Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning

Siddhant Sunil Meshram
Department of Electronics and Telecommunication
SIES Graduate School of Technology
Nerul, India
siddhant.meshram@ieee.org

2020 IEEE Bombay Section Signature Conference (IBSSC) | 978-1-7281-8993-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/IBSSC51096.2020.9332159

***Abstract*—Life Expectancy is an important metric to assess the health of a nation. This paper presents a comparative analysis of life expectancy between developed and developing countries with the help of a Supervised Machine Learning model. The prediction model is trained using three regression models, namely Linear Regression, Decision Tree Regressor and Random Forest Regressor. The selection of model is done on the basis of R^2 score, Mean Squared Error & Mean Absolute Error. Random Forest Regressor is selected for the development of the prediction model for life expectancy, as it had R^2 score as 0.99 and 0.95 on training & testing data respectively, along with 4.43 and 1.58 as the Mean Squared Error & Mean Absolute Error. The comparative analysis is done on the basis of HIV/AIDS, Adult Mortality and Expenditure on Healthcare, as they are the important features suggested by the model. The study undertaken suggests that, developed countries have high life expectancy as compared to developing countries. India has high adult mortality as compared to considered developed countries because of the low expenditure on healthcare. The insights from this analysis can be used by Government and Healthcare sectors for the betterment of society.**

Keywords—Random Forest Regressor, Machine Learning, Life Expectancy

I. Introduction

Life Expectancy of a country is reflected by its socio-economic conditions. The health and well-being of a population can be explained through life table statistics [1]. Life expectancy is the most frequently used life table statistics, which is the average number of years of life remaining [2]. The overall mortality level of a population can be reflected by life expectancy at the time of birth. In 2016, at birth, the Global Life Expectancy was 72.0 years (74.2 years for females and 69.8 years for males), as observed by WHO. It ranged from 61.2 years in the WHO African Region to 77.5 years in the WHO European Region, giving a ratio of 1.3 between the two regions [3]. The datasets or the life tables can be referred to calculate life expectancy of plants, animals or humans [4]. These datasets or tables can be used to predict any country's life expectancy.

This paper is aimed at predicting average life expectancy of a country on the basis of given features and produce some comparative analysis between developed countries like United States of America (USA), United Kingdom (UK), France (FRA), Germany (GER), Australia (AUS) and developing countries like India (IND), Russia (RUS), South Africa (SA), Brazil (BRA), China (CHN). The significance of life expectancy of a country rests on several factors like economic circumstances, regional variations, education, sex differences, physical illnesses, mental illnesses, alcohol intake, GDP, expenditure on healthcare system and other demographic factors. Life expectancy has seen some monumental improvements over the course of 20th and 21st centuries among the developed countries of the world [1].

Machine learning (ML) comprises elements of mathematics, statistics, and computer science. In the advancement of artificial intelligence (AI), ML has played a vital role. Also, the development of 'intelligent products' using variety of data has been employed in both academics and industry. The 21st century has seen explosions in the availability of big data, ML, and data science which have been used in developing different learning methods and helped in offering tremendous potential to enhance medical research and clinical care, especially as providers increasingly employ electronic health records [5]. ML being a dominant technology with its accuracy on predictions for many set of problems is highly used in enhancing life expectancy by monitoring health and reducing mortality rates [6].

II. Data Exploration

The dataset is extracted from WHO and United Nations website & obtained from Kaggle [7]. The data undertaken for study provides a timespan from 2000 to 2015. It has 23 columns out of which 22 are features & 1 is desired output. Out of the 22 features, year & status of the country are redundant features, so those were ruled out of consideration. All the rows and columns were grouped by countries. The rows of countries with blank or no values were filled with mean values of the corresponding features of that country. Several data visualizations are made through scatter plots, to check the correlation between the features & desired output.

The correlation matrix, which is visualized with a heatmap (as shown in Fig.1) plotted with the help of 'seaborn' library of python, tells us about the correlation between Life Expectancy and other features. The legend on the right of the visualization tells that warmer colors show high and positive correlation and colder colors show low and negative correlation. The linear historical relation between the desired output and the features can be estimated through a correlation matrix. In multivariate analysis, it plays an important role, as it elaborates relationship between different components [8]. Looking at Fig.1, one can deduce that, adult mortality, infant deaths, measles, under-five deaths, HIV/AIDS, thinness 10-19 and 5-9 years has a negative correlation with Life Expectancy. Also, Percentage Expenditure on healthcare, BMI, GDP, Income composition of resources and schooling has a positive correlation with Life Expectancy.

