Life Expectancy Predictor (LEP)

Matt Shipmaker & Lehel Keresztely

Department of Computer Science, University of Victoria

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 mortality rates, 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 a country's socio-economic success. Countries with high life expectancy have more citizens living a long healthy life, as well as less mortalities in the early years of a child's life.

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

- National economic circumstances
- National mental health levels
- Education levels
- Variations in regions

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

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. 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, more efficient health infrastructure, and more education throughout impoverished communities. There is still work to be done, as the infant mortality rate for WHO African Region (51 per 1000 live births), is over six times higher than that in the WHO European Region (8 per 1000 live births) [1]. http://apps.who.int/gho/data/node.main.525?lang=en

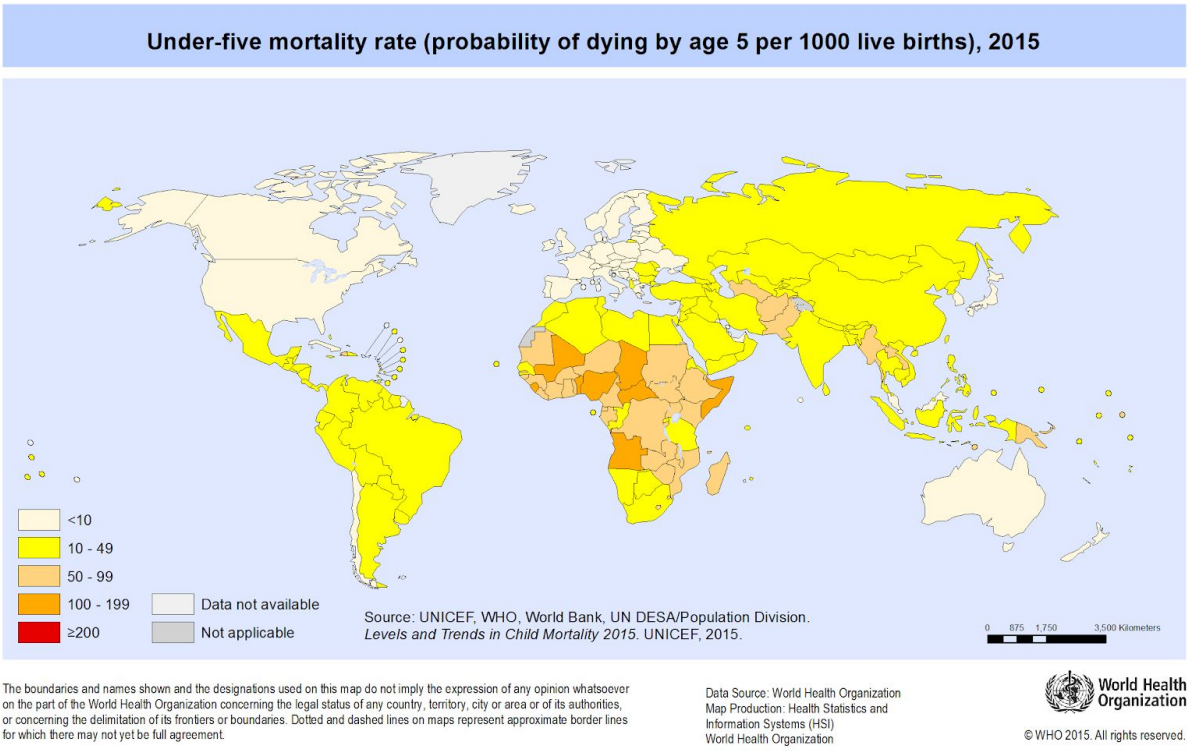

Figure 1.1 Worldwide under-5 mortality rate (probability of dying by age 5) per 1000 live births, 2015

1.2 Malaria Mortality Rate

http://apps.who.int/gho/data/node.main.A1368?lang=en

1.3 WASH Mortality Rate

2 Data Collection

The data we collected is from the WHO. The datasets we use are infant mortality rate, mortality rate from unsafe water, death rate from malaria, and life expectancy. More datasets can be used in the future, for a more robust 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 than csv file formatted data. This design choice limits the amount of datasets we collect from initially, but quickens the process of data preprocessing.

Future considerations include collecting more datasets, and collecting missing data from other sources if possible. Creating functions to process the API calls are also being considered.

3 Data Preprocessing

The data gathered was in csv file format. We chose csv file format because python csv library is very easy 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 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 to be filtered by year or country within the SQL query. We could remove countries with no data values very easily.

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 omitted from the training. 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 training.

Before training the data mining algorithm, each feature set was scaled from 0 to 1. 0 being the minimum value, and 1 being the max 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 allow for any feature or any 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 100000 epochs. Using around 100 training sets took about 2-3 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 because 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 plans allows for user feature selection, to facilitate experimentation with different datasets.

6 Data Analysis

To analyze our dataset with the prediction algorithm, we ran the linear regression algorithm 1200 times over our dataset. 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 seen in Table 6.1.

