Life Expectancy Predictor (LEP)

Abstract

Life expectancy is one of the best indicators of a country's success. A high life expectancy usually means the country and its citizens are living well. Through feature selection, we try to find if mortality rates are a good predictor of life expectancy. From apriori inspection, we can assume mortality rates must have an impact on life expectancy. This report and program we created is built using three mortality rate attributes, but can fundamentally use any features for prediction. The prediction program could be able to determine future life expectancy of a country based on current trends in the data.

Table of Contents

1 Introduction

The life expectancy of a country is a great indicator of the country's socio-economic success. Countries with high life expectancy have more citizens living a long healthy life, less mortalities in the early years of a child's life, less prone to diseases, and employs cleaner water and hygiene standards.

There are many factors that contribute to a high life expectancy, including:

- National economic circumstances
- National mental health levels
- Education levels
- Variations in regions
- Access to medicine and hospitals

When choosing which metrics to consider for predicting the life expectancy of a country, we determined that three unique sets of data should be used: infant mortality rate[1], malaria mortality rate[2], and WASH mortality rate[3]. All our data in this project comes from the World Health Organization (WHO)[4].

1.1 Infant Mortality Rate

The infant mortality rate of a country is a very good indicator to the average life expectancy of a country - in fact, the most commonly used measure of determining life expectancy is the Life Expectancy at Birth (LEB) statistic[5]. This measures the life expectancy of someone who is currently at an infant age, as they are the most susceptible to fatal illnesses: in 2017, 4.1 million (75% of all under-five deaths) occurred within the first year of life[1].

For our data analysis, we used the World Health Organization (WHO) data.

Overall, the infant mortality rate has decreased worldwide through the use of better technology, a more efficient healthcare infrastructure, and more education throughout impoverished communities. There is still work to be done, as the infant mortality rate for the African Region (51 per 1000 live births), is over six times higher than that in the European Region (8 per 1000 live births)[1].