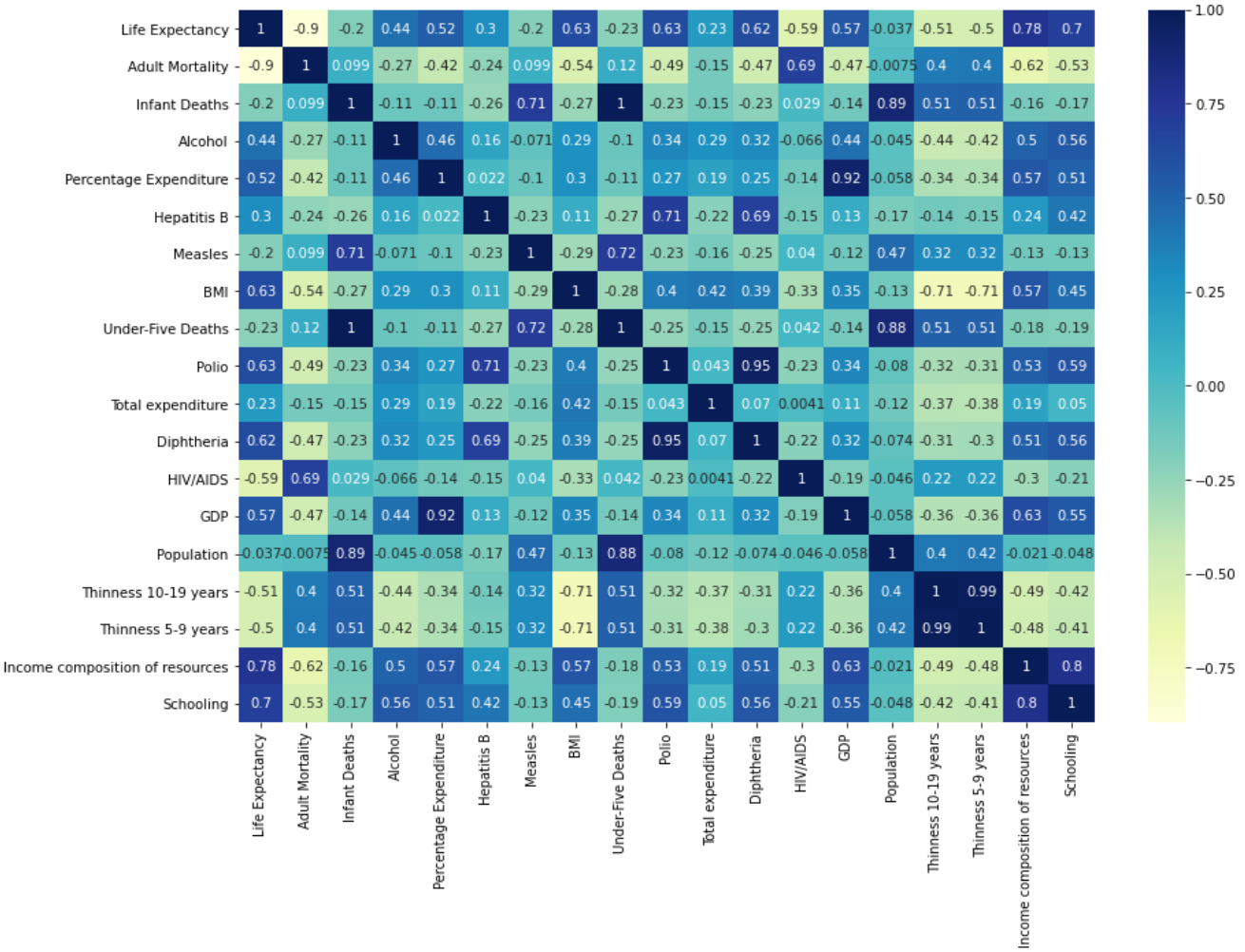

Fig. 1. Correlation Matrix visualization through heatmap

III. Methodology

In order to predict life expectancy, different regression models are used, namely Linear Regression, Decision Tree Regressor & Random Forest Regressor. The simplest and most classic linear method for regression is the linear regression or ordinary least squares (OLS) [9]. Linear regression is referred as a change in the value of the dependent variable with respect to other independent variables [10]. The benefit of using this algorithm is that, it has no parameters, but it cannot control a model's complexity [9]. Decision trees models are widely used for classification & regression tasks. The decisions of these models are generally based on how they learn a hierarchy of "if-else" questions. These models have a drawback of overfitting the training data [9]. So, in order to overcome this drawback, the random forests are introduced. Random forests are a collection of many decision trees which are likely to overfit on part of data. This overfitting is reduced by averaging the results of all the trees & helps in retaining the predictive power of trees [9]. The selection of the model was done on the basis of R^2 score, mean squared error & mean absolute error. This is further discussed in Result & Discussion Section of this paper.