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

Infant Mortality Rate	WASH Mortality Rate
0	0

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

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

Table 6.1 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 parameter with values between 0 and 100. See Figure 6.1 for the visualized outcome.

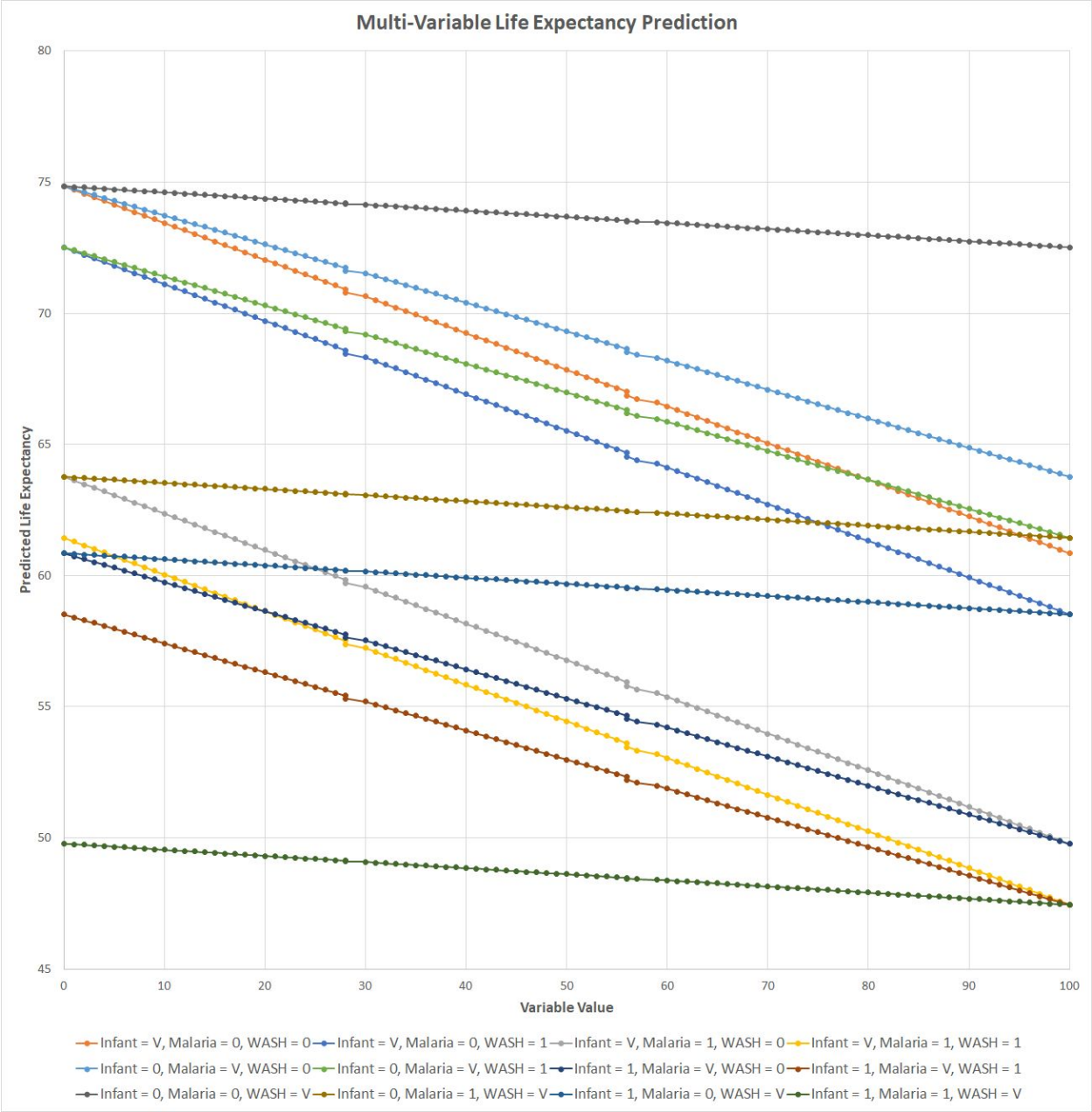

Figure 6.1 Life expectancy given by individual variable parameters and constant parameters

6.1 Analysis Results

Looking at the scatter plot given in Figure 6.1, 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.

6.1.1 Infant Mortality Rate

When setting the infant mortality rate as a variable and constraining all other parameters, we can see that

6.1.2 Malaria Mortality Rate

6.2.3 WASH Mortality Rate

5 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.

6 References

[who] https://www.who.int/
[1]
[infant mort rate] http://apps.who.int/gho/data/node.main.525?lang=en
[malaria mort rate] http://apps.who.int/gho/data/node.main.A1368?lang=en
[wash mort rate] https://apps.who.int/gho/data/view.main.SDGWSHBOD392v?lang=en\
[life expect rate] https://apps.who.int/gho/data/node.main.688?lang=en