School of Chemical and Biomedical Engineering

CH0494: Data Science and Artificial Intelligence

Mini Project

Team Members: Neo Si Yang (U1922021J), Muhamad Ar Iskandar (U1922535L)

Project contribution: Neo Si Yang (50%), Muhamad Ar Iskandar (50%)

BACKGROUND

Life expectancy is known as the key metric for assessing population health. It represents mortality along the entire life course and tells us the average lifespan of a human. Life expectancy has increased significantly over many years since the Age of Enlightenment (17th to 19th centuries). With improvements in healthcare and technology, overall life expectancy has increased from less than 30 years to 72 years globally. However, in many countries of the developing world such as Sub-Saharan Africa in particular, the life expectancy remains as low as 50 to 60 years. In 2019, the population of the Central African Republic has the lowest life expectancy of 53 years [1]. Over a decade, there have been a lot of studies undertaken to improve life expectancy in developing countries. In general, four main factors which affect the life expectancy of a country are economic, social, mortality and health. The main focus in this mini project is on developing countries. In this mini project, we will be using a data set provided by the World Health Organisation. The data set is a CSV file which contains the life expectancy data across all countries (developed and developing) from 2000 to 2015. Statistical results are given in the data set such as number of infant deaths per 1000 population and expenditure on health as a percentage of Gross Domestic Product per capita, etc. The data set can be downloaded from the following URL: https://www.kaggle.com/kumarajarshi/life-expectancy-who [2].

OBJECTIVES

- To investigate what is the most important factor a developing country should focus on to improve its life expectancy.
- To determine the feasibility of predicting the effect of a factor in the future based on past data
- Use data science techniques (linear regression) to solve a real-life problem through data analysis.
- To conclude potential findings based on the results of the mini project.

PROCEDURES

1. **Importing Libraries**

   The first step involved importing the following libraries into Jupyter notebook which are Numpy, Pandas, Matplotlib, Seaborn and Sklearn. Sklearn is the most widely used package for the machine learning process. The following sub-packages were used: LinearRegression and mean_squared_error.

2. **Reading the data**

   The data was read by Pandas and stored in the **lifeData** variable. After importing the data set, a quick look was taken using the head function. The shape, dtypes and info function was also executed to understand the basic details of the data set such as the number of rows and columns and the total number of integers, float and object variables.

3. **Extracting the required factors**

As the main focus is on developing countries, only data sets from developing countries were used. The drop function was used to remove information on developed countries. The following code was used: **lifeData.drop(lifeData[lifeData['Status'] != 'Developing'].index, inplace = True)**. This code removed rows of information if they do not belong to developing countries. After which, the describe function was used to check if the column 'Status' contains only the developing category. After which, a heatmap was plotted to show the correlation between every factor and life expectancy. Upon taking a closer look, all the factors that have shown a strong correlation with life expectancy were extracted and stored into another variable called **lifeNumData**. The factors chosen are:

Economic: Income Composition of Resources **(ICOR)**
Social: Schooling
Health: HIV/AIDS
Mortality: Adult Mortality

4. **Cleaning the data set**

To prevent any wrong prediction or classification for any given model used, all null data have to be removed using the drop function. The downside of this function was that the whole row of data would be eliminated even if the rest of the features are filled and informative. After the data set was cleaned, it was stored into another variable called **modlifeData**.

5. **Sorting of data set in chronological order.**

The data set **'modlifeData'** was sorted from the earliest to latest year, particularly from 2000 to 2015.

6. **Finalised heatmap**

After cleaning and sorting the dataset into chronological order, a heatmap was plotted to show the correlation between the four factors (Income composition of resources, Schooling, HIV and Adult Mortality) and life expectancy. The finalised heatmap is shown in Figure 1.

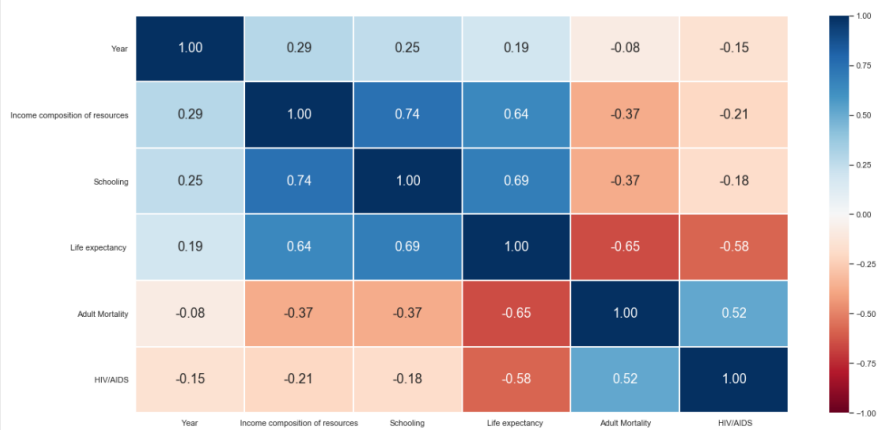

Figure 1: A heatmap showing the correlation between different factors and life expectancy

From Figure 1, it shows that adult mortality and HIV/AIDS have a negative correlation of -0.65 and -0.58 respectively while schooling and income composition of resources (ICOR) have a positive correlation of 0.69 and 0.64 respectively. This means that if the number of years of schooling and ICOR increases, life expectancy also increases. On the other hand, if adult mortality rate and the number of deaths due to HIV/AIDS increases, life expectancy decreases.

