizations, the GHO (Global Health Organization) and the UNESCO (United Nations Educational Scientific and Cultural Organization).

GHO Predictors:

country - country name,
country_code - Three-letter country identifier,
region - Global region of the country,
year - year, Life_expect - life expectancy at birth,
life_exp60 - life expectancy at age 60,
infant_mort - death rate up to age 1,
age1-4mort - death rate between ages 1-4,
alcohol - consumption of alcohol in liters,
bmi - mean BMI,
age5-19thinness - prevalence of thinness among children,
age5-19 obesity - prevalence of obesity among children
adult_mortality - mortality rate of both sexes, per 1000 population,
hepatitis - hepatitis B immunization among 1 year olds %,
measles - measles-containing vaccine first dose among 1 year olds %,
polio - polio immunization among 1 year olds %,
diphtheria - diphtheria immunization among 1 year olds %,
basic_water - population using at least basic drinking water services,
doctors - medical doctors per 10,000, Hospitals - total density per 100,000 population: hospitals,
gni_capita - gross national income per capita,
gghe_d - domestic general gov't health expenditure as %,
che_gdp - current health expenditure as %

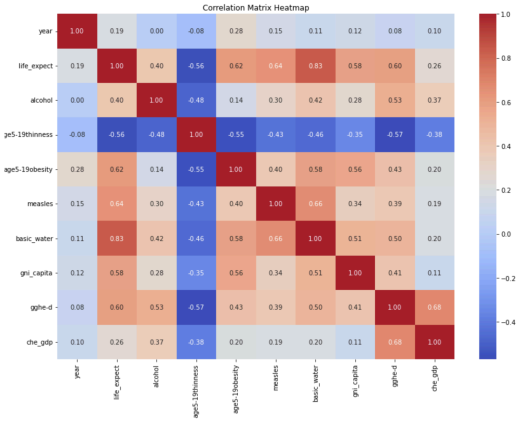
UNESCO predictors:
une_pop - Population (thousands), *une_infant* - infant mortality rate, *une_life* - Life expectancy at birth, total (years), *une_hiv* -Prevalence of HIV %, *une_gni* - GNI per capita, *une_poverty* - Poverty headcount ratio at $1.90 a day, *une_edu_spend* - Government expenditure on education, *une_literacy* - Adult literacy rate, *une_school* - Mean years of schooling

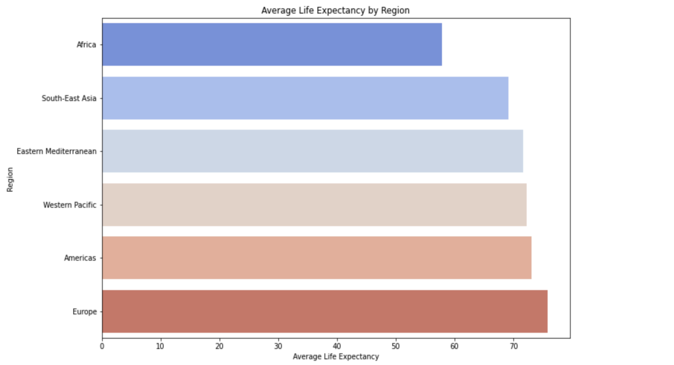
Data Cleaning and Preprocessing
As we began to tackle this task, we noticed that several variables had an abundance of missing data points scattered throughout the different predictors. The reason behind these gaps in data can be attributed to problems in the data collection process. With this in mind, we decided to set a precedent by removing any countries with eight or more missing data points in a single column. "Albania" "Argentina" "Democratic People's Republic of Korea" "Djibouti" "Libya" "Montenegro" "Myanmar" "Somalia" "South Sudan" "Sudan" "Syrian Arab Republic" "Zimbabwe" were removed for this reason. We continued our data-cleaning processes using a technique called linear interpolation to fill the NAs**.** Upon reevaluation, it was recognized that variables such as life_exp60, une_life, une_infant, infant mortality, adult mortality, and mortality rates for ages 1-4 more closely resembled dependent variables rather than independent variables. This stemmed from the observation that these mortality rates are intrinsically linked with life expectancy, creating an artificial correlation that doesn't necessarily reflect causation. Consequently, models incorporating these mortality rates skewed towards reliance on these variables, potentially obscuring the influence of other independent variables and oversimplifying the complexity of life expectancy determinants.

