Project 2: Life Expectancy Prediction

Kesav Adithya Venkidusamy

Bellevue University - Master of Science in Data Science

DSC680-T301 Applied Data Science (2231-1)

Professor Catherine Williams

10/16/2022

Table of Contents

Business Problem 3

Background/History 3

Data Explanation 3

Data Dictionary 4

Data Preparation 5

Data Visualization 6

Methods 12

Analysis 12

Modeling Analysis 12

Feature Analysis 14

Conclusion 16

Assumptions 17

Limitations 17

Challenges 17

Future Uses/Additional Applications 17

Recommendations 17

Implementation Plan 17

Ethical Assessment 18

References 18

Business Problem

Everything has an expiration date; humans are no exception either. The term “life expectancy” refers to the number of years a person can expect to live. By definition, life expectancy is based on an estimate of the average age that members of a particular population group will be when they die. In this project, I aim to explore the parameters affecting the life span of individuals living in distinct countries and learn how the life span can be estimated with the help of machine learning models. I will also focus on exploring the parameters that greatly impact the lifespan of an individual

Background/History

We’re in an unprecedented era where humans are living longer with increased access to modern science and healthcare. It’s no secret, though, that life expectancy varies widely across the globe. Life expectancy depends on several factors, the two most important being gender and birth year. Generally, females have a slightly higher life expectancy than males due to biological differences. Other factors that influence life expectancy include:

- Race and ethnicity
- Family medical history
- Risky lifestyles

Data Explanation

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 of all countries. The datasets are made available to the public for health data analysis.

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

The data set related to life expectancy, and health factors for 193 countries have been collected from the same WHO data repository website, and its corresponding economic data was collected from the 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 the health sector resulting in an improvement in human mortality rates, especially in developing nations in comparison to the past 30 years. Therefore, in this project, we have considered data from the years 2000-2015 for 193 countries for analysis. The data was collected from WHO and United Nations websites with the help of Deeksha Russell and Duan Wang.

Data Dictionary

Characteristics

Data Set Characteristics	Multivariate
Attribute Characteristics	Categorical, Integer
Associated Tasks	Classification
Number of Instances	2938
Number of Attributes	22
Missing Values	Yes
Area	Health

Attributes information

Feature Name	**Feature Description**	**Feature Type**
Country	Country Observed	Discrete
Year	Year Observed	Continuous
Status	Status of the country; Developed or Developing Status	Discrete
Life expectancy	Life expectancy in age	Target
Adult Mortality	Adult Mortality Rates on both sexes (probability of dying between 15-60 years/1000 population).	Continuous
Infant deaths	Number of Infant Deaths per 1000 population	Continuous
Alcohol	Alcohol recorded per capita (15+) consumption (in liters of pure alcohol).	Continuous
Percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%).	Continuous
Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)	Continuous
Measles	Number of reported Measles cases per 1000 population	Continuous
BMI	Average Body Mass Index of the entire population	Continuous
Under-five-deaths	Number of under-five deaths per 1000 population	Continuous
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)	Continuous
Total expenditure	General government expenditure on health as a percentage of total government expenditure (%)	Continuous
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)	Continuous

HIV/AIDS	Deaths per 1 000 live births of HIV/AIDS (0-4 years)	Continuous
GDP	Gross Domestic Product per capita (in USD)	Continuous
Population	The population of the country	Continuous
thinness 1-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)	Continuous
thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9(%)	Continuous
Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)	Continuous
Schooling	Number of years of Schooling(years)	Continuous

Data Preparation

The problem statement of this project is to identify the dataset feature(s) which are mostly related to or affecting the life expectancy of a person. The data set contains approximately 3000 entries. Given the limited size of the data set, the model should only be expected to provide a modest improvement in the identification of attrition vs a random allocation of the probability of attrition. The dataset consists of 22 features of which 20 are numerical and two are categorical with "life expectancy" being the target.

The target variable "life expectancy" is a continuous variable with values ranging between 36.3 years and 89 years with a maximum number of values lying at 72.1 years. The mean for this variable lies at 69.22 years. The details are shown in figure 2.

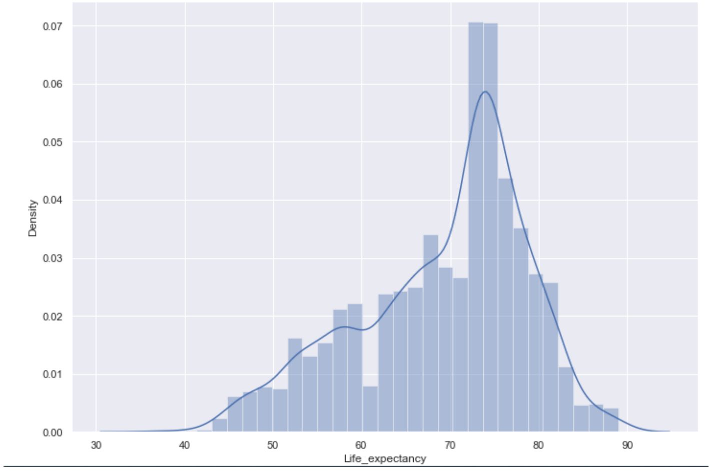


Figure 1. Target Variable "Life Expectancy" Analysis

As part of the EDA process, a null check has been performed initially on all the variables and found a null value present for the below features. The null value for all these variables is replaced with the forwarding fill method.

- Hepatitis B
- Alcohol
- Adult Mortality
- Polio
- Total Expenditure
- GDP
- Population
- Schooling
- Income Composition of resources
- Life Expectancy
- Diphtheria
- Thinness 5 to 9 years
- Thinness 1 to 19 years
- BMI

Then, a duplicate check has been performed, and found no duplicate in the dataset. None of the features has been removed from the dataset as I believe each feature correlates with the target variable "life expectancy".

Data Visualization

As mentioned before, the dataset contains 20 numerical features and 2 categorical features, and 'life expectancy' is the target variable. Below are some of the charts plotted to analyze these variables.

Histogram

The histogram chart is used to identify the distribution of numerical features present in the dataset. All the numerical features are considered for the histogram as shown in figure 2. The observation of each histogram is shown in figure 3.

