Life Expectancy Analysis and Prediction

FDS ASSIGNMENT

Anusha Chandrasekaran

185001020

I. Problem Statement

The term “life expectancy” refers to the number of years a person can expect to live. Shifts in life expectancy are often used to describe trends in the mortality rate. Being able to predict how populations will age has enormous implications for the planning and provision of services and support. Life expectancy and thereby mortality rate is an important feature that helps in the Demographic Analysis of a Country. This study aims to explore the relationships between different features and life expectancy and predict the life expectancy of different countries based on relevant features.

II. Literature Survey

This section of the report describes and elaborates related works that have tried to solve a similar problem statement. Models proposed by Mhd. Khafiroh Zamzamy Sormin *et al* [1] uses Cyclical Order Weight / Bias to train. They claim to have used five architecture models to predict the life expectancy.

On the other hand, Hendricks, Audrey and Graves, Philip E [2] propose two regression models which differ in the raining data given to them. One model is said to be composed of variables that are proven to be of primary importance when predicting life expectancy, and the other is a model consisting of features that were left out.

Kyle J Foreman et al.[3] claim to have modelled 250 causes and cause groups organized by the Global Buren of Diseases, Injuries and Risk Factor Study to generate predictions of worldwide life expectancy for 2017-2040. Separate models were developed for some causes of death because of their unique nature. They have predicted the life expectancy and mortality trends of all countries until the year 2040.

III. Scope

This study explores the relationship between life expectancy and a few features present in the dataset. In the scope of this study, the relationship between life expectancy and geography specific features such as air pollution in specific locality, etc is not explored. Thus, it only considers the unified life expectancy of the country and thus cannot be localized any further.

The features used don’t include any recent medical data such as Cancer, Tuberculosis, etc. Thus, only the economic side of this problem statement was explored.

IV. Dataset Description

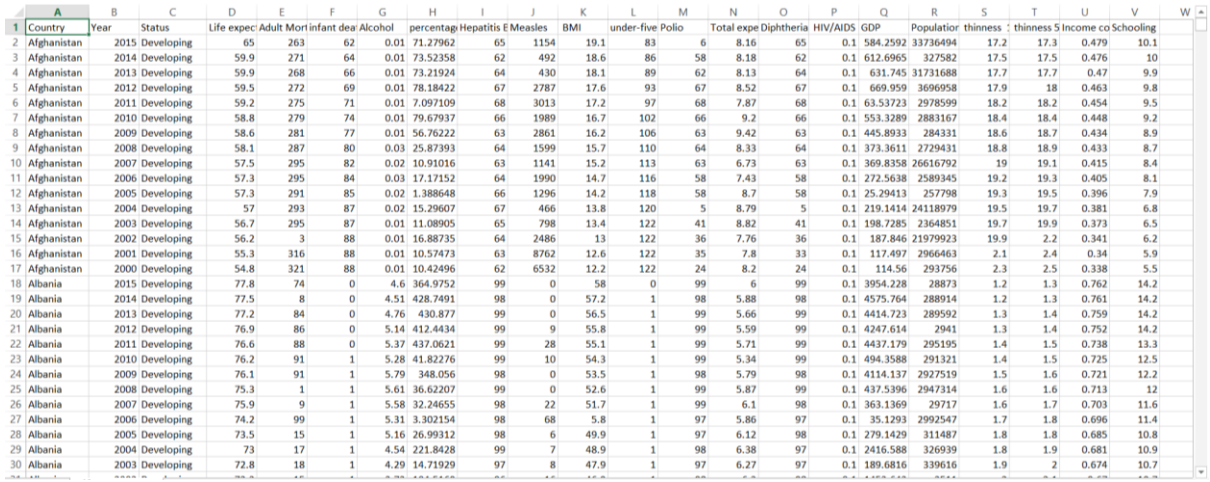
The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries The datasets are made available to public for the purpose of health data analysis.

This dataset was obtained from Kaggle (https://www.kaggle.com/kumarajarshi/life-expectancy-who). It consisted of data from 193 countries - 22 columns and 2938 rows with 20 numerical features.

The dataset consisted of:

- Country - Country
- Year - Year
- Status - Developed or Developing status
- Life expectancy - Life Expectancy in age
- Adult Mortality - Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- infant deaths – Number of infant deaths per 1000 population
- Alcohol - Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- percentage expenditure - Expenditure on health as a percentage of Gross Domestic Product per capita(%)
- Hepatitis B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Measles - Measles - number of reported cases per 1000 population
- BMI – Average Body Mass Index of Entire Population
- under-five deaths - Number of under-five deaths per 1000 population
- Polio - Polio (Pol3) immunization coverage among 1-year-olds (%)
- Total expenditure - General government expenditure on health as a percentage of total government expenditure (%)
- Diphtheria - Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
- HIV/AIDS - Deaths per 1 000 live births HIV/AIDS (0-4 years)
- GDP – Gross Domestic Product per capita (in USD)
- Population – Population of the country
- thinness 1-19 years – Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- thinness 5-9 years - Prevalence of thinness among children for Age 5 to 9(%)
- Income composition of resources - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling – Number of years of Schooling(years)

