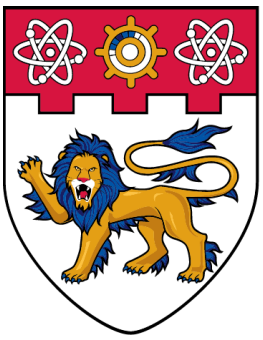
School of Chemical and Biomedical Engineering



NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

CB0494 Introduction to Data Science and Artificial Intelligence

Mini Project: Life Expectancy

Tutorial group: C24

Group 11

No.	Name	Matriculation No.
1	Lew Sun Chun	U1921129H
2	Lim Junwei Darien	U1921811F

Name	Jupyter Notebook (100%)	Report (100%)	Presentation (100%)	Overall (300%)
Lew Sun Chun	50%	50%	50%	150%
Lim Junwei Darien	50%	50%	50%	150%

Background

Life expectancy has always been used as an indicator of a country's standard of living, quality of healthcare services and lifestyle of the citizens. High life expectancy usually is inferred that the country is developed, has high standards of living and healthcare services, whereas low life expectancy would mean the opposite. Other factors could also play a part in affecting the life expectancy such as literacy of the population, budget spent on healthcare, etc.

In the dataset provided by World Health Organisation(WHO), the factors that were being studied in relation to life expectancy includes immunization factors (Polio, Hepatitis B, Diphtheria), mortality factors (adult mortality, infant deaths), economic factors (GDP, Total expenditure, Status of country), social factors (Schooling) and other health factors (BMI, HIV/AIDS). Various factors would be chosen to conduct our analysis and the reasons for choosing them will be explained later in the report.

Objectives

This report aims to determine the best predictors to predict life expectancy using the chosen variables(percentage expenditure, BMI, GDP, Schooling) from the dataset using appropriate machine learning tools. In addition, we would like to find out how anomaly detection and removal would affect the prediction. Lastly, we would also like to find out how other categorical factors can affect life expectancy.

Methodology & Results

1. Selecting variables to work with

The given dataset has many different kinds of data, and it is important to select the various factors we would like to work on to predict life expectancy. Of all the various factors, we would expect the following factors to be more useful in predicting the life expectancy of a country:

Status: Developing or Developed

We would expect the status of the country to be very influential to the life expectancy of a country. A country that is developed is more likely to have a higher life expectancy since its infrastructure (hospitals, emergency networks, etc.) would be much more advanced and people are much more likely to afford healthcare.

Percentage expenditure: Expenditure on health as a percentage of GDP per capita

The percentage expenditure on health should affect the life expectancy since we would expect the greater the spending on healthcare, the greater the life expectancy of a country. However, for countries with low GDP per capita, the percentage expenditure would not really affect life expectancy since the actual amount spent on healthcare is low, to begin with.

GDP: GDP per capita

We would expect countries with higher GDP to have a greater life expectancy as the residents would be wealthier and more likely to afford healthcare.

Schooling: Number of years of schooling

We would expect that countries with a greater number of years of schooling would likely have a greater life expectancy since the people would be more educated and mindful about their health. The more educated would also tend to have a higher income and thus would be more likely to afford healthcare as well.

BMI: The BMI of a person

BMI is a major health indicator that measures a person's height to weight ratio. There is a range of '18-25' where a person would be considered healthy. We would expect countries with healthy levels of BMI to have a greater life expectancy.

Since 'Status' is the only variable out of all the variables we are interested in that has an 'object' datatype, while the rest being numerical data, we would like to do our analysis of numerical and categorical data separately. As such, we would extract the numerical and categorical variables separately.

2. Cleaning and Sorting of data

After extracting all the numerical variables, there is null data in the variables that we have chosen, and the number of entries for each variable is different. These null entries are unwanted and should be removed to prevent errors when doing our analysis. Thus, .dropna() syntax is applied to the data set to remove entries with null data. After the .dropna() syntax, all the variables have identical numbers of non-null entries at 2458. This data frame can then be used for further analysis.