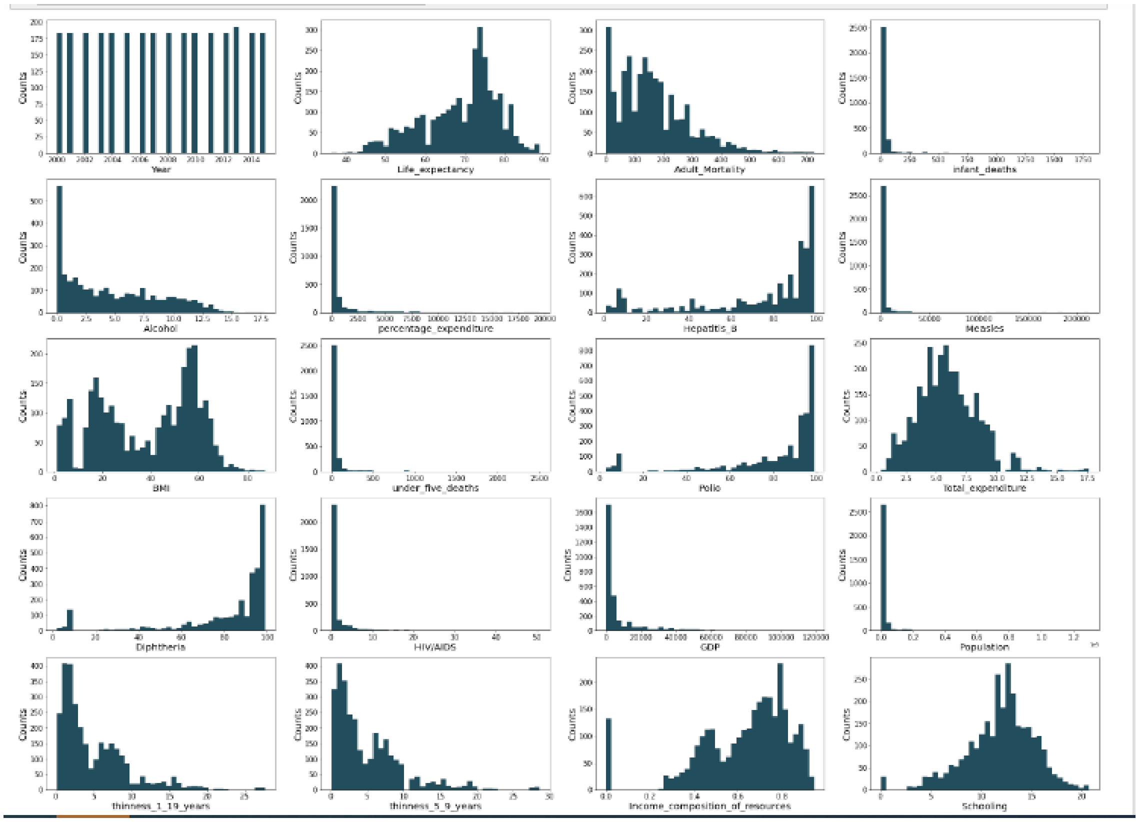
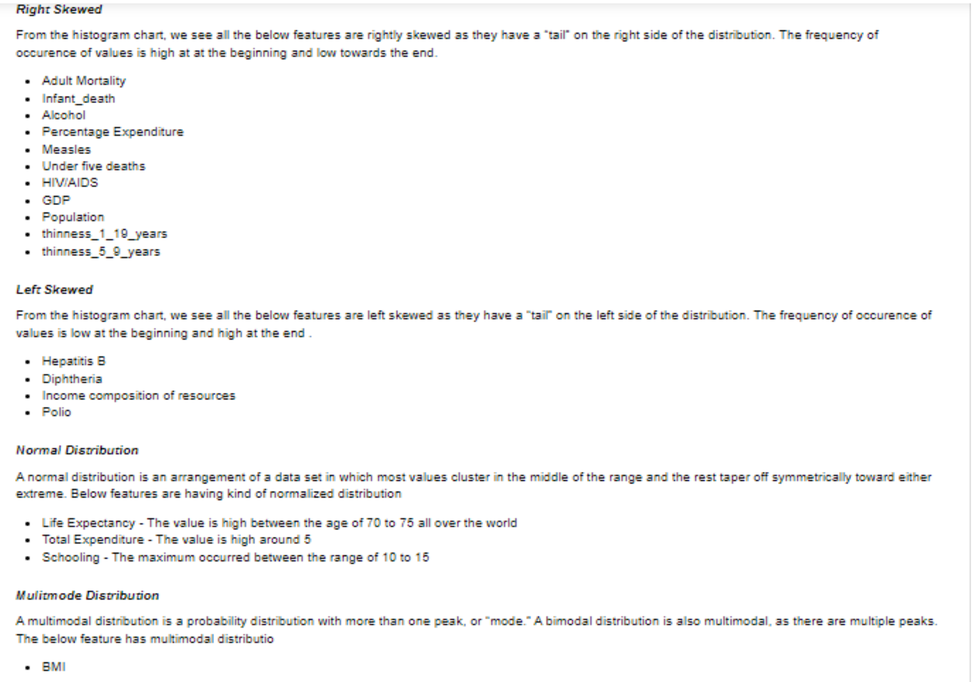


Figure 2 Histogram for numerical variables

Observation



Right Skewed

From the histogram chart, we see all the below features are rightly skewed as they have a "tail" on the right side of the distribution. The frequency of occurence of values is high at at the beginning and low towards the end.

- Adult Mortality
- Infant_death
- Alcohol
- Percentage Expenditure
- Measles
- Under five deaths
- HIV/AIDS
- GDP
- Population
- thinness_1_19_years
- thinness_5_9_years

Left Skewed

From the histogram chart, we see all the below features are left skewed as they have a "tail" on the left side of the distribution. The frequency of occurence of values is low at the beginning and high at the end .