Each row depicted the life expectancy and numerical feature details of a country for a specific year. It also includes a column which describes if the country is a developed or a developing nation.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Country	Year	Status	Life expec	Adult Mort	infant dea	Alcohol	percentag	Hepatitis E	Measles	BMI	under-five	Polio	Total expe	Diphtheria	HIV/AIDS	GDP	Populatior	thinness :	thinness 5	Income co	Schooling	
2	Afghanistan	2015	Developing	65	263	62	0.01	71.27962	65	1154	19.1	83	6	8.16	65	0.1	584.2592	33736494	17.2	17.3	0.479	10.1	
3	Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358	62	492	18.6	86	58	8.18	62	0.1	612.6965	327582	17.5	17.5	0.476	10	
4	Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924	64	430	18.1	89	62	8.13	64	0.1	631.745	31731688	17.7	17.7	0.47	9.9	
5	Afghanistan	2012	Developing	59.5	272	69	0.01	78.18422	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9	18	0.463	9.8	
6	Afghanistan	2011	Developing	59.2	275	71	0.01	7.097109	68	3013	17.2	97	68	7.87	68	0.1	63.53723	2978599	18.2	18.2	0.454	9.5	
7	Afghanistan	2010	Developing	58.8	279	74	0.01	79.67937	66	1989	16.7	102	66	9.2	66	0.1	553.3289	2883167	18.4	18.4	0.448	9.2	
8	Afghanistan	2009	Developing	58.6	281	77	0.01	56.76222	63	2861	16.2	106	63	9.42	63	0.1	445.8933	284331	18.6	18.7	0.434	8.9	
9	Afghanistan	2008	Developing	58.1	287	80	0.03	25.87393	64	1599	15.7	110	64	8.33	64	0.1	373.3611	2729431	18.8	18.9	0.433	8.7	
10	Afghanistan	2007	Developing	57.5	295	82	0.02	10.91016	63	1141	15.2	113	63	6.73	63	0.1	369.8358	26616792	19	19.1	0.415	8.4	
11	Afghanistan	2006	Developing	57.3	295	84	0.03	17.17152	64	1990	14.7	116	58	7.43	58	0.1	272.5638	2589345	19.2	19.3	0.405	8.1	
12	Afghanistan	2005	Developing	57.3	291	85	0.02	1.388648	66	1296	14.2	118	58	8.7	58	0.1	25.29413	257798	19.3	19.5	0.396	7.9	
13	Afghanistan	2004	Developing	57	293	87	0.02	15.29607	67	466	13.8	120	5	8.79	5	0.1	219.1414	24118979	19.5	19.7	0.381	6.8	
14	Afghanistan	2003	Developing	56.7	295	87	0.01	11.08905	65	798	13.4	122	41	8.82	41	0.1	198.7285	2364851	19.7	19.9	0.373	6.5	
15	Afghanistan	2002	Developing	56.2	3	88	0.01	16.88735	64	2486	13	122	36	7.76	36	0.1	187.846	21979923	19.9	2.2	0.341	6.2	
16	Afghanistan	2001	Developing	55.3	316	88	0.01	10.57473	63	8762	12.6	122	35	7.8	33	0.1	117.497	2966463	2.1	2.4	0.34	5.9	
17	Afghanistan	2000	Developing	54.8	321	88	0.01	10.42496	62	6532	12.2	122	24	8.2	24	0.1	114.56	293756	2.3	2.5	0.338	5.5	
18	Albania	2015	Developing	77.8	74	0	4.6	364.9752	99	0	58	0	99	6	99	0.1	3954.228	28873	1.2	1.3	0.762	14.2	
19	Albania	2014	Developing	77.5	8	0	4.51	428.7491	98	0	57.2	1	98	5.88	98	0.1	4575.764	288914	1.2	1.3	0.761	14.2	
20	Albania	2013	Developing	77.2	84	0	4.76	430.877	99	0	56.5	1	99	5.66	99	0.1	4414.723	289592	1.3	1.4	0.759	14.2	
21	Albania	2012	Developing	76.9	86	0	5.14	412.4434	99	9	55.8	1	99	5.59	99	0.1	4247.614	2941	1.3	1.4	0.752	14.2	
22	Albania	2011	Developing	76.6	88	0	5.37	437.0621	99	28	55.1	1	99	5.71	99	0.1	4437.179	295195	1.4	1.5	0.738	13.3	
23	Albania	2010	Developing	76.2	91	1	5.28	41.82276	99	10	54.3	1	99	5.34	99	0.1	494.3588	291321	1.4	1.5	0.725	12.5	
24	Albania	2009	Developing	76.1	91	1	5.79	348.056	98	0	53.5	1	98	5.79	98	0.1	4114.137	2927519	1.5	1.6	0.721	12.2	
25	Albania	2008	Developing	75.3	1	1	5.61	36.62207	99	0	52.6	1	99	5.87	99	0.1	437.5396	2947314	1.6	1.6	0.713	12	
26	Albania	2007	Developing	75.9	9	1	5.58	32.24655	98	22	51.7	1	99	6.1	98	0.1	363.1369	29717	1.6	1.7	0.703	11.6	
27	Albania	2006	Developing	74.2	99	1	5.31	3.302154	98	68	5.8	1	97	5.86	97	0.1	35.1293	2992547	1.7	1.8	0.696	11.4	
28	Albania	2005	Developing	73.5	15	1	5.16	26.99312	98	6	49.9	1	97	6.12	98	0.1	279.1429	311487	1.8	1.8	0.685	10.8	
29	Albania	2004	Developing	73	17	1	4.54	221.8428	99	7	48.9	1	98	6.38	97	0.1	2416.588	326939	1.8	1.9	0.681	10.9	
30	Albania	2003	Developing	72.8	18	1	4.29	14.71929	97	8	47.9	1	97	6.27	97	0.1	189.6816	339616	1.9	2	0.674	10.7	

