PROJECT NAME: LIFE EXPECTANCY

TEAM NAME: OPTIMISTIC BOOSTERS

GROUP MEMBERS:

1. Daphney Ezinne Igwe
2. Oluwanifesimi Okesola
3. Ayeni, Damilola Motunrayo
4. Grace Okewale
5. Stephanie Bomett
6. Medinat Apanpa-Kolade
7. Kama Mary Ofuru
8. Justina Achuka
9. Ednahngetich10
10. Patience Mba
11. Oluwabunmi Bankole
12. Patience Chizoba Mba
13. adachukwu.jay
14. asarebediakofrancisca81
15. Kehinde Onyeche-Ogazi
16. lidadarlsy
17. Sandraobeten
18. azeezabolanle40
19. jackyogingo
20. owansharry
21. Kama Mary
22. Itunu Gbadamosi

**Introduction**

What is life expectancy?

Life expectancy is a statistical measure of the average time an organism is expected to live.

Life expectancy depends on many factors like age, income, availability of clean water, pollution etc.

Some Important use cases for the life expectancy metric include clinical research, pricing insurance premiums.

**Problem Statements**

Although there have been a lot of studies undertaken in the past on factors affecting life expectancy, it was found that, the test data set used were only treated with one model. Also, some important data set variables were also excluded, coupled with an extreme short test period of a year.

The objective of this study is to build a reliable predictive model based on multiple linear regression while considering data from a period of 2000 to 2015 for all the counties and important immunizations like Hepatitis B, Polio and Diphtheria.

**Data Wrangling**

Data set was sourced from Kaggle, a free open-source website. Which contains various kinds of data.

The data consists of 2938 rows and 22 columns.

In this data set, life expectancy which is measured in age is our response variable and the regressors include:

* Country
* Year
* Status
* Adult Mortality
* Infant deaths
* Alcohol
* Percentage expenditure
* Hepatitis B
* Measles
* BMI
* Under-five deaths
* Polio
* Total expenditure
* Diphtheria
* HIV/AIDS
* GDP
* Population
* Thinness 1-19 years
* Thinness 5-9 years
* Income composition of resources and Schooling

**Variable Description**  
All columns are numerical except ‘Status column’ which has 2 categories: Developing/Developed and, Country which contains the names of 193 countries used in the data set.

Some of the regressors are defined below.

* Schooling: Number of years of Schooling (years)
* Income composition of resources: Human Development Index in terms of income composition of resources.
* Population: Population of the country.
* GDP: Gross Domestic Product per capita (in USD).
* HIV/AIDS: Deaths per 1 000 live births HIV/AIDS (0-4 years).
* Diphtheria: Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
* Polio: Polio (Pol3) immunization coverage among 1-year-olds (%)
* Alcohol: Alcohol, recorded per capita (15+) consumption (in liters of pure alcohol)

**Data visualization**

Quick observation and hypothesis can be made from various graphical presentation of the dataset. Below are some of the deduced facts:

* 80% of the dataset were mined from developing countries.
* A huge difference between life expectancy in developed vs developing countries.
* Interestingly BMI and Alcohol are positively related to life expectancy.
* Adult mortality is negatively related to life expectancy.
* Relationship of each regressor with the response variable

**Data Preprocessing**

Various standard methods were employed to process the given dataset to a suitable standard for modelling.

these methods include:

* Dealing with missing values, using calculated mean values.
* Encoded the categorical data, status using label encoder.
* Converted the categorical variable country, into a dummy variable.
* Dropped column year cos it was not useful.
* Removed the outliers.
* Split the model into test (20%) and train data (80%)
* Standardized our data using standard scalar.

**Modeling**

There are various modelling algorithms, some open-source and others paid.

In fulfilling the requirements of this project, four open-source algorithms were use in testing our predicted results,

* Random Forest Regression (RFR),
* Extra Trees Regression (ETR),
* Ridge Regression (RR) and
* Lasso Regression (LR).

The test results obtained include:

* Y-predictions
* Mean Squared Error and
* Accuracy Score