- Hepatitis B
- Diphtheria
- Income composition of resources
- Polio

Normal Distribution

A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. Below features are having kind of normalized distribution

- Life Expectancy - The value is high between the age of 70 to 75 all over the world
- Total Expenditure - The value is high around 5
- Schooling - The maximum occurred between the range of 10 to 15

Mulitmode Distribution

A multimodal distribution is a probability distribution with more than one peak, or "mode." A bimodal distribution is also multimodal, as there are multiple peaks. The below feature has multimodal distributio

- BMI

Figure 3: Observations from histogram charts

Violin Graph

A violin graph has been plotted for the categorical feature 'Status' and our target variable 'Life Expectance' to find the correlation between them. The details are shown in figure 4. From the graph, I could see developing countries have a low life expectancy and developed countries have high life expectancy all over the world.

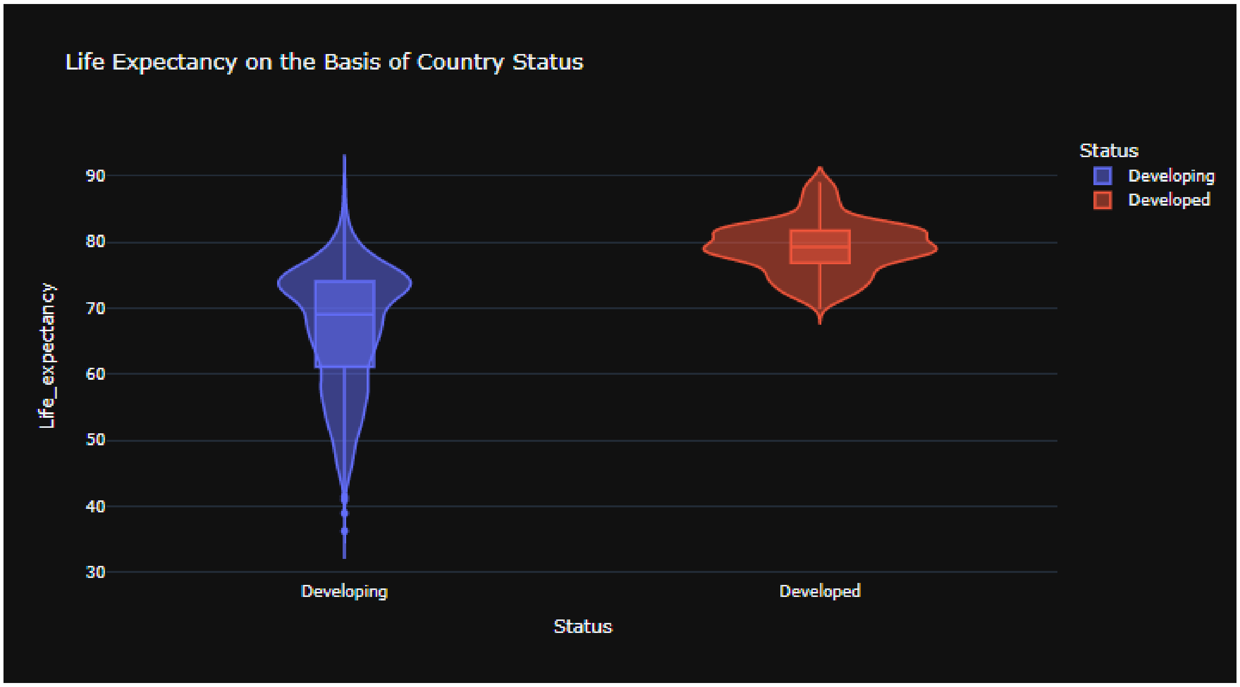


Figure 4 Violin chart for Categorical Variable Status

Map Chart

A map chart as shown in figure 5 has been plotted between the categorical variable "country" and the target variable "life expectancy" to show the density of life expectancy across countries. Looking at the chart, we see that life expectancy is high for developed countries compared to developing countries. In addition, we could also notice that life expectancy increases over the years across the countries.

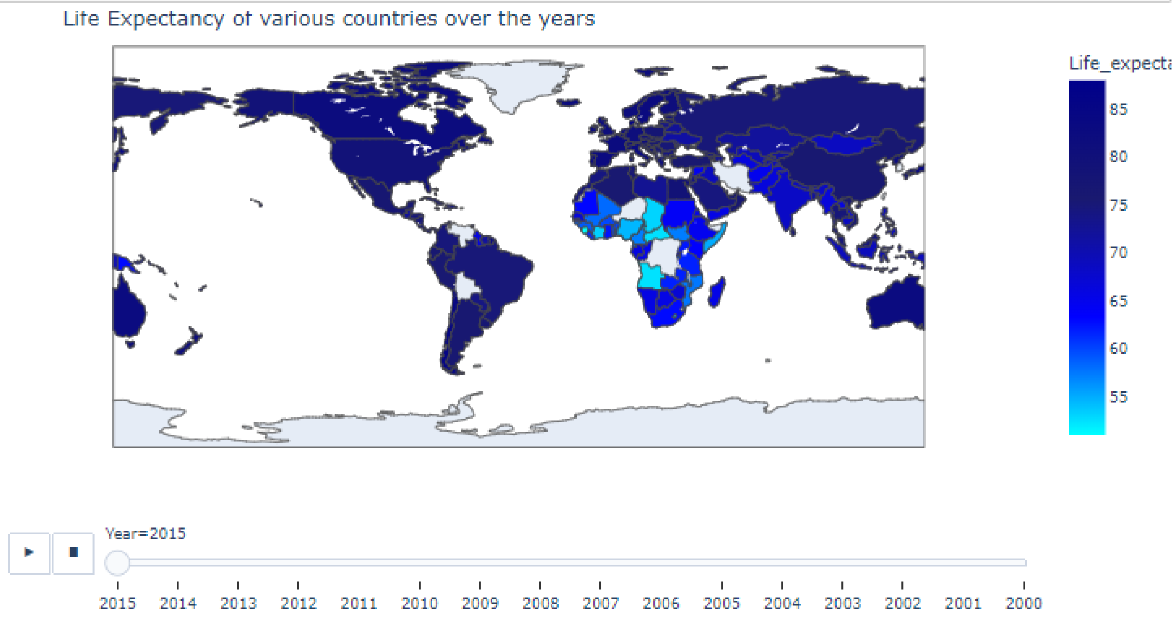


Figure 5: Map Chart showing the density of life expectancy

Heat Map

A heat map has been created to understand the correlation between various features present in the dataset. The details are shown in figure 6. The observations of the heat map are shown in figure 7.