3. Correlation of life expectancy with other factors

Visualization

Seaborn pair plot syntax is used to visualize the correlation of how each variable go against life expectancy:

sb.pairplot(cleaneddata)

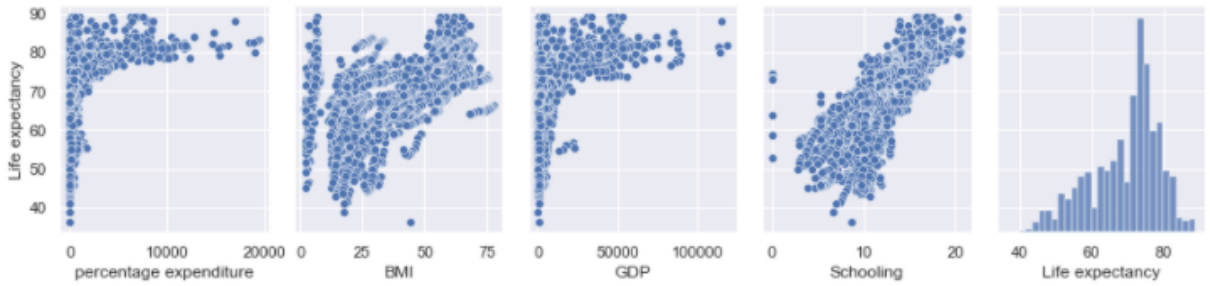


Figure 1. The extracted row of pair plot of life expectancy against the other factors

From Figure 1, it can be seen that percentage expenditure and GDP have a low correlation with life expectancy, while BMI and schooling would have a much stronger correlation with life expectancy.

Confusion Matrix

The visualization can be comprehended in terms of the correlation coefficient.

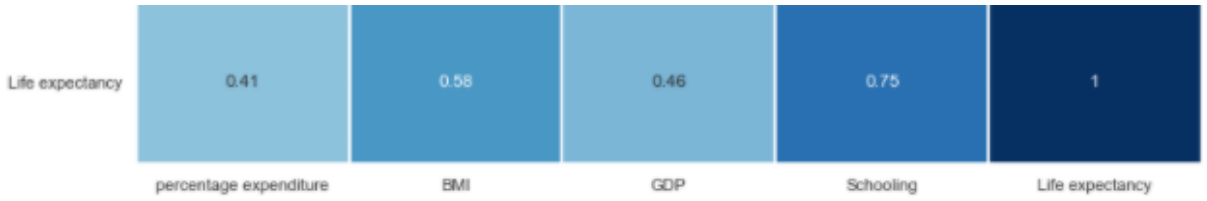


Figure 2. Extracted confusion matrix

In Figure 2, schooling and BMI have a higher correlation coefficient than percentage expenditure and GDP. Schooling has the highest correlation coefficient with life expectancy. We defined the correlation value of 0.5 and below to be low correlation against life expectancy. Thus, according to the heatmap, Percentage expenditure and GDP should be excluded from our data analysis, as they have a low correlation with life expectancy and we would not expect these factors to be able to predict life expectancy well.

4. Anomaly Detection and Removal

Isolation Forest would be used to do anomaly detection instead of Local Outlier Factor (LOF), as we focused on the detection and removal of global outliers instead of local outliers. It is also computationally faster than LOF. The isolation forest module from **pyod** library would be used as it would enable us to do a multivariate model in the later stage. The module can be installed by typing "pip install pyod" in the command prompt.

Uni-variate detection

BMI and Schooling against life expectancy are used in the analysis, and the results for the correlation coefficient for BMI before and after anomaly detection is shown in Figure 3 below.

```
Graph 1: Scatterplot of Life expectancy against BMI
Graph 2: Visualising anomalies of  Life expectancy against BMI
Graph 3: Scatterplot of  Life expectancy against BMI after anomaly removal
```

