IS517 Final Project Report
World Health

By Abhinav Choudhry and Kerstin Wolf

May 9, 2021

1 Introduction

A lot has changed in a little over a year due to the COVID-19 pandemic that has swept the globe. The pandemic has challenged us all, and we have lost so many people to this virus. Due to COVID-19, public health has been brought to the forefront of everyone's minds unlike ever before on a global scale. This pandemic has also highlighted the differences in public health between countries as well. Life expectancy is not the same in every country. Citizens in one country may have a significantly higher life expectancy than citizens in another country. Keeping on the topic of public health, we decided to focus this project on life expectancy around the world and the various health factors that play into it.

2 Research Questions

For this project, we focused on three research questions applied in the context of countries of the world:

1. How well do healthcare factors do in comparison with economic factors in predicting life expectancies (both healthy life expectancy and life expectancy at birth)?
2. Are we able to successfully predict life expectancy at birth and healthy life expectancy at birth by a) health infrastructure and b) health behaviors?
3. Is there a significant difference between factors that predict healthy life expectancy versus life expectancy, and if so, what factors cause this?
4. Which statistical methods deliver the best performance on this dataset?

Healthy life expectancy (HALE) can potentially differ widely from life expectancy. While someone may live to be 80 years old, they may not be living a healthy life at that point. Citizens of one country may be expected to live to 80 years old, but their healthy life expectancy

can be much shorter. WHO defined healthy life expectancy as the "average number of years that a person can expect to live in 'full health' by taking into account years lived in less than full health due to disease and/or injury." The average age when most citizens began to decline in health and face numerous health problems and injuries is where healthy life expectancy is determined for that country.

3 Datasets

A large number of datasets were used in this project. Our primary source of data was from the Kaggle World Health Statistics 2020 dataset. This dataset was made up of 39 CSV files with one numeric attribute in each file and no null values. Each file focused on a different topic relating to health such as alcohol consumption, number of medical doctors, air pollution death rate, and so on. The number of rows per file was variable and was dependent on the number of countries multiplied by the number of years of available data multiplied by the number of sexes of the population that the data was collected for (including both sexes). The years of data collected had a fairly wide range, but most of the years fell within 2000 to 2019. As for sex, for some of the CSV files there were three different categories dividing the data: female, male, and both sexes.

The secondary datasets used in conjunction with the main dataset include a dataset on the prevalence of obesity among adults from the World Health Organization, plus GDP Ranking and Health Expenditure data from the World Bank database. Each of these datasets were used to make an attribute in the final combined dataset. The obesity dataset focused specifically on adults that are eighteen years old or older with a body mass index (BMI) of thirty or more. This dataset had a unique structure compared to the others in that all the columns were different years and the first couple of rows weren't observations so much as misplaced columns labeling what is being measured, the population being measured, and the sex of the population being measured. This dataset contained 196 unique observation rows. The GDP dataset was much more straightforward with only four columns: country abbreviation, ranking, country name, and GDP. This dataset contained 205 rows. The expenditure dataset is structured like a combination of the two. A few columns are the country name, code, and indicator, and the rest are years, but all rows are observations for individual countries. In this dataset, the values are the health expenditure as a percentage of the GDP for each country. This dataset contained about 264 unique row observations because it also included groups of various countries.