The 'sklearn' or 'sci-kit learn' library of python or popularly known as machine learning library is used to split the data into training and testing sets. For all the three models, the dataset is divided into 80-20, i.e. 80% for training and 20% for testing. Then the training set is trained with the help of 'fit' function of the same python library. A pickle file of the trained model is created by importing 'pickle' library of python. This pickle file is then used while creating a web user interface of the model. With the help of 'flask', yet another library from python, along with a HTML code are used to create a web user interface (as shown in Fig.2).

IV. Result and Discussion

The model selected for the prediction of life expectancy was 'Random Forest Regressor', as it had an R^2 score of 0.99 & 0.95 on training & testing data respectively. Also, the Mean Squared Error (MSE) and Mean Absolute Error (MAE) of this model were 4.43 & 1.58 respectively. These R^2, MSE & MAE values were found out to be better as compared to the values in Linear Regression & Decision Tree Regressor models. The feature importance according to the selected model is shown in the Fig.3.

Looking at that horizontal bar plot, features like Adult Mortality (adult mortality rates of both sexes ,probability of dying between 15 and 60 years per 1000 population), HIV/AIDS (Deaths per 1000 live births, 0-4 years), Income composition of resources, schooling (Number of years of Schooling), percentage expenditure (Expenditure on health as a percent of Gross Domestic Product per capita), thinness (Prevalence of thinness) and BMI (Average Body Mass Index of entire population) are most important features, which may have a great impact on life expectancy of a country.