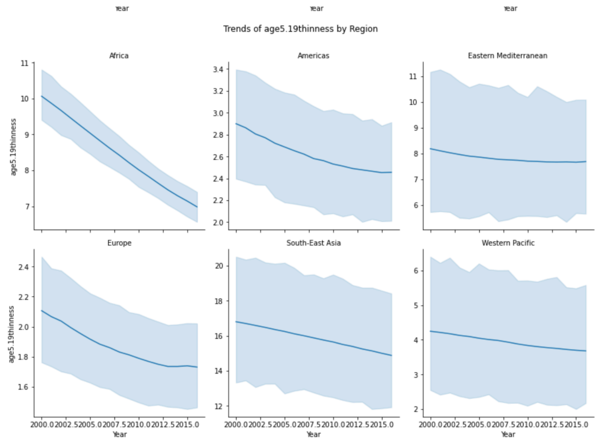


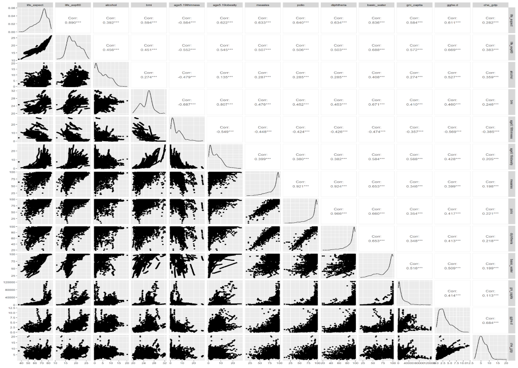
The Variance Inflation Factor (VIF) Values graph determines the multicollinearity of each variable to see if they need to be removed or not. If the VIF is above 5, the variable needs to be removed since the variance is relatively too large. In this case, we removed variables ‘measles’, ‘polio’, and ‘diphtheria’.

Exploratory Data Analysis (EDA)









Heatmap: The heatmap shows the correlation between each variable. The more red the squares are, the more correlated it is. The more blue the squares are, the less correlated it is. This helps us see which relationship we should be able to analyze based

Histogram: This is a histogram of life expectancy in each generalized region. The graph lets us see if the environment plays a significant role in life expectancy. Based on the data on the graph, countries in Africa have a lower average life expectancy than all the other regions. However, we cannot assume that this is due to regional factors.

Faceted Line Plot: These graphs represent malnutrition over the years in each. When comparing the six regions, you can see how the regions of Africa, Southeast Asia, and the Eastern Mediterranean all start with a relatively high percentage of children aged 5-19 who are considered to be thin. This correlates with the life expectancy of these 3 regions and age 5-19 thinness directly correlates to a lower life expectancy.

Pair Plots: With the pair plots, there are different types of representation of the data, the squares with the ‘corr’, state the correlation between the two variables that it is comparing. Then line distribution graphs show the distribution of the two variables, and the scatter plots show the overall relationship between the two variables. All in all, the plots show the relationship between all the variables in multiple visual representations.