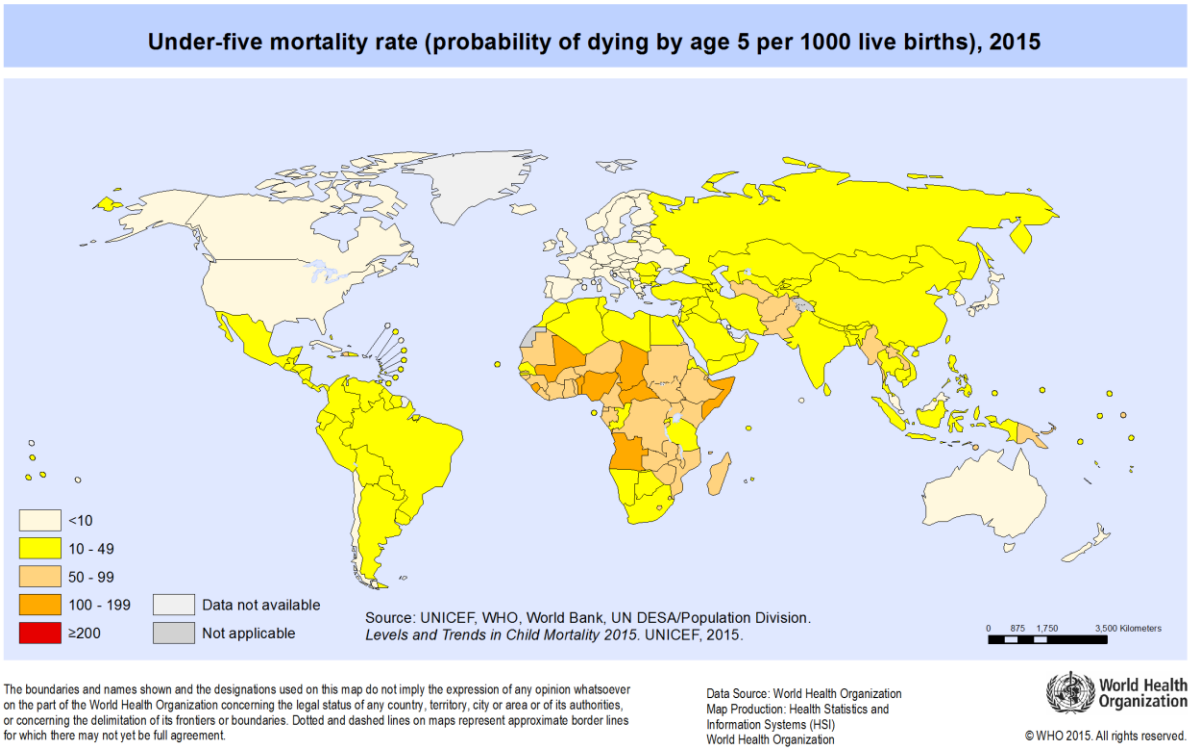


Figure 1.1 Worldwide under-5 mortality rate (probability of dying by age 5) per 1000 live births, 2015[6]

1.2 Malaria Mortality Rate

Malaria is one of the world's deadliest and most prolific diseases. It affects over 90 countries around the world - and with 216 million cases of malaria in 2016 the disease is a global problem[7]. Malaria is spread through mosquitoes, when one mosquito bites an infected individual, the insect acts as a host to the disease and can possibly infect all other organisms it bites.

Malaria is most common throughout tropical areas and regions around the equator, and even though there is a vaccine available, most are not able to afford it as malaria is commonly associated with poverty. We chose malaria as a parameter for predicting the life expectancy of a country since it is a factor that affects many people around the world.

1.3 WASH Mortality Rate

The WASH mortality rate is the mortality rate attributed to unsafe water, unsafe sanitation and lack of hygiene: exposure to unsafe Water, Sanitation and Hygiene for All (WASH)[3]. Hundreds

of millions of people around the world do not have access to clean drinking water - leaving them susceptible to diarrheal illnesses and cholera.

We chose to use the WASH mortality rate parameter in our prediction model as it is important to see how much something as simple as clean water contributes to a country's average life expectancy.

2 Data Collection

The data we collected is from the WHO. The datasets we use are infant mortality rate[1], malaria mortality rate[2], the mortality rate from unsafe water, sanitation, and hygiene[3]; and average life expectancy per country[8]. Additional datasets can be used in the future for a more robust and accurate prediction program.

Our initial design plan was to use the WHO APIs to collect all data, however the XML/JSON APIs were more difficult to collect data with than CSV file formatted data. This design choice limits the amount of datasets we collected from initially, but improves data processing and preprocessing time.

Future considerations include collecting more datasets, and collecting missing data from other sources if possible. Creating functions to process the API calls could also be considered to improve prediction accuracy.

3 Data Preprocessing

The data gathered was in CSV file format. We chose a CSV file format because the Python CSV library is very straightforward to use. The CSV files were cleaned to remove explanatory text for easier and cleaner processing - the data itself was cleaned to remove extra whitespace and other unnecessary punctuation.

The data was inserted into individual database tables, with the associated country's ID and year ID as foriegn keys. This allowed the joining of multiple datasets, and implementation of filters by year or country within the SQL query. Pruning the dataset becomes very simple.

Countries with no data for a dataset was particularly troublesome. In our case, countries with very high life expectancy did not have complete feature data, and thus were omitted from the training dataset. A byproduct of this is the max life expectancy prediction was lower than the real

max life expectancy - future considerations include generating missing values to be able to incorporate all countries in the training algorithm.

Before training the data mining algorithm, each feature set was scaled from 0 to 1: “0” being the minimum value, and “1” being the maximum value of the feature set.

