---

title: "Life Expectancy Prediction"

author: "India\_Team"

date: "2022-08-28"

output:

html\_document: default

pdf\_document: default

---

---

output:

pdf\_document: default

html\_document: default

---

# Vision

Life Expectancy measures the average time a human being expects to live based on different parameters. Everything has an expiration date; humans are no exception either. The research proposes the prediction of the life expectancy of a person given the essential parameters with the help of machine learning models. Life expectancy is affected by socioeconomic, genetic, and environmental factors such as age, gender, environment, medical history, lifestyle choices, ethnicity, race, and many more. Although the prediction of life expectancy is not a hundred percent accurate, it allows helping in determining the course of treatment and procurement of medical services. It also paves a path for a risk-support model for the elderly according to the predicted risk.

# Stakeholders

1\. This study targets the general public as an estimate of life expectancy can enable an individual to work upon the factors that affect the longevity of their life.

2\. This study also aims to help the Government health bodies and private healthcare institutions as it could aid in the understanding of the amount of investment that should be made in the health sector and highlight the areas to work on to get maximum life expectancy of their citizens.

# Research Questions

The proposed study aims to answer the following research questions:

1. Which area should a country focus on to efficiently improve the life expectancy of its population?

2. What is the overall average life expectancy in most countries?

3. What is the ideal population size for any country to have a higher life expectancy?

4. What is the effect of different communicable/non-communicable diseases on life expectancy?

5. How much do lifestyle choices such as drinking alcohol impact a person's expected years of age?

6. What is the effect of different natural and unnatural death?

7. What is the influence of indirectly related factors such as GDP per capita on the expected life of a human?

8. How do Infant and Adult mortality rates affect life expectancy?

# Dataset range

Data creation is done by first extracting the following datasets :

1. Life expectancy WHO, (2000-2015)

2. Mortality caused by road traffic injury, (2000-2019)

We will be looking for several other datasets from the World Bank which depicts causes of deaths due to numerous circumstances - worldwide from 2000 to 2019. These dataset variables can be merged with the dataset by WHO and ​​further exploratory analysis can be run on those columns to get meaningful values.

The table below illustrates some of the variables used to predict the life expectancy.

| Variables | Description |

|---------------------------------|---------------------------------------|

| Country | Name of countries used for analyzing life expectancy. Total of 193 countries are used for analysis. |

| Year | It has the year in which the particular data points are recorded. |

| Infant deaths | . |

| Adult Mortality |. |

| GDP | Expenditure on health as a percentage of Gross Domestic Product per capita(%) |

| Deaths -XXX | Deaths caused because of different diseases/incidents. |

| Status | . |

|Traffic Mortality| . |

# Exploratory Data Analysis

Using a correlation map, the potency of the relationship between features and the dependent variable can be determined. The given map demonstrates the relation between some features and the life expectancy.

![Fig 1: Correlation map between variables](D:\UTS College Files\Sem 1\Statistical Thinking for Data Science\_Spring\_2022\AT\_2\23-08-22\correlation\_WHO.png)

After merging the Mortality caused by road traffic injury data to the existing data by WHO, a strong correlation is visible between Traffic Mortality and Life Expectancy, which can be seen in below correlation map.

The following scatter plots between the life expectancy and dataset features for a particular country confirms the correlation.

![Fig 2: Life Expectancy vs Population in Afghanistan](D:\UTS College Files\Sem 1\Statistical Thinking for Data Science\_Spring\_2022\AT\_2\23-08-22\Copy of lifeExpvsPop\_AFG.png)

![Fig 3: Life Expectancy vs GDP in Afghanistan](D:\UTS College Files\Sem 1\Statistical Thinking for Data Science\_Spring\_2022\AT\_2\23-08-22\Copy of lifeExpvsGDP\_AFG.png)

# Modelling Techniques Used

Numerous algorithms can help predict life expectancy, but not all of them can produce promising outcomes. There is never a perfect algorithm; there is always a trade-off between explainability, complexity, and accuracy. We will employ classical regression algorithms such as Multiple Linear Regression, Support Vector Regression, k Nearest Regression, and XG Boost for predicting life expectancy.

# Issues being anticipated

1. Issues might arise due to the dataset being imbalanced while training the regression model.

2. We might not have enough data for training and testing the model, which might result in Overfitting/Underfitting models.