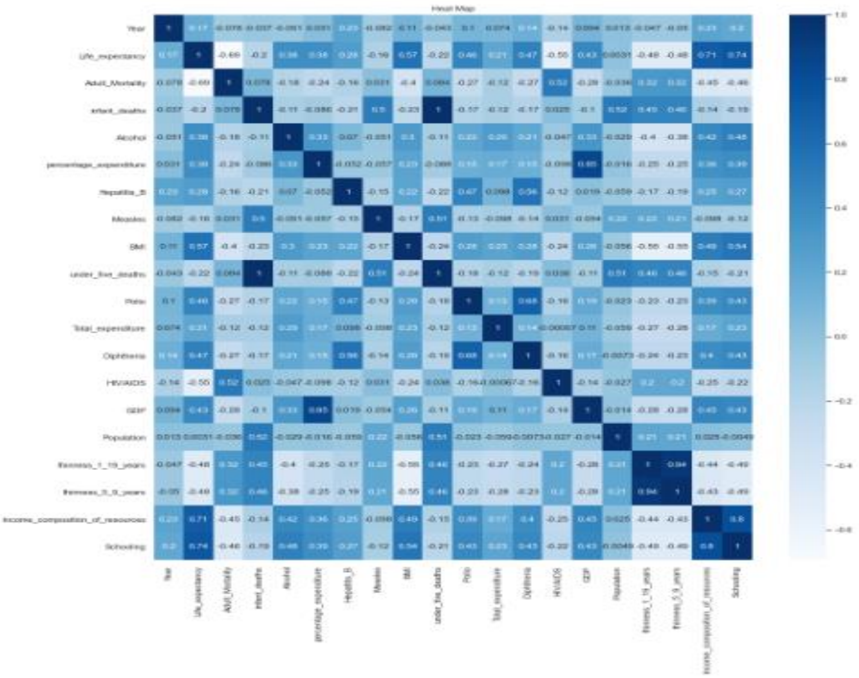
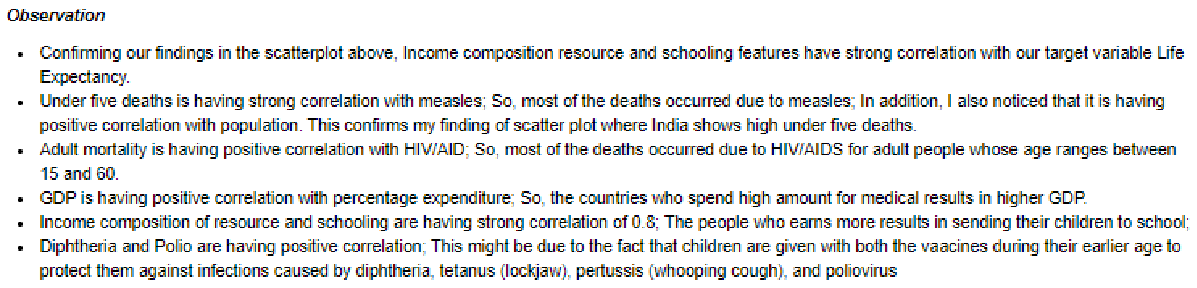


Figure 6: Heat Map to show the correlation



Observation

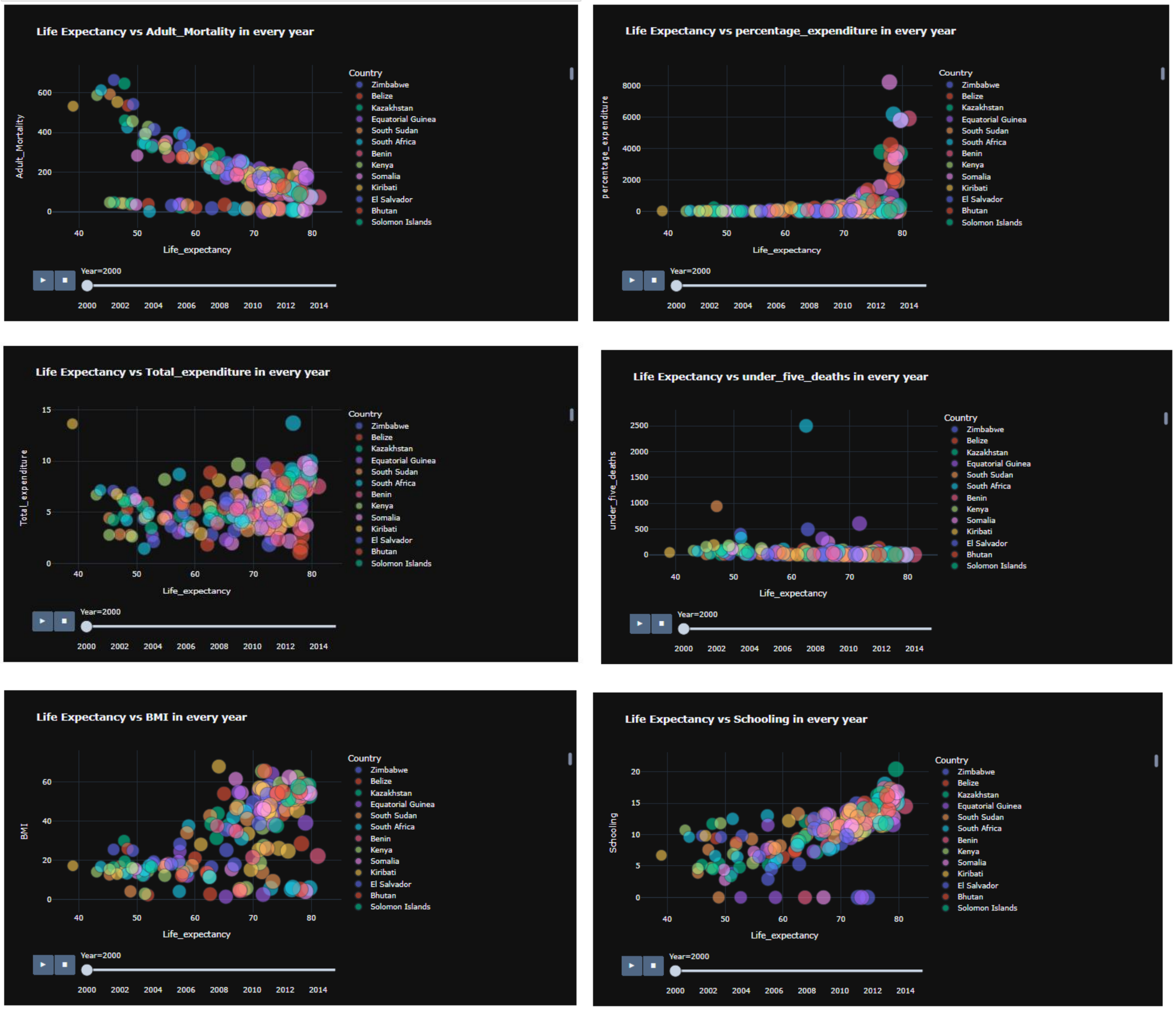
- Confirming our findings in the scatterplot above, Income composition resource and schooling features have strong correlation with our target variable Life Expectancy.
- Under five deaths is having strong correlation with measles; So, most of the deaths occurred due to measles; In addition, I also noticed that it is having positive correlation with population. This confirms my finding of scatter plot where India shows high under five deaths.
- Adult mortality is having positive correlation with HIV/AID; So, most of the deaths occurred due to HIV/AIDS for adult people whose age ranges between 15 and 60.
- GDP is having positive correlation with percentage expenditure; So, the countries who spend high amount for medical results in higher GDP.
- Income composition of resource and schooling are having strong correlation of 0.8; The people who earns more results in sending their children to school;
- Diphtheria and Polio are having positive correlation; This might be due to the fact that children are given with both the vaacines during their earlier age to protect them against infections caused by diphtheria, tetanus (lockjaw), pertussis (whooping cough), and poliovirus

