# **PROJECT REPORT**

**1.Choice of Dataset**

This dataset is related to life expectancy, health factors for 193 countries has been collected from the WHO data repository website and its corresponding economic data was collected from 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 health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single dataset

The motivation behind choosing the dataset was **formulating a regression model to predict life expectancy** based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries. Important immunization like Hepatitis B, Polio and Diphtheria will also be considered. In a nutshell, this project will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well. From the observations this dataset to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting **measures to be given importance in order to efficiently improve the life expectancy of humans**. Another aim is to come up with a standard formula(model) to predict the life expectancy based on these factors.

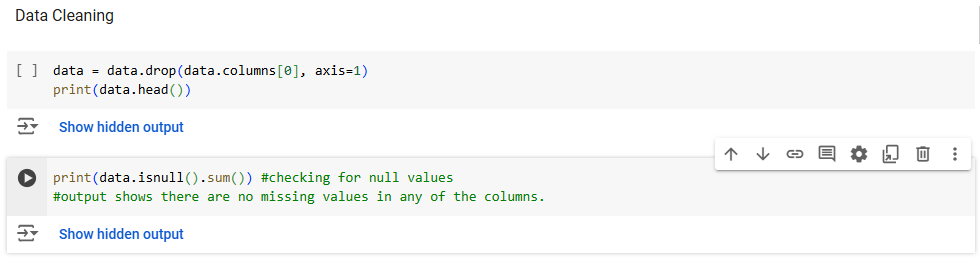
**2.Understanding Dataset**

**Description:** The final merged file(final dataset) consists of 20 Columns and 2938 rows which meant 19 predicting variables. All predicting variables are divided into several broad categories ​Immunization related factors, Mortality factors, Economical factors and Social factors

Link to the Final Dataset: <https://drive.google.com/file/d/1QJc5Em9szRmGdFsj83DCbr2vyhPY67NU/view?usp=drive_link>

**Data-preprocessing:**

* Dropped the unnecessary ‘year’ column
* Checked for any missing values , also duplicates



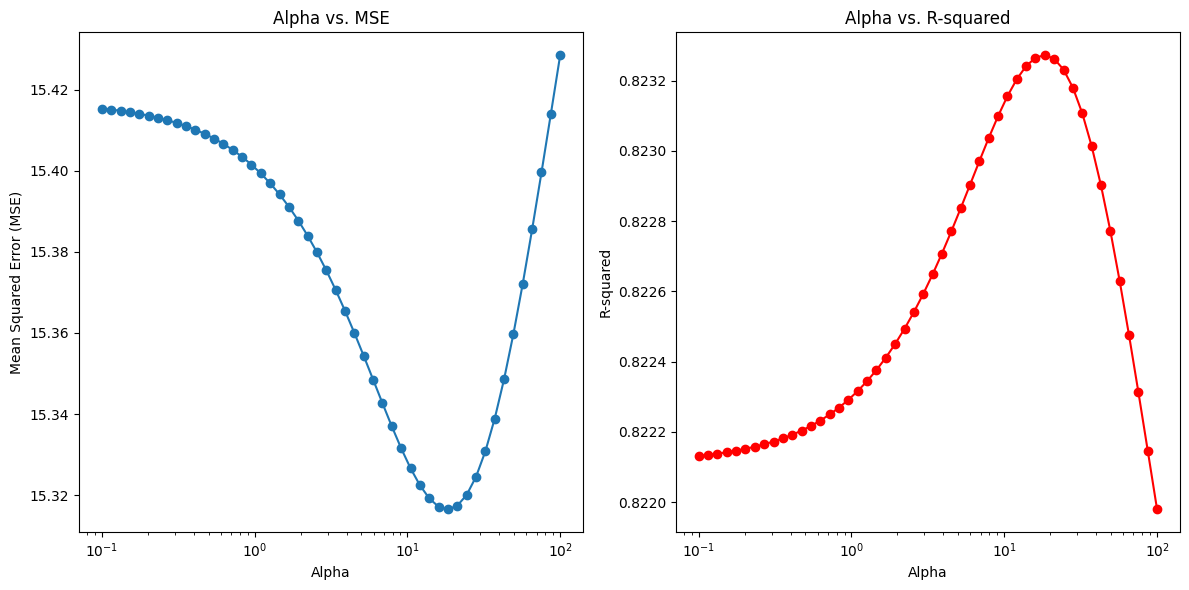
**3.Algorithm Usage**

**Selection of Algorithm and Justification:** Firstly, I started with simple regression models and evaluated their performance on the data (with test/train ratio=0.8):