4 Preprocessing

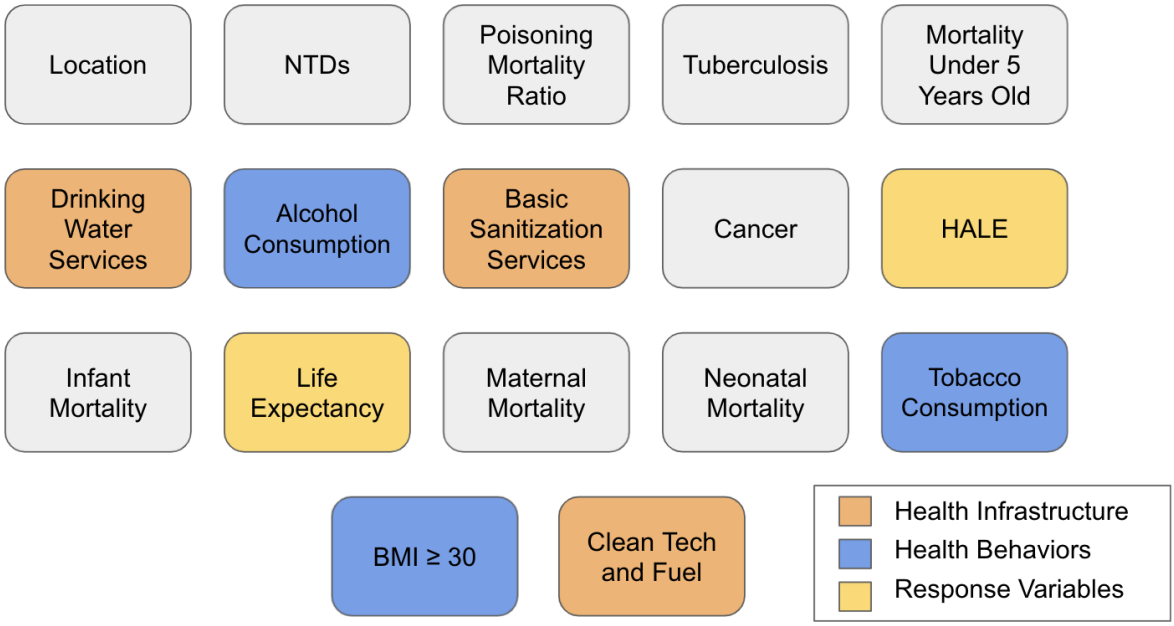
Data preprocessing played a massive role in this study and how well the models ran. While the four datasets we started with were fairly clean, new issues arose with the merging of the data files. Prior to merging, though, a significant amount of time was spent selecting the data and attributes that would be used from the Kaggle World Health Statistics 2020 dataset. With 39 separate CSV files making up the entire dataset, each file had to be looked at and thoroughly evaluated. Of the 39 attributes, only 22 made it through a thorough inspection as attributes that had value to the project, had believably accurate data, contained data for the desired years, and had a sufficient number of rows for those years. While none of the potential attributes contained null values, there were a number that contained questionable values of zero. We determined that in many instances some zeros represented that the data for that country wasn't available or that the country decided to falsely report certain details. For example, we eliminated an attribute on crude suicide rate due to so many countries not reporting or falsely reporting their numbers. Other attributes were also eliminated in this stage if they didn't contain enough rows for the desired years. We knew early on that there would be a number of difficulties with having a base dataset that has so many attributes but so few observations, and we were cautious to try to maintain as many usable rows as we could. This is why a number of attributes were immediately eliminated if each individual year of that attribute contained less than 100 observations.

After the initial evaluation, the 22 attributes from the base dataset that met the requirements were then written into an Excel document where the years for each attribute were then ranked by the number of observations. For example, not all attributes had the same number of rows per year. It was very common to find for one attribute that there were 200 rows for one year, but then only 105 rows for another year. This is why the ranking was important. This document allowed us to see which years appeared in the attributes the most and relatively how many rows there would be. Using this document, we selected data from 2015 to make up our training set and the 2016 data to make up the test set. We also further eliminated a couple more attributes to get to 19 attributes in total.

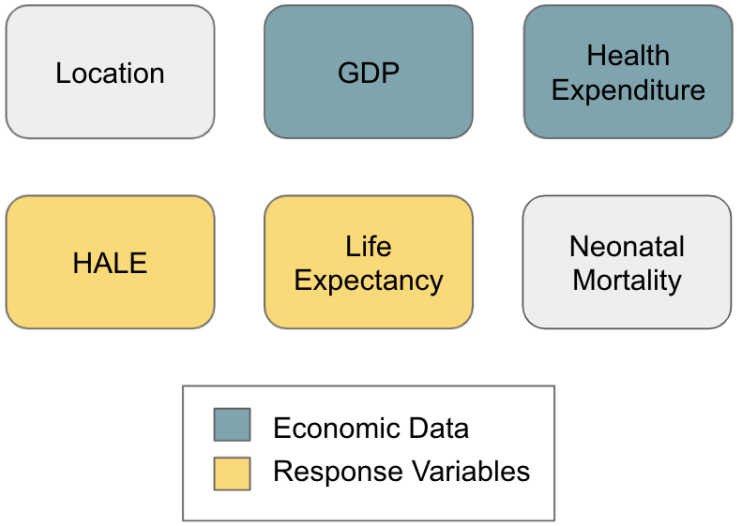
The process of selecting out the desired data from each file and merging them all into a single training set and test set was another challenge that arose. With the 19 attributes we had selected, we found that not all contained the same countries, and this was leading to a training