Figure 7 Heat Map Correlations and observation

Target Variable Analysis

The following scatter plots are created to analyze our target variable “life expectancy”. The details are shown in figure 8.

- **Life Expectancy vs Adult Mortality over years for each country:** We see the adult mortality rate is low for developed countries compared to developing countries. Due to the development of medicine in developing countries, the adult mortality rate has increased over the period. However, life expectancy has increased and adult mortality has decreased over years in the countries
- **Life Expectancy vs Percentage Expenditure over years for each country:** The medical spending contribution by developing countries is low resulting in low life expectancy compared to the spending by developed countries which results in high life expectancy. I could see the same pattern for all the years.
- **Life Expectancy vs Total Expenditure in every year and country:** Most of countries the total expenditure lies below 10 and only limited countries are having expenditures greater than 10. Most of the values for life expectancy lie in the range of 40 and 90. The United States is having high total expenditure and life expectancy. Life expectancy has increased over years for all the countries
- **Life Expectancy vs under_five_deaths in every year and country:** Most countries are having values of less than 500. I also noticed that India is having high under 5 deaths. This might be due to population as India is the 2nd largest population country in the world. However, over the year under-five deaths have been reducing in India.
- **Life Expectance vs BMI in every year and country:** Over the years, the BMI across the countries has increased which results in high life expectancy. This is because people are more cautious about their health nowadays compared to older years.
- **Life Expectance vs Schooling in every year and country:** Schooling also plays a major role in improving life expectancy. Schooling gradually increases over the years which results in high life expectancy.


Life Expectancy vs Adult_Mortality in every year

Life Expectancy vs percentage_expenditure in every year

Life Expectancy vs Total_expenditure in every year

Life Expectancy vs under_five_deaths in every year

Life Expectancy vs BMI in every year

Life Expectancy vs Schooling in every year

Figure 8: Scatter Charts

Regression Plot

The regression plot creates a regression line between 2 parameters and helps to visualize their linear relationships. The regression plots for the features (BMI, Income composition of resources, adult mortality, GDP, and infant deaths) have been created to understand the relationship with the target variable “life expectancy”. The details are shown in figure 9. I could see life expectancy increase for an increase in BMI, GDP, and income composition of resources and a decrease for an increase in adult mortality and infant deaths as expected.

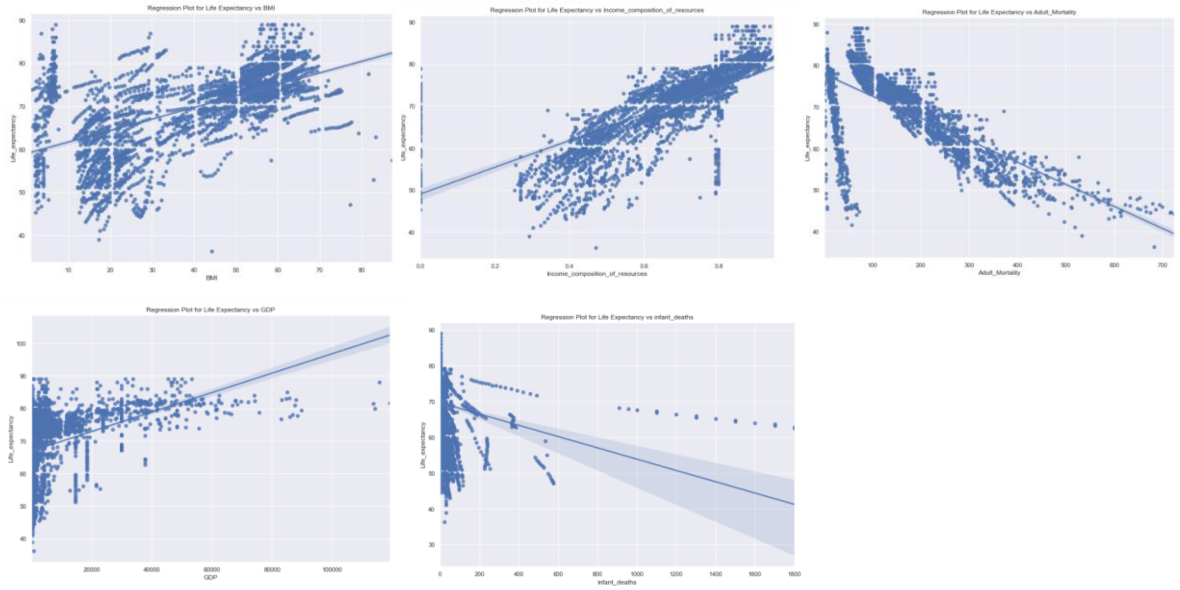


Figure 9: Regression Plot

Methods

Since the target variable "Life Expectancy" is a continuous variable, the linear regression model has been executed on the dataset to determine the features that are mostly related or correlated to our target. Linear regression is commonly used for predictive analysis and modeling. Simple Linear Regression is a type of regression algorithm that models the relationship between a dependent variable and a single independent variable. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line.

