# Project 2: Proposal & Data Selection

Kesav Adithya Venkidusamy

Bellevue University - Master of Science in Data Science

DSC680-T301 Applied Data Science (2231-1)

Professor Catherine Williams

10/02/2022

## Topic

Life Expectancy Prediction

## Business Problem

Everything has an expiration date; humans are no exception either. The term “life expectancy” refers to the number of years a person can expect to live. By definition, life expectancy is based on an estimate of the average age that members of a particular population group will be when they die.

We’re in an unprecedented era where humans are living longer with increased access to modern science and healthcare. It’s no secret, though, that life expectancy varies widely across the globe. Life expectancy depends on several factors, the two most important being gender and birth year. Generally, females have a slightly higher life expectancy than males due to biological differences. Other factors that influence life expectancy include:

- Race and ethnicity
- Family medical history
- Risky lifestyles

In this project, I aim to explore the parameters affecting the life span of individuals living in distinct countries and learn how the life span can be estimated with the help of machine learning models. I will also focus on exploring the parameters that greatly impact the life span of an individual.

## Datasets

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors of all countries. The datasets are made available to the public for health data analysis.

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

The data-set related to life expectancy, and health factors for 193 countries have been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nation website. Among all categories of health-related factors, only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in the health sector resulting in an improvement in human mortality rates, especially in the developing nations in comparison to the past 30 years. Therefore, in this project, we have considered data from the years 2000-2015 for 193 countries for analysis. The data was collected from WHO and United Nations website with the help of Deeksha Russell and Duan Wang.

**Characteristics**

| | |
|---|---|
| **Data Set Characteristics** | Multivariate |
| **Attribute Characteristics** | Categorical, Integer |
| **Associated Tasks** | Classification |
| **Number of Instances** | 2938 |
| **Number of Attributes** | 22 |
| **Missing Values** | Yes |
| **Area** | Health |

**Attributes information**

| Feature Name | Feature Description | Feature Type |
|---|---|---|
| Country | Country Observed | Discrete |
| Year | Year Observed | Continuous |
| Status | Status of the country; Developed or Developing Status | Discrete |
| Life expectancy | Life expectancy in age | Target |
| Adult Mortality | Adult Mortality Rates on both sexes (probability of dying between 15-60 years/1000 population). | Continuous |
| Infant deaths | Number of Infant Deaths per 1000 population | Continuous |
| Alcohol | Alcohol recorded per capita (15+) consumption (in liters of pure alcohol). | Continuous |
| Percentage expenditure | Expenditure on health as a percentage of Gross Domestic Product per capita(%). | Continuous |

| | | |
|---|---|---|
| Hepatitis B | Hepatitis B (HepB) immunization coverage among 1-year-olds (%) | Continuous |
| Measles | Number of reported Measles cases per 1000 population | Continuous |
| BMI | Average Body Mass Index of the entire population | Continuous |
| Under-five-deaths | Number of under-five deaths per 1000 population | Continuous |
| Polio | Polio (Pol3) immunization coverage among 1-year-olds (%) | Continuous |
| Total expenditure | General government expenditure on health as a percentage of total government expenditure (%) | Continuous |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%) | Continuous |
| HIV/AIDS | Deaths per 1 000 live births HIV/AIDS (0-4 years) | Continuous |
| GDP | Gross Domestic Product per capita (in USD) | Continuous |
| Population | The population of the country | Continuous |
| thinness 1-19 years | Prevalence of thinness among children and adolescents for Age 10 to 19 (%) | Continuous |
| thinness 5-9 years | Prevalence of thinness among children for Age 5 to 9(%) | Continuous |
| Income composition of resources | Human Development Index in terms of income composition of resources (index ranging from 0 to 1) | Continuous |
| Schooling | Number of years of Schooling(years) | Continuous |

## Methods

Since the target variable “Life Expectancy” is a continuous variable, I will run a linear regression on the dataset to determine the features that are mostly related or correlated to our target. Linear regression is commonly used for predictive analysis and modeling. Simple Linear Regression is a type of regression algorithm that models the relationship between a dependent variable and a single independent variable. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

Feature selection is one of the most important parts of any ML model. We always want to select those features which have the maximum effect on our final output. I will perform the following operations for feature selection:

- Lasso Method
- OLS Regression
- SK Learn

## Ethical Considerations

One of the ethical considerations for this project is the consideration of results from the analysis in decision-making. Some of the conclusions made from this project's study could be incorrect or misrepresented due to insufficient data. So, users of the model need to be careful in inferring outcomes and applying the actions in real-world scenarios.

The dataset considered for this analysis contains health information. So, it is necessary to ensure no personal and confidential information is present in the dataset. The dataset neither has personal nor confidential information as it is extracted from the World Health Organization website, and is available for public use.

## Challenges/Issues

One of the earliest challenges I might face is during the data preparation step of the model building. Identifying the correct features that contribute to the target, planning on how to handle the insufficient and null data, and deciding the next steps if the data is imbalanced to name a few. Another challenge would be choosing the correct technique/method to be applied for feature selection.

## Reference

KUMARRAJARSHI (2017). Life Expectancy (WHO). https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Arya Andhika (2019-08-12): Life Expectancy Prediction using Regression. https://rstudio-pubs-static.s3.amazonaws.com/534874_2bdd7c6645804fd1b240e1ca3a9eb9d6.html

World Health Organization: https://www.who.int/data/gho, https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates

Shashank Gupta. https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression

Caitlin McDonnell (2018-02-04). Machine learning to predict life expectancy. https://towardsdatascience.com/what-really-drives-higher-life-expectancy-e1c1ec22f6e1