and test set riddled with null values. Further evaluation of which attributes were contributing to the most null values caused us to eliminate a couple more attributes. This allowed our training and test sets to reach an acceptable row count even after omitting all rows containing nulls. Some basic data cleaning was also conducted at this time to remove any standard deviation ranges after particular attribute values so that the data could be properly read into R.

With the addition of the obesity dataset from WHO, we reached our final base training and test set. Both sets contain 17 attributes relating to health behaviors and health infrastructure with 194 rows. When omitting the null values, the training set contained 124 rows and the test set contained 123 rows.



The remaining two datasets, World bank GDP Ranking and Health Expenditure, were utilized in creating a separate secondary training and test set in order to answer our first research question. A single attribute was pulled from each dataset and then merged with the location, healthy life expectancy, life expectancy, and neonatal mortality attributes from our base training and test sets. Some minor data cleaning was involved to successfully merge these attributes since the country names were written differently in each dataset. For this secondary training and test set, there were six attributes with 194 rows for each set. When omitting the null values, this training set contained 174 rows and the test set contained 173 rows.



5 Modeling Results

For the modelling, we concentrated mainly on healthcare factors and economic factors in line with our research questions. Though preliminary analysis revealed that various types of mortalities had high correlation with life expectancies, we considered them to be themselves dependent on primary infrastructural, economic, and behavioral factors. Therefore, we ran our models only on these. In the lasso model, however, we also ran it on all parameters. *It is likely that some values might differ slightly between runs based on the seeds used but the models are stable between runs and differences in error rates are not large.*

At the preliminary stage, we had also included a binary variable that classified countries on the basis of whether their HALE was higher and lower than the medial. However, as there is no such WHO classification, we decided to focus only on numerical prediction.

Linear Regression		
Data	**Training Error**	**Test Error**
All Economic Data for HALE	8.17%	7.84%
GDP for HALE	8.62%	8.27%
Health Expenditure for HALE	8.16%	7.83%
All Economic Data for Life Expectancy	8.11%	7.82%
GDP for Life Expectancy	8.61%	8.31%

Health Expenditure for Life Expectancy	8.10%	7.82%
Health Infrastructure for HALE	4.42%	4.37%
Basic Drinking Water Services for HALE	5.10%	4.95%
Basic Sanitization Services for HALE	4.77%	4.82%
Population Reliant on Clean Fuel and Tech for HALE	5.38%	5.21%
Health Infrastructure for Life Expectancy	4.47%	4.45%
Basic Drinking Water Services for Life Expectancy	5.17%	4.99%
Basic Sanitization Services for Life Expectancy	4.97%	5.01%
Population Reliant on Clean Fuel and Tech for Life Expectancy	5.35%	5.26%
Health Behavior for HALE	6.94%	6.54%
Alcohol Per Capita Consumption for HALE	8.07%	7.74%
Prevalence of Tobacco Smoking Among Population 15 Years Old or Older for HALE	8.56%	8.25%
Percent of Population with BMI Greater Than or Equal to 30 for HALE	7.99%	7.58%
Health Behavior for Life Expectancy	7.09%	6.73%
Alcohol Per Capita Consumption for Life Expectancy	8.12%	7.80%
Prevalence of Tobacco Smoking Among Population 15 Years Old or Older for Life Expectancy	8.64%	8.31%
Percent of Population with BMI Greater Than or Equal to 30 for Life Expectancy	7.98%	7.57%
Health Behavior and Health Infrastructure for HALE	4.09%	4.11%
Health Behavior and Health Infrastructure for Life Expectancy	4.20%	4.23%

Random Forests

Data	Training Error	Test Error
Economic Factors for HALE	3.56%	5.11%
Economic Factors for Life Expectancy	3.50%	5.08%
Health Infrastructure for HALE	2.04%	2.87%
Health Infrastructure for Life Expectancy	2.07%	2.93%
Health Behavior for HALE	3.02%	3.99%
Health Behavior for Life Expectancy	3.10%	11.76%