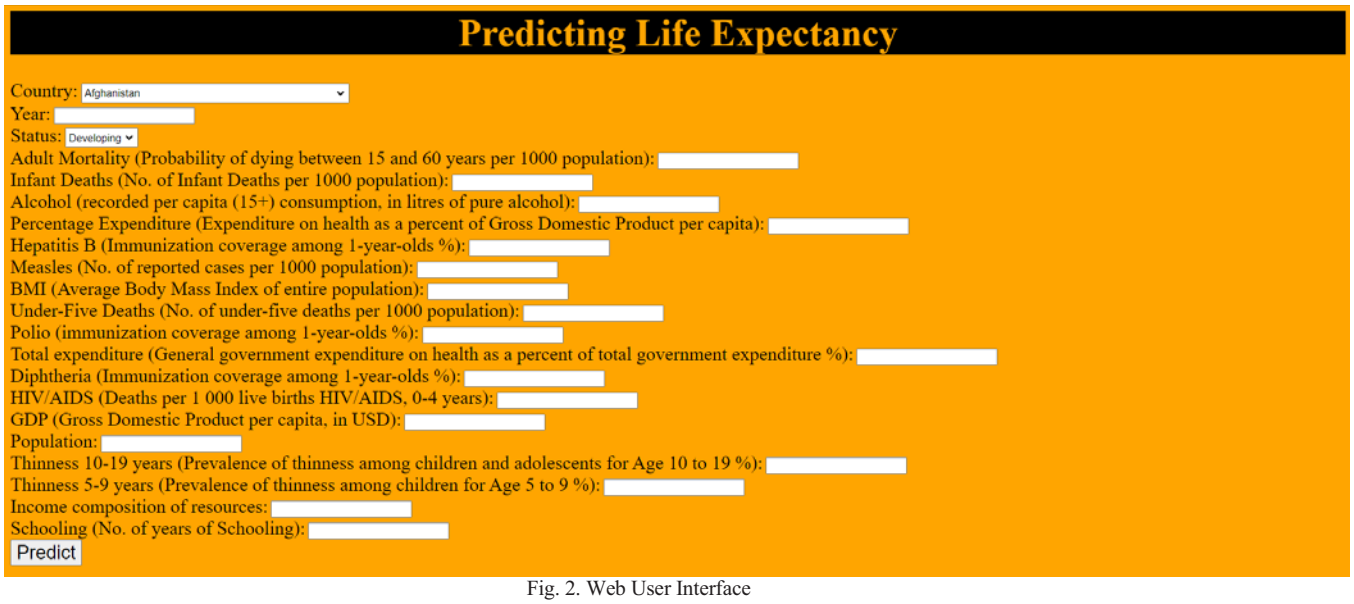

Fig. 2. Web User Interface

The difference between the life expectancy of some of the developing and developed countries can be seen through the visualization (as shown in Fig.4). The data for all the visualization is extracted from the same dataset used for the prediction model. From the graph, we observe that the developed countries have a better life expectancy than the developing countries.

From the Fig.3, important features observed HIV/AIDS, Adult Mortality & Expenditure on Healthcare are considered for comparative analysis. The Fig.6 shows comparison of average deaths due to HIV/AIDS during 2000-2015 among developed & developing countries. The present investigation shows that, South Africa has a higher average death due to HIV/AIDS and hence experience low life expectancy among developing countries. In India, according to [11], the national prevalence of AIDS was about 0.26% in 2016. Following the antiretroviral therapy, the life expectancy increased & total number of HIV-positive persons remains stable at 2.1 million [11]. All the other countries considered in this analysis, have a negligible amount of average deaths due to HIV/AIDS compared to South Africa.

According to analysis, Adult Mortality being the most important feature (refer Fig.3), the comparison of adult mortality among developed & developing countries has been plotted (as shown in Fig.5). For formulating economic policy decisions, mortality statistics can be helpful [12]. According to [13], in India, adults which belong to deprived castes/tribes or have children with a low level of education are mainly illiterate. Due to illiteracy they have a low level of household income & hence, experience high mortality. From the Fig.5, it is observed that Adult Mortality among the developing countries like South Africa, Russia, India and Brazil is high as compared to the developed countries.

In comparison of Expenditure on Healthcare among developed and developing countries, the data of United States of America and United Kingdom were not available. It is observed from the Fig.7, Australia spends highest on their healthcare among the considered developed countries. Germany and France also give equal importance to healthcare. Among the considered developing countries Russia spends the highest. In India, according to [14], Rs. 200 was the maximum annual per capita public health expenditure in 2016, i.e. just 0.9% of GDP, while in private spending it ranked at 18th. In South East Asian countries, except Pakistan, India is recorded as one of the lowest countries, when it comes to public expenditure on health [15]. In most cases, the developing countries spend less money on health as compared to developed countries in both, as percent of GDP and out of their total budget [15]. The same can be deduced form the graph in Fig.7.

Any error in data, can result in wrong predictions. Since the insights are purely based on the available data, there may be other factors or features on which the Life Expectancy depends on.

V. Conclusion

Prediction of Life Expectancy using Machine Learning can be used to get insights into economic and healthcare development of any country. The work undertaken reports that, South Africa has highest deaths per 1000 live births (0-4 years) in HIV/AIDS. Among the considered developed countries, Australia invests more in healthcare, which reflects in their low adult mortality rate. Compared to considered developed countries, India's adult mortality rate is much higher. Among the considered developing countries South Africa spends more on healthcare, but due to high HIV/AIDS, its adult mortality is highest. The present developed model, suggests developing countries have a less life expectancy than the developed countries. This model will help us to increase life expectancy considering the impact of a specific factor on the average lifespan of people of specific country. Additional features can be included in the model by using an enhanced dataset. The various NGO's, Corporate Sectors & Government may use this model and analysis to propose their future plans and policies related to healthcare.