4 Data Mining

Our program uses linear regression to predict a continuous value with multiple features. Benefits of this algorithm allows for any single or number of features to be used. Quick training and prediction is also a plus. Our final implementation used a learning rate of 0.001 and 100,000 epochs. Using around 100 training sets took two to three seconds to train.

The end project allows for a user to input data values from 0 to 1 for each feature, to see what the predicted life expectancy would be for an arbitrary country. Our current design trains the model each time a user submits their values. This design choice was made as it allows for dynamic feature selection in future implementations. In contrast, we considered saving the weights of the trained model for quicker predictions, however, it would not be scalable for dynamic feature selection.

Our future design plan allows for user feature selection, and to facilitate experimentation with different datasets to improve life expectancy prediction accuracy.

5 Data Analysis

To analyze our dataset with the prediction algorithm, we ran the linear regression algorithm 1200 times to calculate predicted life expectancy in different scenarios. In order to visualize the effect of a single variable on the prediction outcome, each data parameter (infant mortality rate, malaria mortality rate, and WASH mortality rate) was individually ran against the data set with the other two variables set as a constant, as seen in Table 5.1.

	Malaria Mortality Rate	WASH Mortality Rate
Variable Infant Mortality Rate For Each 0 to 100	0	0
	0	1
	1	0
	1	1

	Infant Mortality Rate	WASH Mortality Rate
	0	0
Variable Malaria Mortality Rate	0	1
For Each 0 to 100	1	0
	1	1

	Infant Mortality Rate	Malaria Mortality Rate
	0	0
Variable WASH Mortality Rate	0	1
For Each 0 to 100	1	0
	1	1

Table 5.1 Data analysis parameters used to predict individual parameter effect on life expectancy outcome

This way we can see how much of an effect each individual parameter has on the life expectancy outcome, by setting two parameters to a constant value of 0 or 1, we can run the remaining variable parameter with values between 0 and 1, incremented by 0.01 each time. See Figure 5.2 for the visualized outcome.