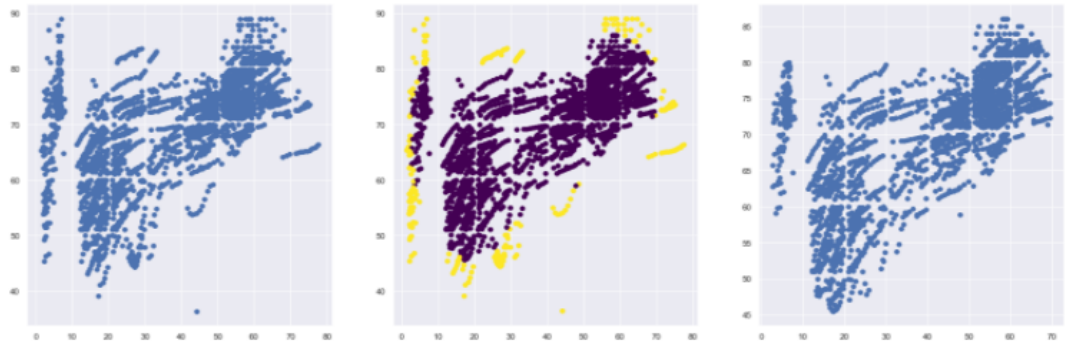


Figure 3. Bivariate plot of BMI against life expectancy before and after anomaly removal

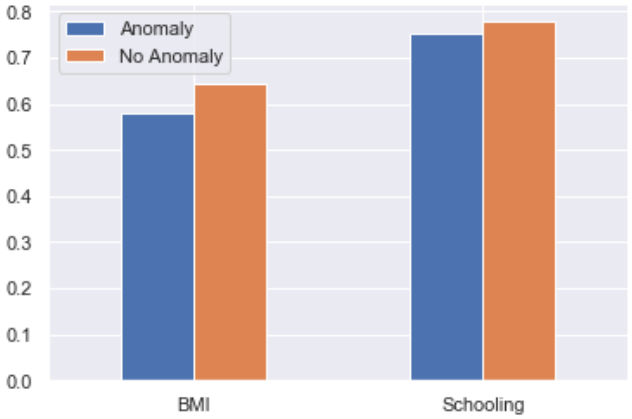


Figure 4. Bar chart of correlation coefficient of BMI and Schooling.

From Figure 4, we have determined that the correlation coefficient does improve the coefficient by a little. However, the results vary differently for the different variables. The correlation coefficient of BMI improved much greater than Schooling. To confirm the reliability of it, we would like to do a linear regression with and without anomaly.

5. Linear Regression

After obtaining the correlation coefficients of how each variable correlates to the Life Expectancy data as well as how anomaly, we would like to predict the life expectancy with each variable, as well as with the multivariate model. Furthermore, to confirm whether the decision to drop the "percentage expenditure" and "GDP" variables was the right choice, we are doing multivariate linear regression with and without these two variables to compare the scores.

Predictor	Train	Test	Train_NA	Test_NA
BMI	0.359395	0.249732	0.443262	0.303522
Schooling	0.601982	0.460856	0.633647	0.510231
Multivariate_2	0.631607	0.500902	0.671769	0.550389
Multivariate_4	0.641891	0.528135	0.657595	0.488001

Figure 5. The goodness of fit (R^2) of the various predictors

In Figure 5, “Multivariate_2” explains the linear regression based on BMI & Schooling against life expectancy whereas “Multivariate_4” BMI, Schooling, percentage expenditure & GDP against life expectancy.

We can identify that schooling is the best variable for predicting Life Expectancy as it has the highest R^2 value compared to BMI. We can also conclude that using a multivariate model helps to predict the life expectancy slightly better, although the improvement over the best performing variable (Schooling) might not be very significant. We also found that the two variables we have dropped were indeed justified since the R^2 values of multivariate_2 and Multivariate_4 are almost the same, suggesting that the variables 'percentage expenditure' and 'GDP' are not good at predicting Life expectancy. Lastly, we found that removing the anomaly denoted as “Train_NA” and “Test_NA” would not provide us with a better result since the R^2 score with and without anomaly are similar. Therefore, it is not recommended to remove the anomaly when doing the regression especially since these anomalies are also part of the dataset and should be kept in.

