# **Title: Life Expectancy Analytics**

|  |  |  |
| --- | --- | --- |
| First name | Last Name | IIT Email |
| Minguk | Kim | [mkim105@hawk.iit.edu](mailto:mkim105@hawk.iit.edu) |
| Boyun | Jang | [bjang7@hawk.iit.edu](mailto:bjang7@hawk.iit.edu) |

**Type of Your Projects (select one or more):** Multiple linear regression

1. **Introduction**

Life expectancy is one of the factors that measure human development index (HDI) in each country besides human education level and living standard. It is used to describe the quality of life in a particular area. The difference in life expectancy is also cited to demonstrate the need for medical care and the improvement of social support.

In order to prove the necessity of medical and social support for each country, we are going to analyze the factors of life expectancy such as the national mortality rate, economic factors, social factors and other health-related factors.

We will present what kind of improvements are needed to improve the life expectancy of the population by predicting the vulnerable diseases and environments that affect the mortality rate in a specific area in order to improve the mortality rate.

1. **Data Sets**

Data sets related to life expectancy and health factors in 193 countries were collected from the same WHO data store website. Economic data was collected on the UN website. Only representative elements were selected from all categories of health-related factors. The data set consists of 22 columns and 2938 rows, which means 20 prediction variables.

The attached dataset “Life Expectancy Data.csv” has been obtained from Kaggle.

|  |  |
| --- | --- |
| Variable | Description |
| Country | Country |
| Year | Year |
| Status | Developed or Developing status |
| Life expectancy | Life Expectancy in age |
| Adult Mortality | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) |
| Infant deaths | Number of Infant Deaths per 1000 population |
| Alcohol | Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol) |
| percentage expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%) |
| Hepatitis B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) |
| Measles | Measles - number of reported cases per 1000 population |
| BMI | Average Body Mass Index of entire population |
| under-five deaths | Number of under-five deaths per 1000 population |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%) |
| Total expenditure | General government expenditure on health as a percentage of total government expenditure (%) |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) |
| HIV/AIDS | Deaths per 1 000 live births HIV/AIDS (0-4 years) |
| GDP | Gross Domestic Product per capita (in USD) |
| Population | Population of the country |
| thinness 1-19 years | Prevalence of thinness among children and adolescents for Age 10 to 19 (% ) |
| thinness 5-9 years | Prevalence of thinness among children for Age 5 to 9(%) |
| Income composition of resources | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) |
| Schooling | Number of years of Schooling(years) |

1. **Research Problems**

The main goal of this project is to predict if the various data elements we collect will affect the average life expectancy.

The following objectives are based on these key objectives.

* The first goal is to predict what vulnerable diseases and environments will affect the mortality rate in a given area in order to improve anticipated life expectancy.
* The second goal is to analyze the data and discover patterns that can increase the life expectancy. For example, we can analyze if increasing medical expenditure or increasing the compulsory schooling period can effect positively the average life expectancy.

1. **Potential Solutions**

The main goal of this project is to predict factors that affect life expectancy and life expectancy.

First, we will make multiple linear regression models.

* **Y Variable** :

Life Expectancy – Life Expectancy in age

* **X Variables** :

1. Country - Country
2. Year - Year
3. Status – Developing or Developing status
4. Adult Mortality - Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
5. Infant deaths - Number of Infant Deaths per 1000 population
6. Alcohol - Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)
7. Percentage of expenditure - Expenditure on health as a percentage of Gross Domestic Product per capita(%)
8. Hepatitis B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
9. Measles - Measles - number of reported cases per 1000 population
10. BMI - Average Body Mass Index of entire population
11. Under-five deaths - Number of under-five deaths per 1000 population
12. Polio - Polio (Pol3) immunization coverage among 1-year-olds (%)
13. Total expenditure - General government expenditure on health as a percentage of total government expenditure (%)
14. Diphtheria - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
15. HIV/AIDS - Deaths per 1 000 live births HIV/AIDS (0-4 years)
16. GDP - Gross Domestic Product per capita (in USD)
17. Population - Population of the country
18. Thinness 1-19 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (% )
19. Thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%)
20. Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
21. Schooling - Number of years of Schooling(years)

Since data size is large enough, we will perform feature selection methods like backward elimination, forward elimination using p-value or AIC/BIC as metric, stepwise regression, subset regression through hold-out evaluation

Also, multicollinearity problem can occur because there are many variables. Therefore, we will use the VIF as a reference to identify correlations between x variables and to remove variables that are highly correlated with other variables.

We will also perform goodness of fit, individual parameter test, residual analysis and compare adjusted R2 and RMSE to see if the model is qualified.

In the hypothesis test,

**Claim Example 1: Northeast Asia has a generally higher life expectancy than Southeast Asia.**

H0: Northeast Asia and Southeast Asia have the same average life expectancy.

H1: Northeast Asia has a higher life expectancy than Southeast Asia.

**Claim Example 2: Developed countries have a generally higher life expectancy than developing countries.**

H0: Average life expectancy is the same for developed and developing countries.

H1: The average life expectancy of developed countries is higher than that of developing countries.

Like above, you can test your hypothesis first, and then you can continue to study why A has a higher life expectancy than B.

You can also use the ANOVA test if necessary to compare the two groups.

1. **Evaluations**

As we learned in class, the data size is 2939, so we split the data. As a way of doing this, our data size is more than 30, so hold-out evaluation should be used. For data size is large, data should be mixed, 80% of the population is used for the training set to build a model, and the remaining 20% is used to test data set to evaluate the model fit.

To confirm the qualification of the created model, f-test and residual analysis will be performed to check the accuracy. We will analyze the data as the most appropriate model for problem solving through confidence interval and multiple criteria such as p-value test, adjust R-squared test and RMSE verification.

1. **Expected Outcomes**
   1. A comparison of the Y and X values reveals factors that affect the expected life expectancy. Therefore, it extracts factors that are influential among social, economic and other health-related factors and suggests improvement of life expectancy by country.
   2. Comparison of life expectancy by country.
   3. The results of the improvements from the past can be seen through factor comparisons by year (2000-2015).
   4. How do infant and adult mortality affect life expectancy?
   5. Does life expectancy have a positive or negative correlation with eating habits, lifestyle, and drinking?
   6. Does life expectancy have a positive or negative relationship to drinking?
   7. What is the impact of vaccination coverage on life expectancy?