V. Methodology

1. Brief look of Dataset

```
: df.head()
```

:

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	G
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.259
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.696
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.744
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.959
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.537

5 rows × 22 columns

2. Checking for missing values

The dataset is found to have missing values as shown below, so the mean of that column was found and filled in the missing values' locations.

```
df.isnull().sum()
```

```
Country                          0
Year                             0
Status                           0
Life expectancy                 10
Adult Mortality                 10
infant deaths                    0
Alcohol                        194
percentage expenditure           0
Hepatitis B                    553
Measles                          0
 BMI                            34
under-five deaths                0
Polio                           19
Total expenditure              226
Diphtheria                      19
 HIV/AIDS                        0
GDP                            448
Population                     652
 thinness  1-19 years           34
 thinness 5-9 years             34
Income composition of resources  167
Schooling                      163
dtype: int64
```

```
df.fillna(df.mean(),inplace=True)
```

3. Non graphical Exploratory Data Analysis (EDA)

The top 10 countries with maximum and minimum life expectancy is found.

	Country	Life expectancy
84	Japan	82.53750
165	Sweden	82.51875
75	Iceland	82.44375
166	Switzerland	82.33125
60	France	82.21875
82	Italy	82.18750
160	Spain	82.06875
7	Australia	81.81250
125	Norway	81.79375
30	Canada	81.68750

