**Life expectancy using linear regression model**

About the data

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 for all countries The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same 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. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo,Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model dataset. The final merged file(final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories:​Immunization related factors, Mortality factors, Economical factors and Social factors.

Features:

* Country
* Year
* Status
* Life expectancy
* 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 features
* Schooling

To predict:

* Life expectancy

Bibliography

https://towardsdatascience.com/why-do-we-set-a-random-state-in-machine-learning-models-bb2dc68d8431

https://www.kdnuggets.com/2019/03/beginners-guide-linear-regression-python-scikit-learn.html

https://towardsdatascience.com/how-different-factors-have-an-influence-on-your-life-expectancy-7b807b04f33e

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