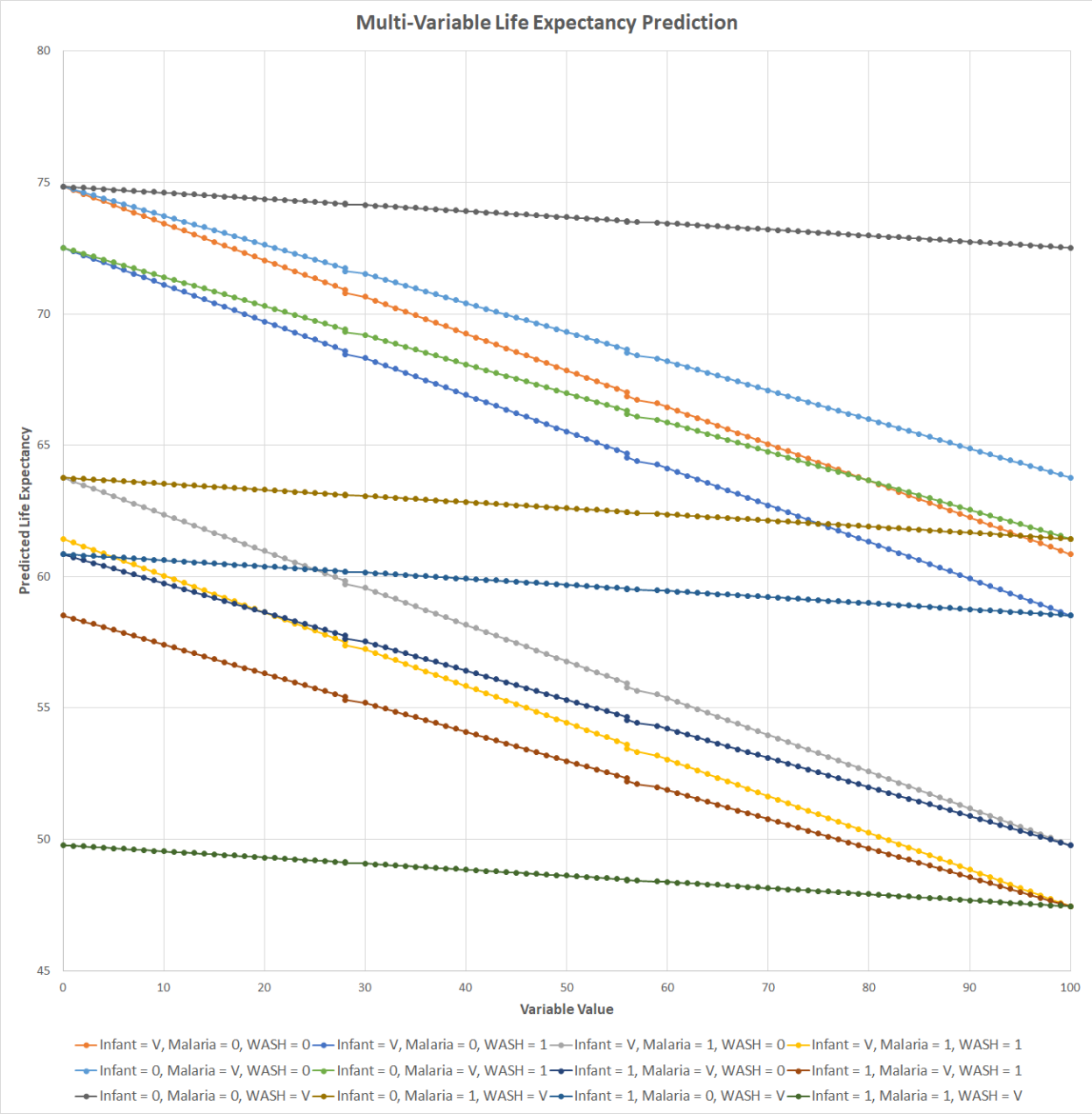


Figure 5.2 Life expectancy given by individual variable parameters and constant parameters

5.1 Analysis Results

Looking at the scatter plot given in Figure 5.2, we can clearly see that certain parameters have a greater effect on the life expectancy prediction than others. This is useful information when determining which national problem is contributing the most to a low life expectancy.

5.1.1 Infant Mortality Rate

When setting the infant mortality rate as a variable and constraining all other parameters, we can see that this variable has quite a significant effect on the life expectancy prediction outcome. This is expected since the life expectancy at birth (LEB) calculation is frequently used to find a countries average life expectancy.

5.1.2 Malaria Mortality Rate

Setting the malaria mortality rate as a variable deems a very wide range of life expectancy results. Malaria has the highest effect on the life expectancy of a country - therefore it is the most important parameter to consider when trying to improve a country's life expectancy. Even with both attributes infant mortality rate and WASH mortality rate set to 0 (very low rate), the outcome is a large range from ~75 years when the malaria mortality rate is low to ~63 year when the malaria mortality rate is high. Looking at Figure 5.2 we can see that the variable malaria mortality rate produces a steep-sloped linear line, indicating its effect on the life expectancy prediction outcome.

5.1.3 WASH Mortality Rate

Surprisingly the WASH mortality rate does not have as significant of an effect on a country's life expectancy as we had predicted. Looking at Figure 5.2 we can see that the variable WASH rate derives a fairly steady linear line, with a small slope.

6 User Interface

Along with the linear regression life expectancy prediction algorithm, we also designed an intuitive front-end web interface for our program. This allows the user to input three variables (infant mortality rate, malaria mortality rate, and WASH mortality rate) into the interface, send these values through the application API, and receive a predicted life expectancy.