Bagging

Data	Training Error	Test Error

Health Infrastructure for HALE	1.10%	2.87%
Health Infrastructure for Life Expectancy	2.00%	2.94%
Health Behavior for HALE	2.89%	11.76%
Health Behavior for Life Expectancy	2.98%	4.10%
Health Behavior and Health Infrastructure for HALE	1.75%	11.76%
Health Behavior and Health Infrastructure for Life Expectancy	1.81%	4.10%

Boosting		
Data	**Training Error**	**Test Error**
Economic Data for HALE	0.14%	6.68%
Economic Data for Life Expectancy	0.10%	6.62%
Health Infrastructure for HALE	0.28%	2.78%
Health Infrastructure for Life Expectancy	0.25%	2.88%
Health Behavior and Health Infrastructure for HALE	0.00%	2.17%
Health Behavior and Health Infrastructure for Life Expectancy	0.00%	2.25%
Health Behavior for HALE	0.02%	4.79%
Health Behavior for Life Expectancy	0.02%	5.00%

Support Vector Machines (Linear Kernel)		
Data	**Training Data**	**Test Error**
Health Infrastructure for HALE	4.38%	4.19%
Health Infrastructure for Life Expectancy	4.44%	4.33%
Health Behavior for HALE	6.97%	6.38%
Health Behavior for Life Expectancy	7.15%	6.65%
Health Behavior and Health infrastructure for HALE	4.01%	3.89%
Health Behavior and Health infrastructure for Life Expectancy	4.11%	4.01%
Economic Factors for HALE	7.58%	7.03%
Economic Factors for Life Expectancy	7.67%	7.17%

The tables above detail the results for various models that we ran. The ones in orange were the only ones with test errors above 10%. Some test errors are lower than training errors but as we are dealing with time series data varying across years, this is understandable. We now interpret them in context of our research questions.

Lasso Results

We ran lasso on all attributes and again on healthcare parameters. The lasso on all parameters finds that mortality factors have a negative effect on both HALE and life expectancy and drinking water and sanitization have positive effects. High BMI has a negative effect.
The lasso on healthcare alone finds high BMI to have the largest negative effect while all infrastructure parameters: drinking water, sanitization, and access to clean fuel and technology have positive effects. Drinking water has the highest coefficient but its absolute value is lower than that of high BMI. This shows how important obesity and overweight characteristics are to life expectancy.

We noticed some anomalous results in the lasso w.r.t substance abuse as alcohol and tobacco did not show negative effect on HALE and life expectancy when we ran lasso on healthcare factors and all parameters respectively.

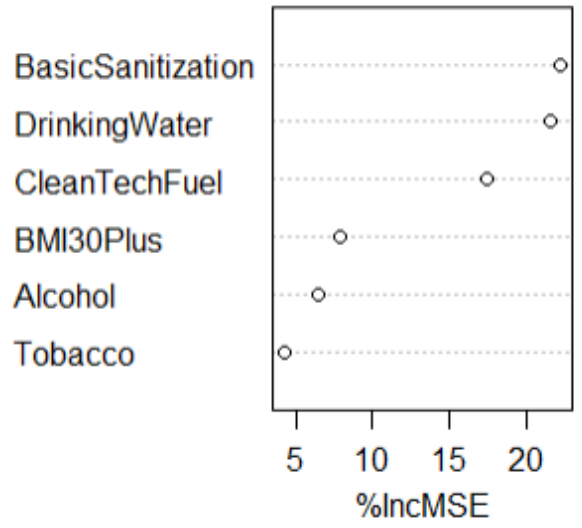
RQ1: Model performance of economic factors in comparison with healthcare factors