3. Issues related to the missing values can also lead to an adverse impact on the accuracy of the ML model.

# Appendix

```{r eval=FALSE}

#Loading the Libraries

library(caTools)

library(tidyr)

library(GGally)

library(dplyr)

library(plotly)

library(ggplot2)

library(gapminder)

library(Amelia)

library(quantmod)

library(tidyverse)

library(reshape)

library(reshape2)

library(stringr)

library(corrplot)

library(RColorBrewer)

library(psych)

#Merging Traffic Mortality Dataset with Life Expectancy (WHO)

traffic <- read.csv("Mortality\_RoadAccident.csv")

head(traffic)

traffic <- traffic[, c(1,45:60)]

traffic <- drop\_na(traffic)

traffic <- melt(traffic, id=c("Country.Name"))

names(traffic)[names(traffic) == 'value'] <- 'Traffic.Mortality'

names(traffic)[names(traffic) == 'variable'] <- 'Year'

d2 <- traffic[order(traffic$Country.Name),]

write.csv(data, "Traffic\_Mortality\_2000\_2015.csv", row.names=FALSE)

d1 <- read.csv("Life Expectancy Data.csv")

d2\_new <- subset(d2, select = -c(Year))

d2\_new <- rename(d2\_new, Country = Country.Name, Year=Year.1)

df3 <- merge(d1, d2\_new)

write.csv(df3,"Life\_Expectancy\_Data\_merged.csv") #Merged dataset

life <- read.csv("Life Expectancy Data\_merged.csv")

head(life)

colnames(life)

describe\_data <- summary(life)

describe\_data

#Plotting missing values

missing.values <- life %>%

gather(key = "key", value = "val") %>%

mutate(is.missing = is.na(val)) %>%

group\_by(key, is.missing) %>%

summarise(num.missing = n()) %>%

filter(is.missing==T) %>%

select(-is.missing) %>%

arrange(desc(num.missing))

missing.values %>%

ggplot() +

geom\_bar(aes(x=key, y=num.missing), stat = 'identity') +

labs(x='variable', y="number of missing values", title='Number of missing values') +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

#Checking the correlation the parameters

Corrdata <- life %>%

select\_if(is.numeric)

#Plotting the correlation among all parameter

ggcorr(Corrdata,

method = c("pairwise","pearson"),

label = T,

label\_size = 2,

label\_round = 2,

hjust = 1,

size = 3,

color = "royalblue",

layout.exp = 5,

low = "green3",

mid = "gray95",

high = "darkorange",

name = "Correlation")

#Population Vs Life Expectancy of each country

population <- plot\_ly(data = life,x=~Life.expectancy,

y = ~Population, color=~Country,

text = ~paste("Population", Population ,'$<br>Country:'

,Country),

marker=list(size=5)) %>%

layout(title="Life Expectancy vs Population of countries in every year")

pop\_country <- population %>% layout(

xaxis = list(

type = "log"

)

)

pop\_country

#GDP vs Life Expectancy of each country

gdp\_country <- plot\_ly(data = life,x=~Life.expectancy,

y = ~GDP, color=~Country,

text = ~paste("GDP", GDP ,'$<br>Country:',Country),

type='scatter',mode='markers') %>%

layout(title="Life Expectancy vs GDP of countries in every year")

gdp\_country <- gdp\_country %>% layout(

xaxis = list(

type = "log"

)

)

gdp\_country

#Plot of diphtheria for each country with year

diph <- plot\_ly(life, x=~Life.expectancy,

y = ~Diphtheria, color=~Country,

text = ~paste("Deaths due to Diphtheria", Diphtheria

,'$<br>Country:',Country),

type='scatter',mode='markers',

marker = list(size=5))

diph <- diph %>% layout(title="Life Expectancy Vs Deaths due to Diphtheria")

diph <- diph %>% layout(

xaxis = list(

type = "log"

)

)

diph

#Infants death of each country

infants\_country <- plot\_ly(life, x=~Life.expectancy,

y = ~infant.deaths, color=~Country,

text = ~paste("Infant Mortality", infant.deaths

,'$<br>Country:',Country),

type='scatter',mode='markers',

marker = list(size=15))

infants\_country <- infants\_country %>% layout(title="Life Expectancy

Vs Infant Deaths based on

Country Year Wise")

infants\_country <- infants\_country %>% layout(

xaxis = list(

type = "log"

)

)

Infants\_country

```

# References

1. Aman Vishnoi. (2022).Heath Dataset https://www.kaggle.com/code/amanvishnoi777/health-dataset/data. Kaggle

2. Mortality caused by Road Traffic Injury. (World Bank). https://data.worldbank.org/indicator/SH.STA.TRAF.P5. The World Bank

3. Ivan Chave. (2022). Cause of Death - Our World in Data. <https://www.kaggle.com/datasets/ivanchvez/causes-of-death-our-world-in-data> Kaggle.

4. Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total). The World Bank.

5. Shashank Gupta.(NA).Life Expectancy Prediction Using Linear Regression.

6. Dr.Vikram Bali1,Dr.Deepti Aggarwal,Sumit Singh,Arpit Shukla. (2021). Life Expectancy: Prediction &Analysis usingML.

7. R Documentation. Merge data frames in R. https://r-coder.com/merge-r/

8. Black\_Raven (James Ng). (2019). Regression Analysis on Life Expectancy.

9. Esteban Ortiz-Ospina. (2017). “Life Expectancy” – What does this actually mean?. Our World in Data.

10. R Documentation. (NA). Exploratory Data Analysis. R for Data Science.

11. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [https://www.R-project.org/](https://www.r-project.org/).

12. <https://r4ds.had.co.nz/index.html>

13. <https://www.geeksforgeeks.org/data-visualization-in-r/?ref=lbp>