6.1 Frontend Implementation

For the user interface we decided to use the Flask[9] framework. This allowed us to easily integrate the linear regression algorithm, data analysis, data querying, and interface into one simple application.
User input is sent as a JQuery Ajax POST request to the Flask API's respective prediction endpoint:

```
$.ajax({
    url: 'predict?' + $.param({
        x: document.querySelector('#infant_mortality_rate').value,
        y: document.querySelector('#malaria_mortality_rate').value,
        z: document.querySelector('#wash_mortality_rate').value,
        year: document.querySelector('#prediction_year').value}),
    type: 'POST',
        contentType: 'application/x-www-form-urlencoded; charset=UTF-8',return go(f, seed, [])
```

Which is then caught by the Flask application:

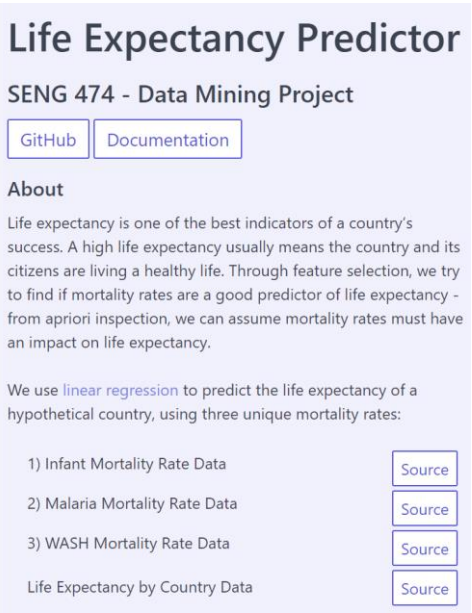
```
@app.route('/predict', methods=['POST'])
def predict():
    x = float(request.args.get('x'))
    y = float(request.args.get('y'))
    z = float(request.args.get('z'))
```

This allows the prediction algorithm to be extremely flexible - new endpoints and functionality can be added to improve prediction accuracy. Since an API is used to get the life expectancy prediction, the frontend can be implemented in various ways, such as a mobile app.

6.2 Using the Interface

The interface was designed to be easy to use. The interfaces is presented as a single-page: every element that a user needs to run the prediction algorithm is accessible from this page. Below is a breakdown of each element:

6.2.1 Sidebar



The application sidebar is a simple panel where important information about how the prediction results are calculated, and where data is sourced from. It provides links for the source code as well as the link to this documentation resource.

6.2.2 Parameter Input

The main program runs in this area - this is where users can input their own parameters to predict the life expectancy of a hypothetical country. Each input area is described below:

1) **Country Name** - any string that is used to name the country, a random one is chosen for the user each time the prediction algorithm is run. Useful when differentiating prediction data in the log.
2) **Infant Mortality Rate** - the mortality rate for infants, normalized from 0 to 100; where 0 is the lowest possible infant mortality rate, and 100 is the highest possible infant mortality rate as derived from the data set. Default 50.
3) **Malaria Mortality Rate** - the malaria mortality rate, normalized from 0 to 100; where 0 is the lowest possible malaria mortality rate, and 100 is the highest possible malaria mortality rate as derived from the data set. Default 50.
4) **WASH Mortality Rate** - the WASH mortality rate, normalized from 0 to 100; where 0 is the lowest possible WASH mortality rate, and 100 is the highest possible WASH mortality rate as derived from the data set. Default 50.

After completing all input parameters, the user can click the “Create Country” button to calculate the predicted life expectancy using their input data - see Figure 6.2.3.