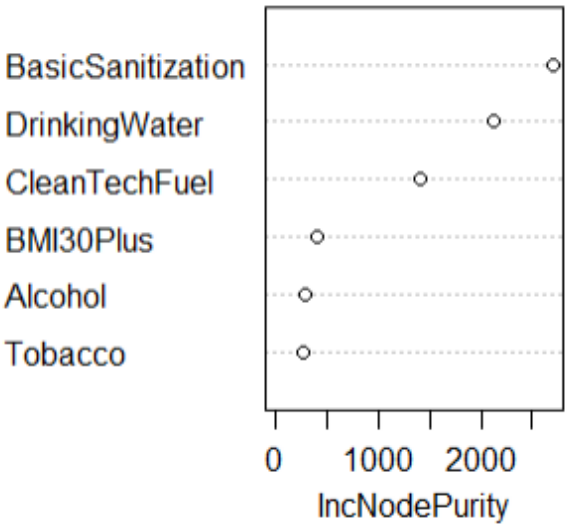
We found that economic factors produce test errors of approximately 8% in linear regression, 5% under random forests, 6.5% in boosting and 7% in Support Vector machines. On the other hand, both health infrastructure and health behavior has much lower test errors in all the models with the errors even being half those of economic models in random forests, bagging, and boosting models. Thus, we can conclude that healthcare factors have superior predictive power for life expectancies.

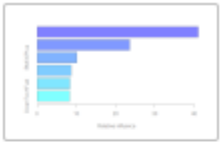
GDP and health expenditure data show some difference in results. Under linear regression, health expenditure has more prediction accuracy and explanatory power under R-square, but under boosting and random forests, GDP appears to have a larger effect. So, we cannot conclusively comment on the relative importance of each.

RQ2: Importance of Health Infrastructure and Health Behaviors as factors

We grouped health factors under health infrastructure and health behavior running them separately as well as together. While both health infrastructure and health behavior prove to be important, the models show health infrastructure to be more important. This is shown by the results of the bagging and boosting models below.







	var <fctr>	rel.inf <dbl>
BasicSanitization	BasicSanitization	34.495051
DrinkingWater	DrinkingWater	23.324179
CleanTechFuel	CleanTechFuel	17.831482
BMI30Plus	BMI30Plus	9.547585
Tobacco	Tobacco	7.449313
Alcohol	Alcohol	7.352391

Among health behaviors, the prevalence of overweight and obesity among the population has the single largest negative impact on life expectancy and the effect is much higher than for alcohol and tobacco use.

RQ3: Healthy Average Life Expectancy and Life Expectancy at Birth

Most of the models showed no major differences in the results for these two dependent variables. However, in random forests and bagging models, we noticed some difference between the predictive performance with life expectancy being poorly predicted by health behaviors in random forests and HALE being poorly predicted by all healthcare factors as well as health behavior alone. As the prediction varied on both HALE and life expectancy and only applied to these models, we cannot conclude that this is a true difference. There were minor differences

w.r.t the impact of individual factors but these were not the primary factors and so we cannot conclude that there is any major difference. We cannot comment whether the fact that no difference exists is due to the factors having the same effect on both HALE and life expectancy or due to calculation methodologies of the source data.

RQ4: Model Performance

We found that all models produced fairly low test errors, even for the economic factors. However, boosting delivered the best results with very low test errors for all factors. Random forests performed well on some models but did poorly on others. Linear regression was effective, but its test errors were not the lowest. Boosting was also very helpful for interpretation giving a clear relative significance of factors as well, though, even lasso and linear regression did the same. The results of lasso suggest that we should also examine individual disease factors affecting mortality. As this was not in our original research questions, it was outside the scope of this project.

We would suggest a perusal of the pdf file for all results as we have only outlined the top highlights.

6 Conclusion and Recommendations

Our analysis shows that both health infrastructure and health behavior are critical to life expectancy. Provision of drinking water, sanitation, and clean fuel technologies are able to account for differences in life expectancies far more accurately than economic factors. This leads us to conclude that economic growth alone cannot ensure improved health outcomes: health infrastructure has to grow for improved outcomes. Adverse health behaviors too have a demonstrable negative effect, but the most important among these is the prevalence of overweight and obesity. This factor was shown to be impactful among all the models we ran. Thus, we can recommend that it is extremely important to tackle the issue of obesity and overweight among the population. This might need intervention in terms of both tackling overnutrition and curbing sedentary behavior.

7 References

Healthy life expectancy (HALE) at birth. (n.d.). World Health Organization. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/66

World Bank Open data (n.d.) www.data.worldbank.org. retrieved May 2021

WHO. World health statistics 2020: monitoring health for the SDGs, sustainable development goals. Geneva: World Health Organization; 2020.

(*Prevalence of obesity among adults, BMI ≥ 30, age-standardized - Estimates by WHO region*, n.d.)