Feature selection is one of the most important parts of any ML model. We always want to select those features which have the maximum effect on our final output. I will perform the following operations for feature selection:

- Pearson Correlation Matrix
- Lasso Method
- OLS Regression

Analysis

Modeling Analysis

Initially, I ran the linear regression on the data without applying any normalization technique. Then, I applied the MinMaxScaler normalization technique which scales the minimum and maximum values to be 0 and 1 respectively and calculated the score. The score and mean square errors for these 2 approaches are given in table 1.

Methods	R2 Score	RMS Error
Without Normalization	83.54%	14.26
With MinMaxScaler Normalization	83.54%	0.00513

Table 1: Linear Regression Scores for various techniques

In linear regression, accuracy cannot be measured due to continuous target variables. If we try to evaluate the accuracy, we will end up overfitting the model. So, the performance of the linear regression model is evaluated using the below metrics.

R2 (R-Squared): R-squared (R2) is a measure of how close the data points are to the fitted line. It is also known as the coefficient of determination. Below is the acceptable range for R2 values.

- R-squared value < 0.3 this value is generally considered a None or Very weak effect size
- R-squared value 0.3 < r < 0.5 this value is generally considered a weak or low effect size
- R-squared value 0.5 < r < 0.7 this value is generally considered a Moderate effect size
- R-squared value r > 0.7 this value is generally considered a strong effect size

Root Mean Square Error (RMSE): Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are

Among the 2 linear regression models performed on the input data, the one with the MinMaxScaler normalization technique applied gives a better result with a low root mean square error value. The R2 value is the same for both techniques. Figure 10 represents a regression plot for the predicted values with and without the normalization technique.

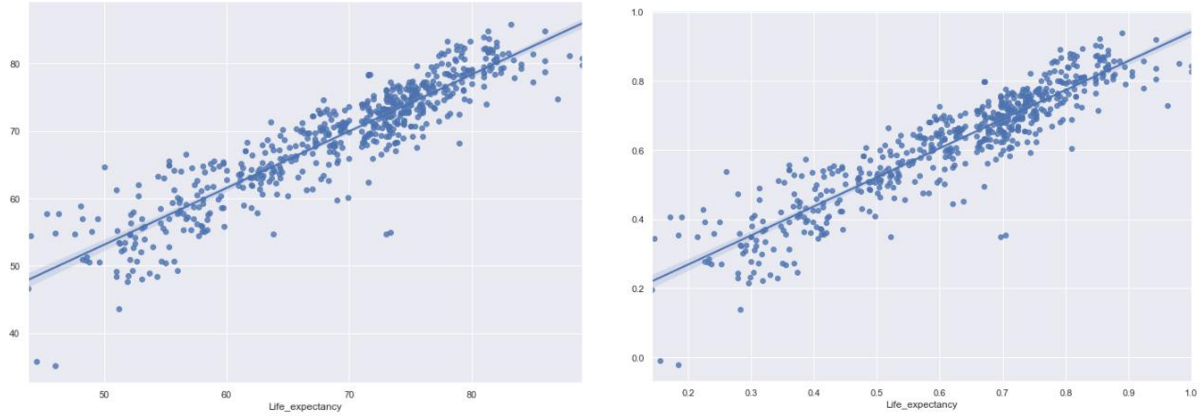


Figure 10: Regression Plot for predicted values without normalization and with normalization

Feature Analysis

Using various methods as mentioned in the method section, I tried to find the best features from the dataset and the following are the best features in the dataset which show a higher impact on the target variable “life expectancy” compared to other features present in the dataset.

Test	Features
Pearson's correlation matrix - Feature correlation to target variable "life expectancy"	**<u>Positive Correlation</u>:** • Schooling • Income composition of resources • BMI **<u>Negative Correlation:</u>** • Adult Mortality • HIV/AIDS • Thinness_1_9_years
Lasso Regression	BMI Polio Diphtheria Income composition of resources Schooling
Ordinary Least Square (OLS)	Year Hepatitis_B GDP Population thinness_1_19_years thinness_5_9_years