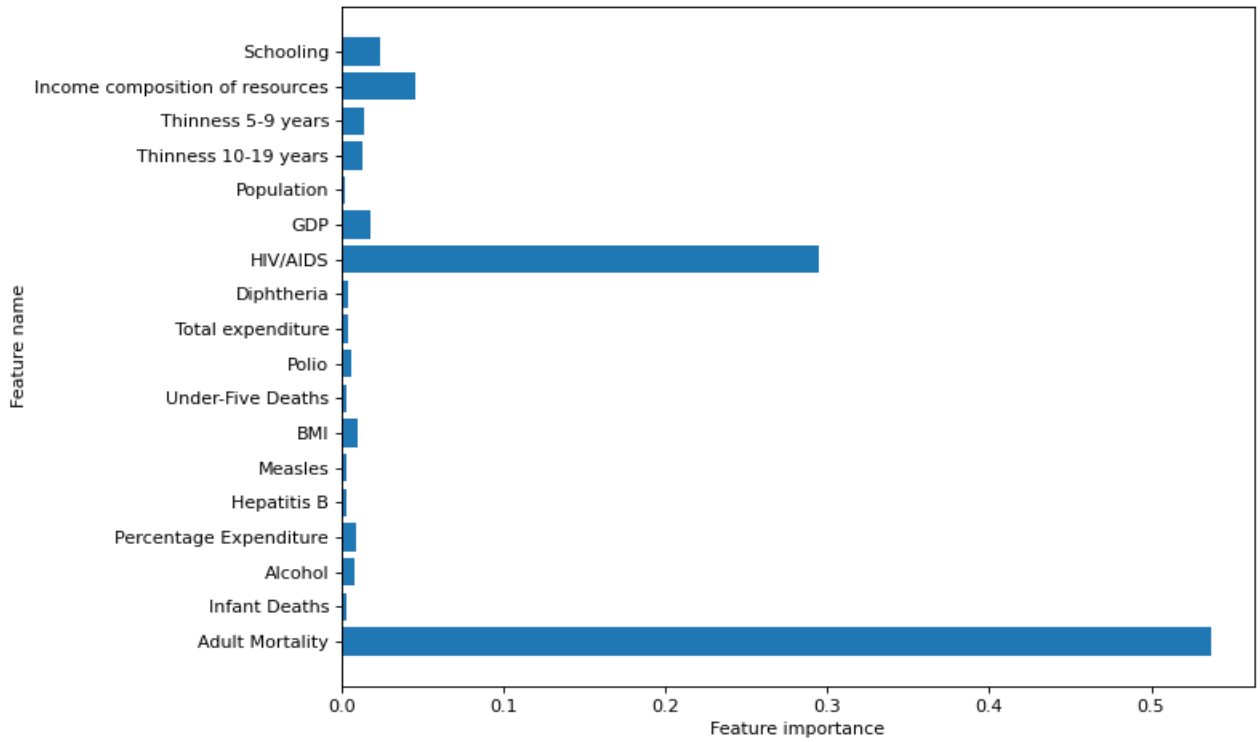

Fig.3 Feature Importance Plot

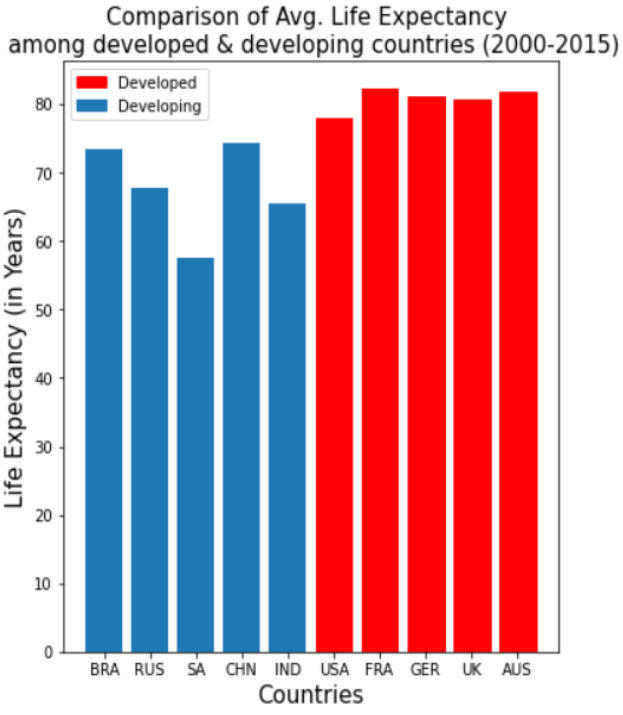

Fig.4

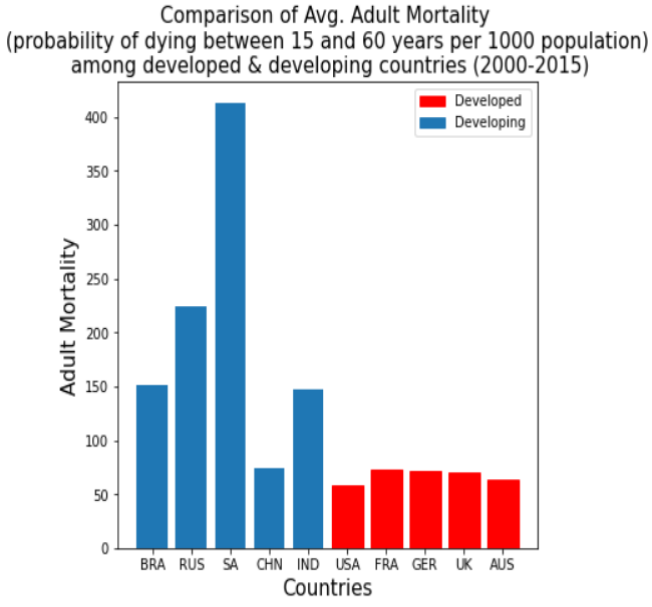

Fig.5

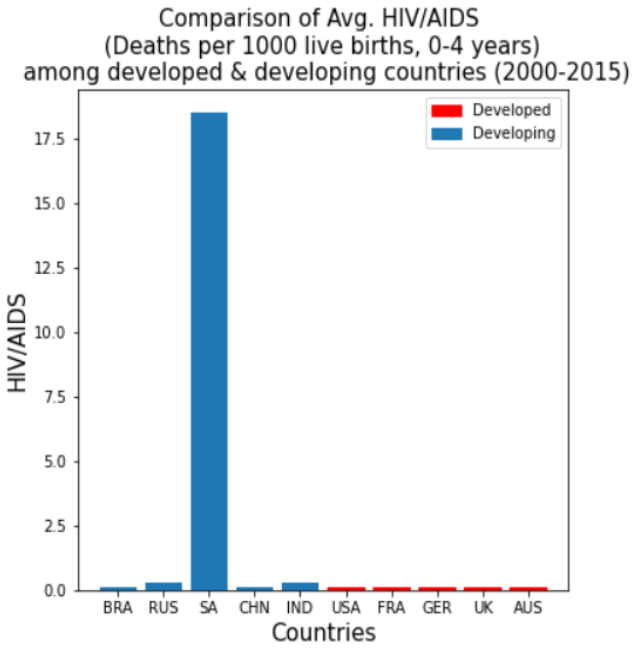

Fig. 6