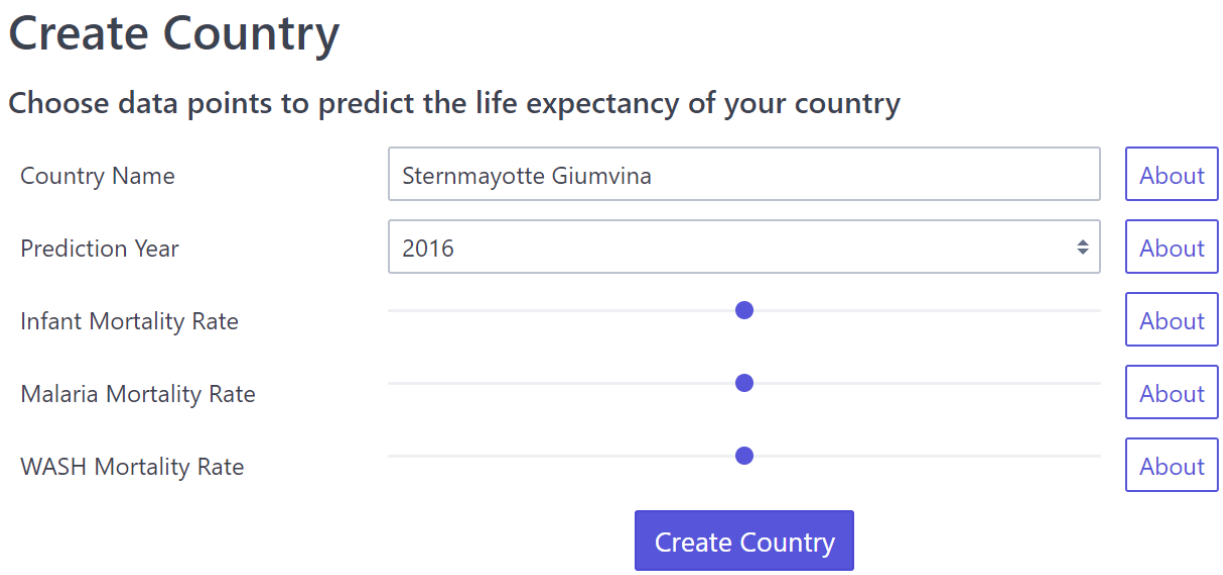


Figure 6.2.3 Input area for user data

6.2.3 Prediction Result

The linear regression prediction result is then returned to the user dynamically, using Ajax document updating. A meter is used to visually indicate how low or high the life expectancy is relative to the lowest or highest possible life expectancy.

6.2.4 Logging

Each prediction run in the current user session is saved in a table with the attributes: country name, life expectancy in years, and the relative percent deviation from average. This way users can compare past prediction results.

Country Name	Life Expectancy (Years)	Relative to Average (%)
Lyri	74.665	14.869%
Ancro Beminew	63.549	-2.232%
Gerchristcroa	48.836	-24.868%
Ancro Beminew	63.94	-1.631%
Canldives Grecoca	64.284	-1.102%
Iofnew	60.669	-6.663%
Sternmayotte Giumvina	61.509	-5.371%

6.3 Future Improvements

The current interface allows for basic usage of the prediction algorithm. In the future the web interface could allow for the addition of different data sets, usage of different prediction years, and dynamic graphing to compare past prediction results.

7 Conclusion

The system as a whole works as intended, however with drawbacks. The system can only use known data values, and not all countries have a full dataset for each feature. Feature selection and extraction takes time to develop the database table and scripts for insertion.

A model with every country will behave differently with a model trained on only a subset of countries. With an incomplete dataset to train the model, it will never be 100% accurate.

Our data analysis showed some interesting results - malaria mortality rate is the number one contributor to low life expectancy among our parameters; useful data to know when tackling a country’s life expectancy problem.

8 References

[1] http://apps.who.int/gho/data/node.main.525
[2] http://apps.who.int/gho/data/node.main.A1368
[3] https://apps.who.int/gho/data/view.main.SDGWSHBOD392v
[4] https://www.who.int/
[5] https://en.wikipedia.org/wiki/Life_expectancy
[6] http://gamapserver.who.int/mapLibrary/Files/Maps/Global_UnderFiveMortality_2015.png
[7] https://apps.who.int/iris/bitstream/handle/10665/259492/9789241565523-eng.pdf
[8] https://apps.who.int/gho/data/node.main.688?lang=en
[9] https://palletsprojects.com/p/flask/