7. Splitting of data into a fixed number of Train and Test sets

Before executing the linear regression model, modlifeData was split into a fixed number of train and test sets. Approximately 75% of the dataset was used in our train model while the remaining 25% was used in our test model. As mentioned in step 5, the dataset had already been sorted in chronological order. The train and test set were arranged in such a way that the train set contains data from the previous years (2000 to 2011) and the test set contains data from the latest years (2012 to 2015). The reason behind this splitting is that if the train model from past years can have high performance on the test sets which are from latest years, this means that the model is suitable to predict data in the future. To execute this, the following code was used: **modlifeData.Year.searchsorted('2012', side='left')**. This is to check the index of the dataset that starts from 2012 onwards so that we would know where to split the dataset into test and train sets. **The splitting of the dataset into a random number of train and test sets cannot be done here because it can mess up the chronological order of the whole data set.** Once the splitting of data was complete, linear regression was used in our data analysis.

8. Linear Regression Analysis

The main technique used in this mini project is linear regression. In this case, the predictors would be one of the variables in from each category (Economic, Social, Health and Mortality) while the prediction of the response (target) would be life expectancy. The goal of linear regression is to obtain a line that best fits the data. This line is known as the least square regression line for which the total prediction errors are as small as possible. The minimised error is computed based on the vertical distance between each of the individual data points and the regression line [3]. Once linear regression was done on every predicting factor against life expectancy, multiple variable regression would be used to predict life expectancy with multiple independent variables (predicting factors). The purpose of this is to determine the overall fit of the model and the relative contribution of each of the predictors to the total variance. The evaluation metrics used to test the goodness of fit of our linear regression models were mean squared error (MSE) and explained variance (R^2). The R^2 values and MSE of each linear regression model, as well as the multiple variable regression model, would be computed and tabulated. By comparing the R^2 values and mean square errors (MSE) between different variable models, we can determine which is the most significant factor that contributes to the increase in life expectancy.

RESULTS AND DISCUSSION

Predictor	R2_Train	R2_Test	R2_diff	MSE_Train	MSE_Test	MSE_diff
HIV/AIDS	0.352814	0.172142	-0.180672	53.903642	46.708769	-7.194873
Adult Mortality	0.407447	0.417283	0.009836	49.353273	32.877602	-16.475671
Schooling	0.440938	0.555163	0.114225	46.563850	25.098246	-21.465604
Income	0.355862	0.584462	0.228600	53.649772	23.445187	-30.204584
Multi Variable	0.744031	0.761986	0.017955	21.319440	13.429018	-7.890423

Figure 2: Tables of Results (R^2& MSE of train and test sets, R^2 difference & MSE difference between train and test sets) **Note: R2_diff = R2_test-R2_train; MSE_diff = MSE_test-MSE_train**

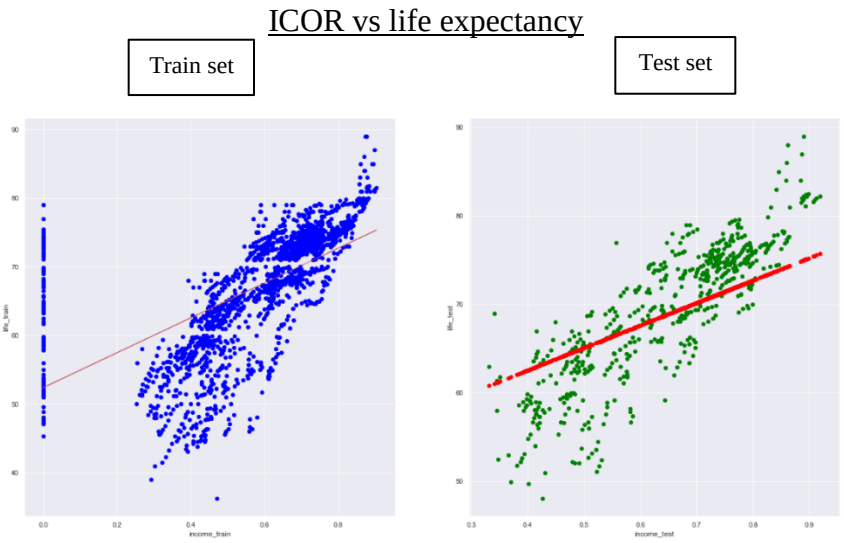

Figure 3: Linear Regression model for ICOR vs life expectancy for both train and test sets

From Figure 2, it can be observed that by executing multi variable regression whereby the regression line consists of more than 1 predictor, the R^2 score for both train and test set are the highest among the other single predictor regression models. Furthermore, the MSE for both train and test sets is the lowest among the other single predictor regression models as well. The reason for this is because more information which has relatively good correlation with life expectancy is being learnt and the corresponding trained regression line will be much better to predict life expectancy.