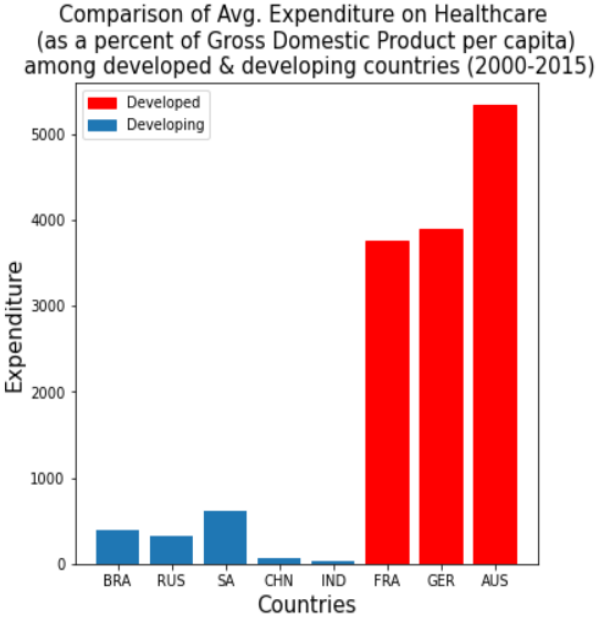

Fig. 7

ACKNOWLEDGMENT

I would like to thank Kumar Rajarshi, Deeksha Russell and Duan Wang for data collected by them on Life Expectancy from WHO and United Nations website.