Model performance:

Linear Regression - Mean Squared Error: 15.417159484436095, R^2 Score: 0.8221111322562291 Ridge Regression - Mean Squared Error: 15.317893770439237, R^2 Score: 0.8232564966462455

The value of alpha in ridge regression was determined by plotting the following graphs and was found to be 18



1. **Minimal Performance Gain**: The difference in MSE and R^2 between Ridge and Linear Regression is minimal, which means regularization provides some benefit but isn’t transformative. This implies that regularization is helping only slightly by stabilizing the model.
2. **Ridge Regression’s Effectiveness**: Ridge works well when slight regularization is enough to improve stability without altering model complexity significantly. If interpretability is essential and you prefer a model close to Linear Regression, Ridge is a solid choice. The model’s coefficients remain largely interpretable while slightly reducing overfitting

Lasso Regression - Mean Squared Error: 16.481828753005992

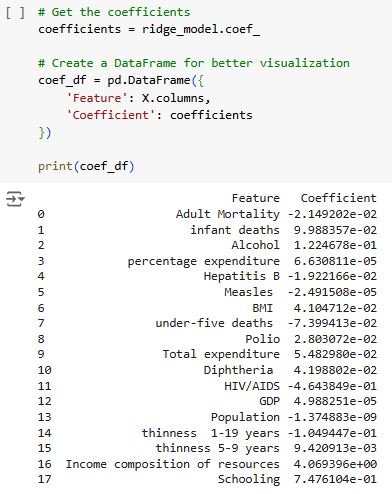
Lasso Regression - R^2 Score: 0.8098265858779754

(performed worse then the two)

As interpretability and simplicity are priorities, Ridge Regression is a good final choice, as it provides stability with a small improvement

**IMPLEMENTATION:**<https://colab.research.google.com/drive/11dklxAdYwqoJFl8kOSDUxbMAA2jwhnqW?usp=sharing>

**4.INFRENCES DRAWN**

****

* Positive coefficients indicate features that **increase life expectancy** as their values increase, while negative coefficients indicate features that **decrease life expectancy** as they increase.
* **Income Composition of Resources** (4.069396): This feature has the strongest positive impact on life expectancy, suggesting that countries with higher income composition or resources see a marked increase in life expectancy.
* What is the impact of schooling on the lifespan of humans? (0.747): Education level has a substantial positive effect on life expectancy, aligning with findings in public health that link education with better health outcome
* **HIV/AIDS** (-0.46438): As expected this feature has a significant negative coefficient, indicating that higher HIV/AIDS prevalence strongly correlates with a reduction in life expectancy.
* Features such as Population, GDP, Percentage expenditure, Measles have very small weights means these features could be removed for model simplification
* How does Adult mortality rates affect life expectancy? (-0.0214): Higher adult mortality is understandably linked with lower life expectancy
* **Thinness 1-19 Years** (-0.104): Indicates that lower BMI in children and adolescents (associated with malnutrition) negatively impacts life expectancy. Hence countries having high value of this must take initiatives to improve nutrition of childrens (especially those from lower backgrounds)
* **Alcohol Consumption**: Interestingly, alcohol has a small positive coefficient. While this might seem counterintuitive, it could imply that in certain contexts, moderate alcohol consumption is a marker of lifestyle factors correlated with longer life expectancy. It doesn’t mean alcohol directly improves life expectancy but rather that people in regions with moderate alcohol intake might also have access to better healthcare and social services.

**Conclusion**: The analysis indicates that **investments in education and income equality**, **targeted health interventions** for high-risk factors (like HIV/AIDS and childhood mortality), and **effective healthcare spending** could have the most meaningful impact on increasing life expectancy. Additionally, some features like GDP or population, though not directly impactful here, could be monitored as situational indicators.