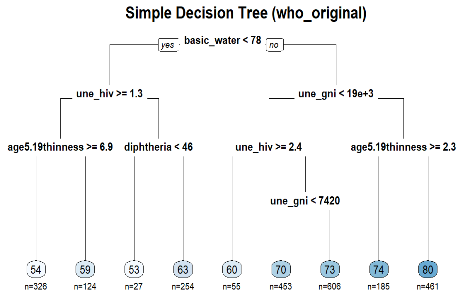
Methodology

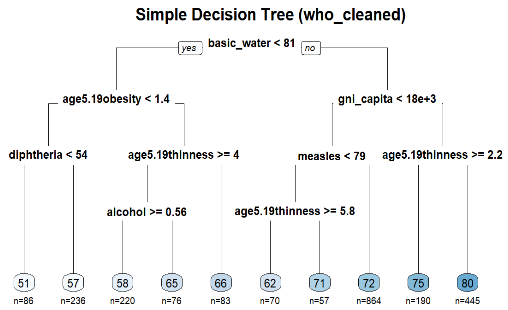
After successfully cleaning and preprocessing our data, we decided to start with a multiple linear regression model, which is the simplest and easiest to interpret. We went ahead and transformed the data using a box-cox transformation, normal transformation, and log transformation. Based on the diagnostic plots, we determined that both the box-cox transformation and log transformation of the cleaned dataset satisfied the basic assumptions of linear regression. This was our model: life_expect ~ alcohol + bmi + age5.19thinness + age5.19obesity + basic_water + gni_capita + gghe.d + che_gdp. After employing a multiple linear regression model, we explored Single Decision Trees on both the original dataset, with its inherent missing values, and a cleaned version devoid of N/As. This comparison aimed to assess the impact of missing data on prediction outcomes and the importance of variables. It leveraged the decision tree's capability to handle datasets with varying levels of completeness and aimed to illustrate the commonalities between them regarding the variables. After our Single Decision Trees, we worked on tuning them as a way to remove statistical noise. Additionally, we created a Random Forest and a gradient-boosting model using the cleaned dataset to examine which variables are more important and also help with getting the most precise predictions for our response variable

Main Results:

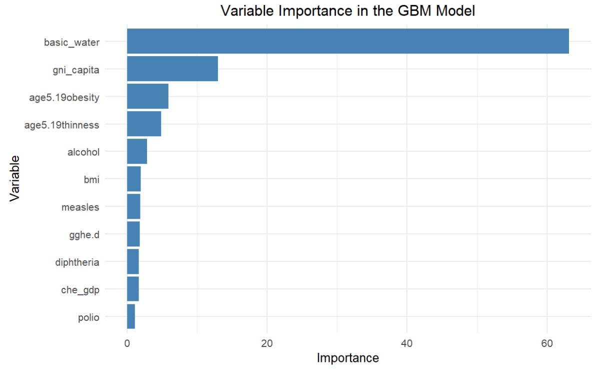
Model	R^2	RMSE	MAPE
Multiple linear regression w/ box-cox	0.500	24.47	31.71
Multiple linear regression w/normalized dataset	0.873	N/A	N/A
Multiple linear regression w/ log transformation	0.996	N/A	N/A
Simple Decision Tree (who_original)	0.821	3.733	4.314
Simple Decision Tree (who_cleaned)	0.812	3.879	4.456
Tuned Decision Tree (who_cleaned)	0.903	2.771	3.060
Random Forest (who_cleaned)	0.959	1.858	1.645
Gradient Boosting (who_cleaned)	0.932	2.323	2.613

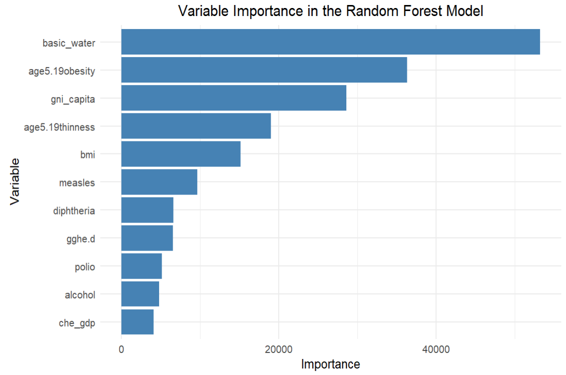
The Random Forest and Gradient Boosting models demonstrated superior predictive accuracy among our approaches, though they sacrificed some interpretability in exchange for this performance. Conversely, the tuned decision tree, enhanced through cross-validation, struck an optimal balance by providing accurate predictions alongside detailed insights into how each variable influences life expectancy. This model outperformed the simpler single decision tree not only in predictive accuracy but also in the depth of its analysis, offering 48 unique predictions compared to the latter's 9, thereby providing a more nuanced understanding of the determinants of life expectancy.