Out of the four factors, it seems that the regression model for Schooling and ICOR perform better than that of HIV/AIDS and Adult Mortality. The R^2 scores for test data set from Schooling and ICOR are higher than the R^2 scores for the test data set from HIV/AIDS and Adult Mortality. Furthermore, the improvement in R^2 test scores from Schooling and ICOR are better compared to the other factors as shown by the positive R^2 difference value. This means that a higher percentage of variability of life expectancy has been accounted for and the remaining percentage of the variability is still unaccounted for. For example, if the R^2 test score for income is 0.58, this means that 58% of the variability of the

data points in life expectancy has been accounted for and the remaining 42% of the variability is still unaccounted for. This explains why the higher the R^2 score, the better the predictive model.

As for MSE, the MSE of train and test data for the regression model of ICOR and Schooling is lower than that of HIV/AIDS and Adult Mortality. Moreover, the reduction of MSE in test data is also higher than that of HIV/AIDS and Adult Mortality as shown by the negative MSE difference value. This means that the average squared error between the predicted value and the actual value has been decreased hence giving better goodness of fit to the predictive model.

In general, the linear regression model for ICOR is the best model to predict life expectancy. Although its R^2 train score is lower than Schooling and its MSE for train set is higher than Schooling, the training model for test set has shown greater improvement compared to Schooling. Its R^2 test score (0.58) is higher than the R^2 test score for Schooling (0.55) and the MSE for the test set (23.4) is lower than that of Schooling (25.1). **This shows that the train data for ICOR from previous years can be used to predict future trends since the train model has a relatively good fit on the test set**. However, there are still many limitations to this regression model. In this mini project, it is assumed that the selected variable and life expectancy are linearly related to each other. In the real world, linear relationships between the dependent and independent variables rarely exist. Hence, a linear regression model is not a very accurate predictive model to predict real-life variables. Furthermore, linear regression is very sensitive to outliers. Outliers can often lead to inaccurate R^2 and MSE values [4]. Additionally, the raw data set is insufficient to generate an accurate predictive model as it only shows data set for the past 15 years. Besides, there are some mistakes in the raw data set. For instance, Canada, Finland, France were misclassified as developing countries. Another flaw in the raw data is that there are too many false claims made by some developing countries. For example, the ICOR for South Sudan is zero from the year 2000 to 2010 which is unlikely to be true. This can lead to an error when performing linear regression. If developed countries are classified as developing countries, this can lead to an overestimated R^2 value which may mislead us to think that the variable chosen is a good predictive model for life expectancy.

A room of improvement to our regression model is that the data cleaning of the raw data. As shown in Figure 3, there is a large number of outliers caused by the zero ICOR values declared by some developing countries. This not only cause the train set to behave very differently from the test set, it also lowers the R^2 value and increases the MSE of the train model of ICOR. By removing them, the overall goodness of fit of the regression model on the train and test set will improve, which lead to an increase in the effectiveness of the regression model for ICOR to predict future life expectancy.

CONCLUSION

In conclusion, a developing country should focus on economic factors such as ICOR to improve life expectancy. A developing country needs to utilise its resources productively so that the citizens can live longer than expected. As supported by an increase in R^2 score and decrease in MSE for test set for ICOR, it proves that it is feasible to use past data to predict future trends for the impact of ICOR on life expectancy.

References

[1] M. Roser, "Life Expectancy," October 2019. [Online]. Available: https://ourworldindata.org/life-expectancy. [Accessed 27 March 2020].

[2] K. Rajarshi, "Kaggle," 2018. [Online]. Available: https://www.kaggle.com/kumarajarshi/life-expectancy-who. [Accessed 15 March 2020].

[3] S. Swaminathan, "Linear Regression — Detailed View," Towards Data Science, 26 February 2018. [Online]. Available: https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86. [Accessed 29 March 2020].

[4] N. Kumar, "Advantages and Disadvantages of Linear Regression in Machine Learning," 9 May 2019. [Online]. Available: http://theprofessionalspoint.blogspot.com/2019/05/advantages-and-disadvantages-of-linear.html. [Accessed 2 April 2020].