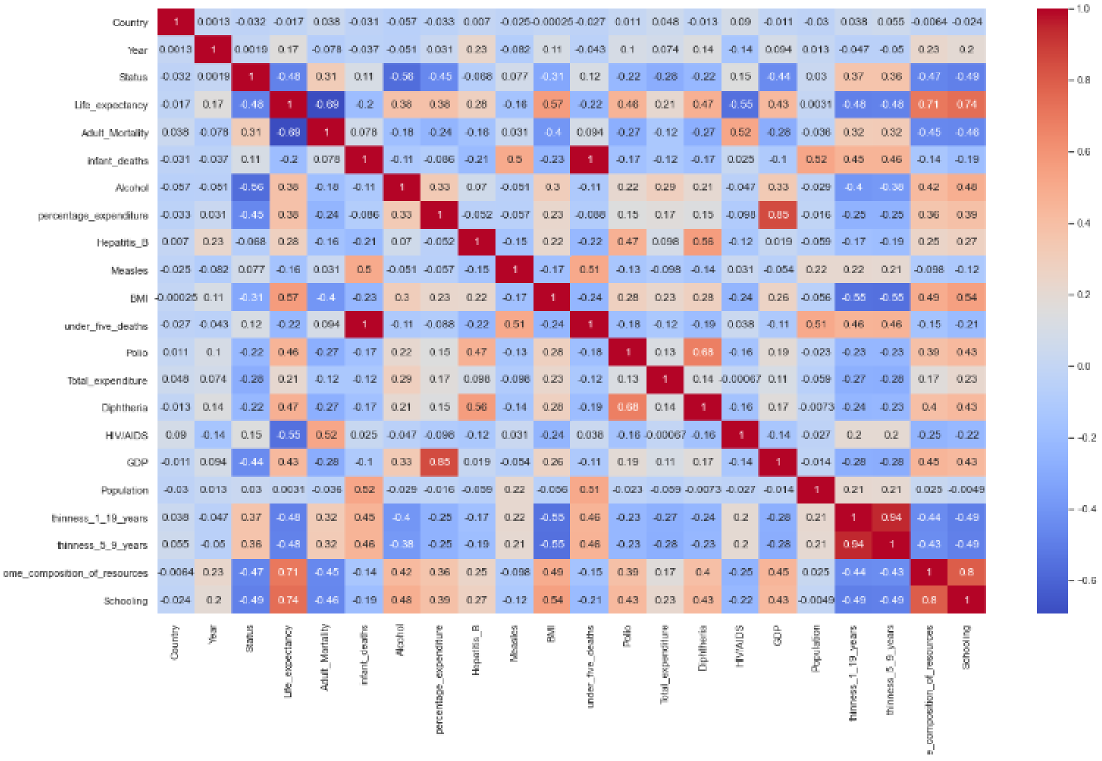
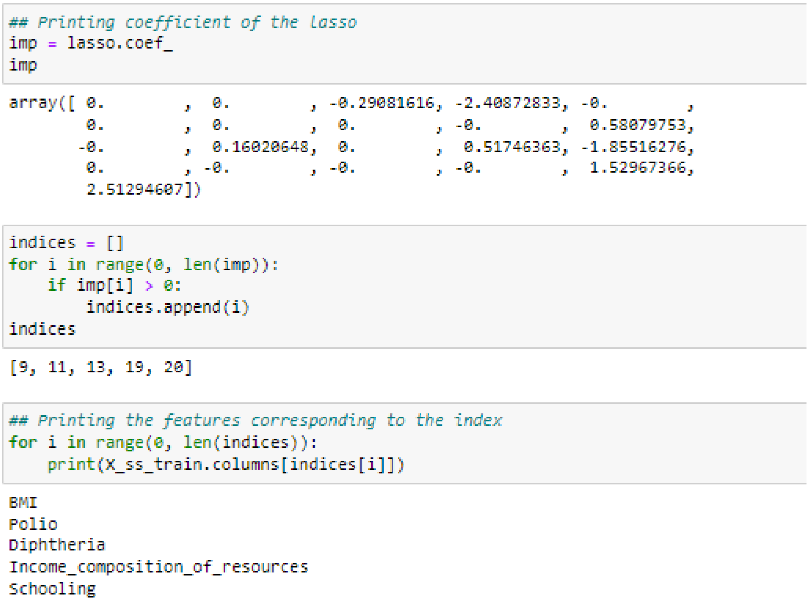


Figure 11: Pearson Correlation Matrix



```
## Printing coefficient of the lasso
imp = lasso.coef_
imp
```

```
array([ 0.        ,  0.        , -0.29081616, -2.40872833, -0.        ,
        0.        ,  0.        ,  0.        , -0.        ,  0.58079753,
       -0.        ,  0.16020648,  0.        ,  0.51746363, -1.85516276,
        0.        , -0.        , -0.        , -0.        ,  1.52967366,
        2.51294607])
```

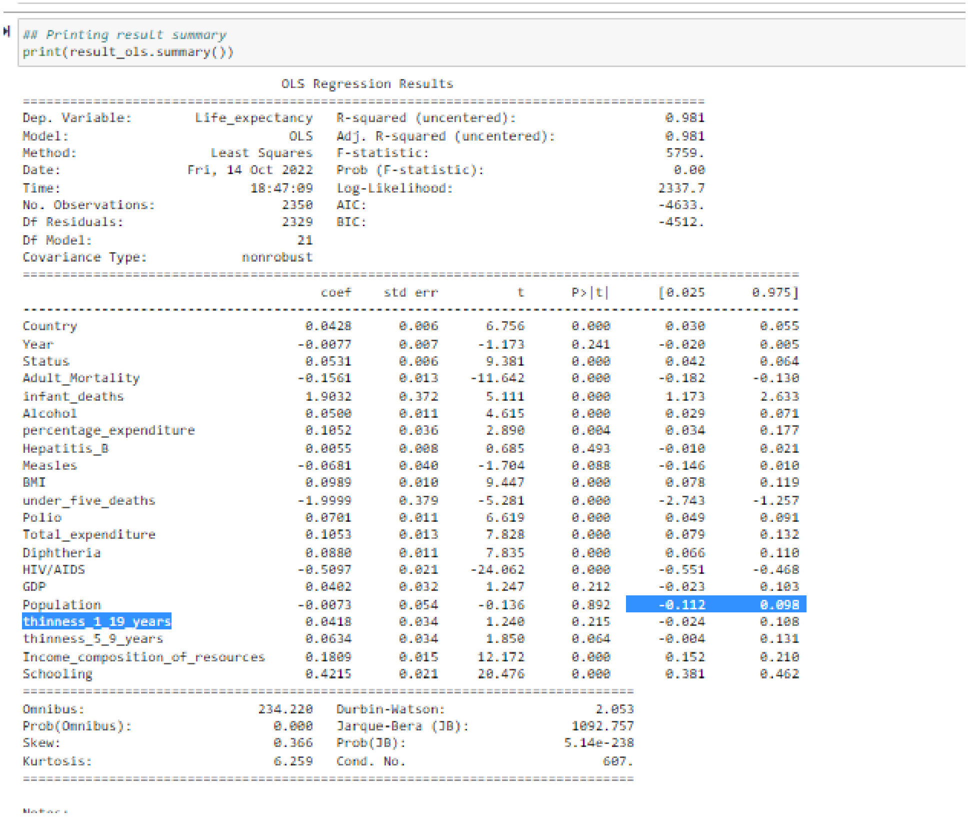
```
indices = []
for i in range(0, len(imp)):
    if imp[i] > 0:
        indices.append(i)
indices
```

```
[9, 11, 13, 19, 20]
```

```
## Printing the features corresponding to the index
for i in range(0, len(indices)):
    print(X_ss_train.columns[indices[i]])
```

```
BMI
Polio
Diphtheria
Income_composition_of_resources
Schooling
```