Simple Decision Tree (who_original)


Simple Decision Tree (who_cleaned)

Looking at the Decision tree (click to be rerouted to our tuned decision tree) and its simple counterparts, it's very clear how important drinking water is in determining whether the life expectancy will be higher or lower. The decision tree covers up any of the non-linear relationships and interactions that the linear regression may have overlooked. For our results, all the paths that originated from greater access to basic_water lead to a higher life expectancy. Following the split on basic_water, other factors like une_hiv, age5.19thinness, diphtheria, and gni_capita serve as secondary criteria. This implies these variables also contribute to life expectancy but are considered after accounting for water access. From the who_cleaned decision tree, we notice that age5.19obesity yields a longer life expectancy for countries without much access to water. This can be explained by noting that countries with a higher obesity rate in children are those that are more economically sound, like countries in Europe and North America.


Variable Importance in the GBM Model


Variable Importance in the Random Forest Model

These bar graphs show that the variable ‘basic_water’ has the highest impact on life expectancy, from both the Gradient Boosting Model and the Random Forest model. We also see a similarity in how the variable of current health expenditure seems to be the least important in both of the models. Our random forest model clearly shows that more variables have more importance, while the GBM model clearly shows a bias toward basic water being the most important variable. GNI_capita, age5.19obesity, and thinness are also important variables according to both models. Many of the immunization variables, such as polio, do not seem to have a strong role in determining life expectancy.

Discussion and Outlook

In this project, the initial challenge was deciding which variables to retain and which to remove. Unlike many datasets, some of the independent variables were highly correlated with the dependent variable, life expectancy. We looked at the VIF specifically to remove any chance of running into a multicollinearity problem. Moreover, Our models were built and validated on a specific dataset. The generalizability of our findings to other populations or contexts remains untested, which could limit the applicability of our main results in further more comprehensive studies.

Throughout our project, our team recognized various strengths, weaknesses, and opportunities for improvement in future collaborations. The main strength of our exploration lies in the diversity of our modeling approach. We could process our data through multiple models and select the one that suited our dataset. A weakness we encountered early and often was the dataset itself. At the project's outset, while selecting a dataset that aligned with our interests, we did not anticipate the extensive data cleaning required. We encountered numerous predictors with N/A values scattered throughout, making them unusable for our analysis. Changes we could make in future explorations include incorporating more data via additional datasets to have a more robust set of data to run through our models. These datasets could focus on a more broad outlook on the determinants of life expectancy, such as lifestyle and social determinants. In addition, while Gradient Boosting and Random Forests showed high predictive accuracy, their complexity reduced model interpretability. This trade-off limited our ability to derive actionable insights from these models as clearly as from simpler models like linear regression or single decision trees. For future projects, we could have a more reliable technique to deal with our missing data points. For example, researching techniques such as multiple imputation could help us to have more accurate data to run our analysis on.

Overall, the random forest model proved to be the best for prediction while the multiple linear regression model was the least accurate. The decision tree with cross-validation proved to be the most important model as it took advantage of both interpretability and accuracy.

Conclusion

Upon completing our analysis, we discovered that access to basic water services is the most significant predictor of life expectancy. We compared two models: one using the original dataset and another using our cleaned dataset. Both models performed similarly well, as indicated by their R-squared and RMSE values, suggesting they make accurate predictions. Through this comparison, it became evident that access to basic water services is the most critical factor. Even after removing the UNE columns, certain predictors—GNI per capita, age 5-19 thinness, and diphtheria immunization rates—consistently played a significant role across both models.