7. Bivariate comparison using categorical variables

We would like to explore further how we can use other categorical variables to predict life expectancy. As such, we have decided to encode each country with its respective continents to show the life expectancies of each continent. To do so, we would require the pycountry module and it could be installed using “pip install pycountry-convert”.

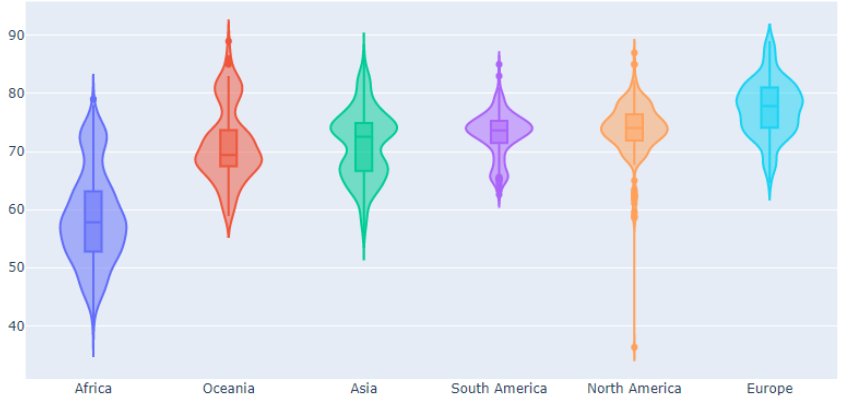


Figure 6. Violin plots on the 6 continents life expectancy

From Figure 6, it can be seen that life expectancy is affected by which continent the country is in. Africa has the lowest life expectancy at around 55 years while the rest of the continents have much higher life expectancies at 70-75 years, with Europe having the highest life expectancy with its median age of 78 years.

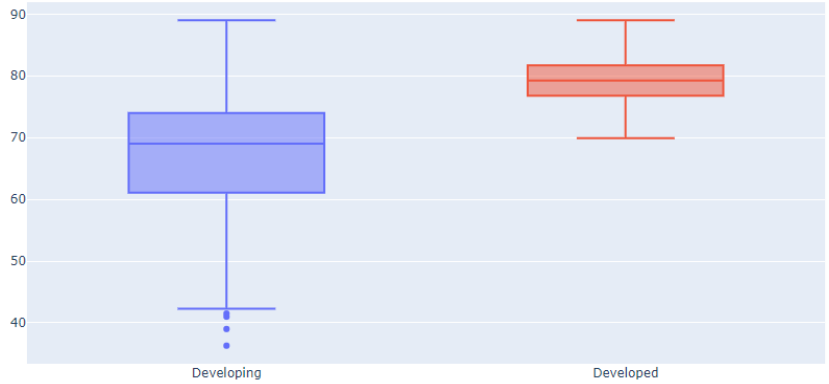


Figure 7. Box plot of developing and developed countries life expectancy

From Figure 7, the status of a country has an impact on life expectancy since the 'Developed' status as a whole is higher than that of 'Developing'. Countries that are 'Developed' do have a greater life expectancy than countries that are 'Developing' in general.

Conclusion

On the whole, schooling has been identified as the best predictors among the other variables based on linear regression. In addition, removing anomalies from the data set would not significantly affect the prediction. One of the reasons why schooling is the best predictor could be that, if the population is more educated, they are able to comprehend the consequences and benefits of a decision. For eg, choosing healthier food vs fast food. They would also be more likely to land a high-paying job, which would make them live comfortably and be able to afford basic needs and healthcare.

We have also performed multivariate linear regression and concluded that the multivariate model does improve the accuracy of the prediction, but not by much. Based on the categorical variable analysis, we found that country status is a major factor in determining life expectancy, while continents can generally predict the range of life expectancy. Therefore, further work can be done on including the categorical data identified in the report (Status and Continents) with the numerical data (BMI & schooling) to improve the prediction of life expectancy using other machine learning tools that have functions to achieve this. One way to do this is to encode the categorical data and perform multivariate linear regression with both numeric and categorical data to predict life expectancy.