Figure 12: Best features using Lasso Regression



```
## Printing result summary
print(result_ols.summary())
```

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:        Life_expectancy   R-squared (uncentered):                   0.981
Model:                            OLS   Adj. R-squared (uncentered):              0.981
Method:                 Least Squares   F-statistic:                              5759.
Date:                Fri, 14 Oct 2022   Prob (F-statistic):                        0.00
Time:                        18:47:09   Log-Likelihood:                          2337.7
No. Observations:                2350   AIC:                                     -4633.
Df Residuals:                    2329   BIC:                                     -4512.
Df Model:                          21
Covariance Type:            nonrobust
===================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Country                            0.0428      0.006      6.756      0.000       0.030       0.055
Year                              -0.0077      0.007     -1.173      0.241      -0.020       0.005
Status                             0.0531      0.006      9.381      0.000       0.042       0.064
Adult_Mortality                   -0.1561      0.013    -11.642      0.000      -0.182      -0.130
infant_deaths                      1.9032      0.372      5.111      0.000       1.173       2.633
Alcohol                            0.0500      0.011      4.615      0.000       0.029       0.071
percentage_expenditure             0.1052      0.036      2.890      0.004       0.034       0.177
Hepatitis_B                        0.0055      0.008      0.685      0.493      -0.010       0.021
Measles                           -0.0681      0.040     -1.704      0.088      -0.146       0.010
BMI                                0.0989      0.010      9.447      0.000       0.078       0.119
under_five_deaths                 -1.9999      0.379     -5.281      0.000      -2.743      -1.257
Polio                              0.0701      0.011      6.619      0.000       0.049       0.091
Total_expenditure                  0.1053      0.013      7.828      0.000       0.079       0.132
Diphtheria                         0.0880      0.011      7.835      0.000       0.066       0.110
HIV/AIDS                          -0.5097      0.021    -24.062      0.000      -0.551      -0.468
GDP                                0.0402      0.032      1.247      0.212      -0.023       0.103
Population                        -0.0073      0.054     -0.136      0.892      -0.112       0.098
thinness_1_19_years                0.0418      0.034      1.240      0.215      -0.024       0.108
thinness_5_9_years                 0.0634      0.034      1.850      0.064      -0.004       0.131
Income_composition_of_resources    0.1809      0.015     12.172      0.000       0.152       0.210
Schooling                          0.4215      0.021     20.476      0.000       0.381       0.462
==============================================================================
Omnibus:                      234.220   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1092.757
Skew:                           0.366   Prob(JB):                    5.14e-238
Kurtosis:                       6.259   Cond. No.                         607.
==============================================================================

Notes:
```

Figure 13: OLS Feature Extraction Summary

Conclusion

The linear regression applied to the normalized data gave the best result and is used to predict life expectancy. Though the R2 value is the same for the linear regression applied to denormalized and normalized data, RMSE has been substantially reduced for the regression applied to normalized data.

Among various methods used to find the best features in the dataset, below are some of the top features having a high impact on the target variable “life expectancy”.

- BMI
- Income composition of resource
- Schooling
- Diphtheria
- Polio
- Population
- GDP
- Hepatitis-B

Assumptions

The dataset contains null values across many features. I have substituted those null values with the ffill method (forward fill) available in the fillna method of pandas. I assume the values substituted would be correct as this method propagates the last valid observation forward.

Among various techniques available for feature extraction, I have considered only Pearson correlation, Lasso, and OLS methods as I assume these methods would give the best and expected results.

Limitations

The dataset considered for this prediction analysis contains only limited rows (~3000 rows). So, the prediction and feature extraction performed during this analysis is based on this limited dataset. If this dataset doesn't accurately reflect the original real-world data, modeling efforts cannot generate any useful insights.

Challenges

There are a couple of challenges I faced during the data preparation step for model building. Identifying the correct features that contribute to the target, planning on how to handle the null values present across various features of the dataset, and deciding on which feature extraction model to use to name a few.

Future Uses/Additional Applications

We see predicting factors affect life expectancy in one way or another. We saw leaving out some features does affect the average value of life expectancy. Similarly, the addition of a few features may also affect the result. With the addition of features to the dataset, this prediction model can be routinely run to identify the factors impacting life expectancy and take necessary action to increase the life span of human beings.

Recommendations

Based on the available data, this model predicts the life expectancy and features impacting the life expectancy with better accuracy. However, this model should be regressed again when more data is available.

Implementation Plan

With the current features available in the dataset, this model can be implemented to predict the life expectancy of a country. In addition, this model can be launched to evaluate various features impacting life expectancy. However, as additional features are added to the dataset, this model must be reevaluated to ensure there is no slippage due to added features.

Ethical Assessment

One of the ethical considerations for this project is the consideration of results from the analysis in decision-making. Some of the conclusions made from this project's study could be incorrect or misrepresented due to insufficient or incorrect data. So, while sharing the outcome of this project with a larger audience, the underlying assumptions and data considerations should be shared.

Another ethical consideration is to ensure no personal and sensitive information is present in the dataset. This dataset is an extract from Global Health Observatory (GHO) data repository governed by WHO and made available for public use. So, this dataset doesn't have any sensitive information.

References

KUMARRAJARSHI (2017). Life Expectancy (WHO). https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Arya Andhika (2019-08-12): Life Expectancy Prediction using Regression. https://rstudio-pubs-static.s3.amazonaws.com/534874_2bdd7c6645804fd1b240e1ca3a9eb9d6.html

World Health Organization: https://www.who.int/data/gho,
https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates

Shashank Gupta. https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression

Caitlin McDonnell (2018-02-04). Machine learning to predict life expectancy.
https://towardsdatascience.com/what-really-drives-higher-life-expectancy-e1c1ec22f6e1