REFERENCES

[1] J. Y. Ho and A. S. Hendi, " Recent trends in life expectancy across high income countries: Retrospective observational study," BMJ Clinical Research, vol. 362, pp. 1-13, August 2018. [Online]. Available: https://www.bmj.com/content/362/bmj.k2562

[2] E. Arias, "United States Life Tables, 2009," National Vital Statistics Reports, vol. 62, no. 7, pp. 1-63, January 2014. [Online]. Available: https://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_07.pdf

[3] WHO. Global Health Observatory (GHO) data. Accessed: July 30, 2020. [Online]. Available: https://www.who.int/gho/mortality_burden_disease/life_tables/situation_trends_text/en/

[4] A. Mandal, "What is Life Expectancy?," News-Medical, February 2019. Accessed: July 30, 2020. [Online]. Available: https://www.news-medical.net/health/What-is-Life-Expectancy.aspx

[5] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC Medical Research Methodology, vol. 19, no. 64, pp. 1-18, March 2019. [Online]. Available: https://doi.org/10.1186/s12874-019-0681-4

[6] V. Malpe and P. Tugaonkar, "Machine LearningTrends in Medical Sciences," 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), pp. 495-499, August 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8653756

[7] Kaggle. Life Expectancy (WHO). Accessed: July 30, 2020. [Online]. Available: https://www.kaggle.com/kumarajarshi/life-expectancy-who/data

[8] T. Pham-Gia and V. Choulakian, "Distribution of the Sample Correlation Matrix and Applications," Open Journal of Statistics, vol. 4, no. 5, pp. 330-344, August 2014. [Online]. Available: https://www.scirp.org/html/2-1240369_48571.htm

[9] A. C. Müller and S. Guido, Introduction to Machine Learning with Python, 1st ed., O'Reilly Media, Inc., October 2016

[10] K.-Y Lee, K.-H Kim, J.-J Kang, S.-J Choi, Y.-S Im, Y.-D Lee, and Y.-S Lim, "Comparison and Analysis of Linear Regression & Artificial Neural Network," International Journal of Applied Engineering Research, vol. 12, no. 20, pp. 9820-9825, January 2017. [Online]. Available: https://www.ripublication.com/ijaer17/ijaerv12n20_77.pdf

[11] R. S. Paranjape and S. J. Challacombe, "HIV/AIDS in India: an overview of the Indian epidemic," Special Issue: The Mouth and AIDS: Lessons Learned and Emerging Challenges in Global Oral Health. Seventh World Workshop on Oral Health & Disease in AIDS, vol. 22, pp. 10-14, April 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/odi.12457

[12] N. Saikia and F. Ram, "Determinants of adult mortality in India," Asian Population Studies, vol. 6, pp. 153-171, July 2010. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/17441730.2010.494441

[13] N. Saikia, J. K. Bora, and M. Luy, "Socioeconomic disparity in adult mortality in India: estimations using the orphanhood method," Genus, vol. 75, no. 7, pp. 1-17, February 2019. [Online]. Available: https://genus.springeropen.com/articles/10.1186/s41118-019-0054-1

[14] H. Sudhakara and T. Rajendraprasad, "Healthcare Expenditure in India -An Analysis," Shanlax International Journal of Economics, vol. 5, pp. 26-34, December 2016. [Online]. Available: https://www.researchgate.net/publication/333186302_Healthcare_Expenditure_in_India_-An_Analysis

[15] S. K. Hooda, "Changing pattern of public expenditure on health in India issues and challenges," ISID - PHFI Collaborative Research Centre Institute for Studies in Industrial Development, pp. 1-37, March 2013. [Online]. Available: http://isid.org.in/pdf/wp154.pdf