Top 10

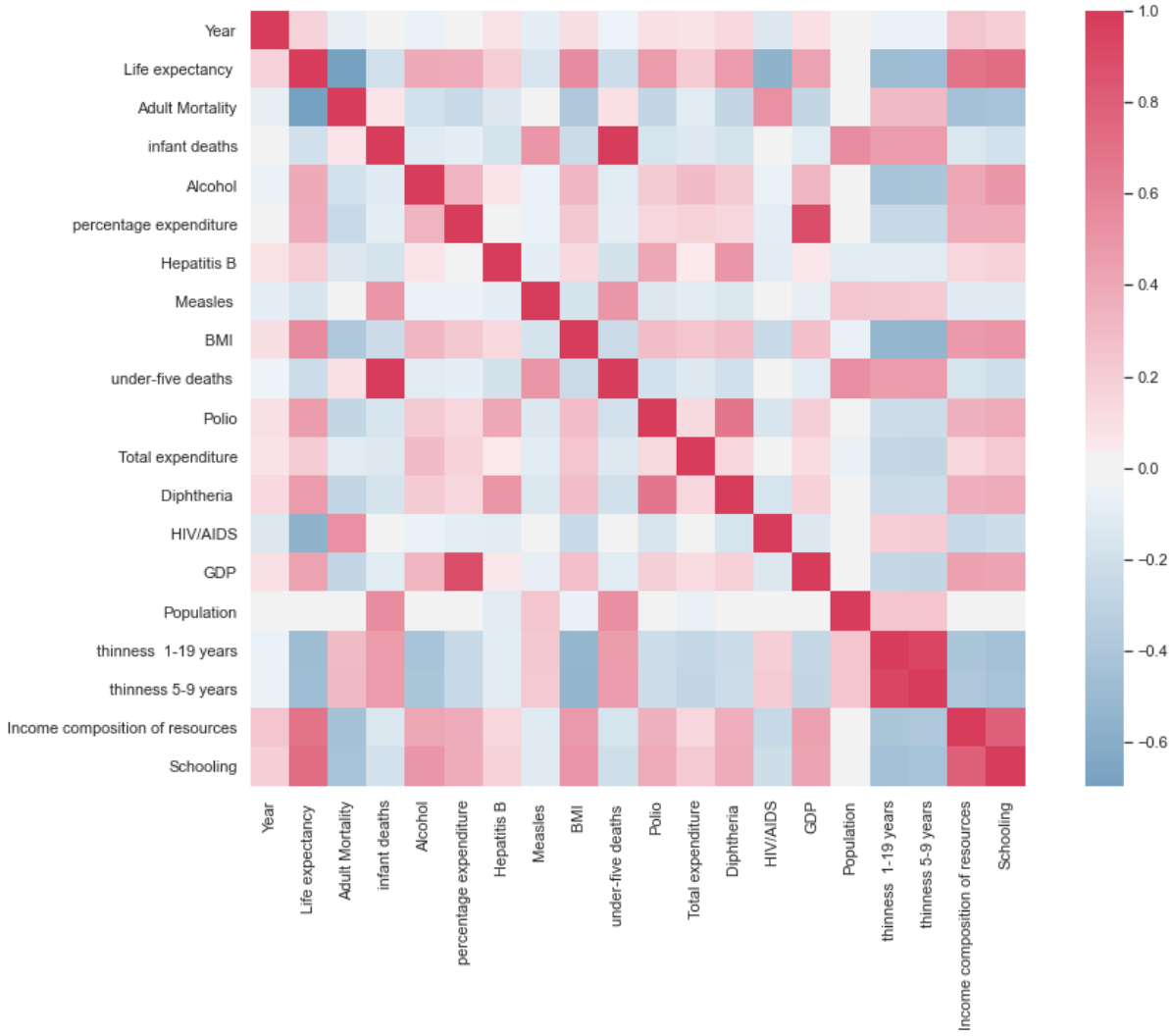
	Country	Life expectancy
152	Sierra Leone	46.11250
31	Central African Republic	48.51250
94	Lesotho	48.78125
3	Angola	49.01875
100	Malawi	49.89375
32	Chad	50.38750
44	Côte d'Ivoire	50.38750
192	Zimbabwe	50.48750
164	Swaziland	51.32500
123	Nigeria	51.35625

Bottom 10

The average life expectancy of a developed and a developing nation was found.

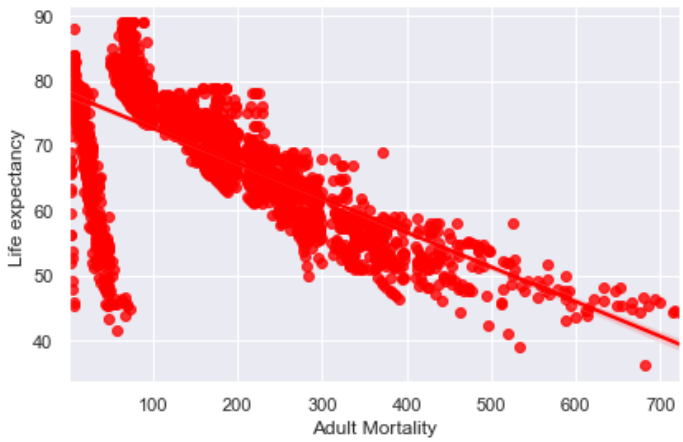
	Status	Life expectancy
0	Developed	79.197852
1	Developing	67.120177

4. Graphical Exploratory Data Analysis (EDA)

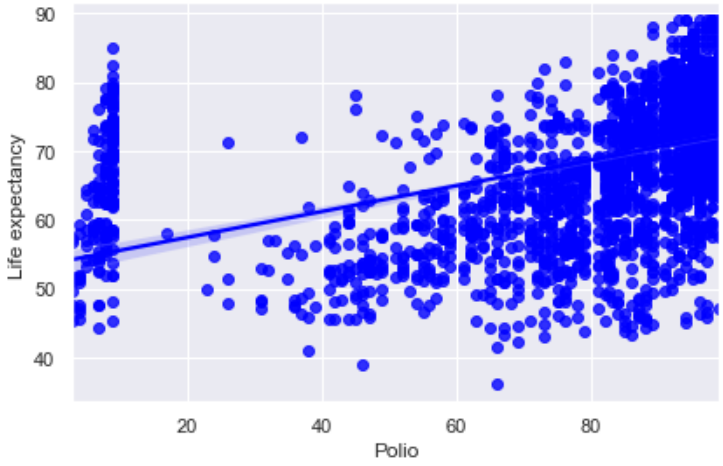


Plotting of a heatmap (correlation matrix) helped identify which features life expectancy depended on the most. It also shows the correlation of each feature to every other feature and the intensity of the colour denotes the correlation degree. Here, blue implies inversely related to a high degree, while red implies directly related to a high degree.

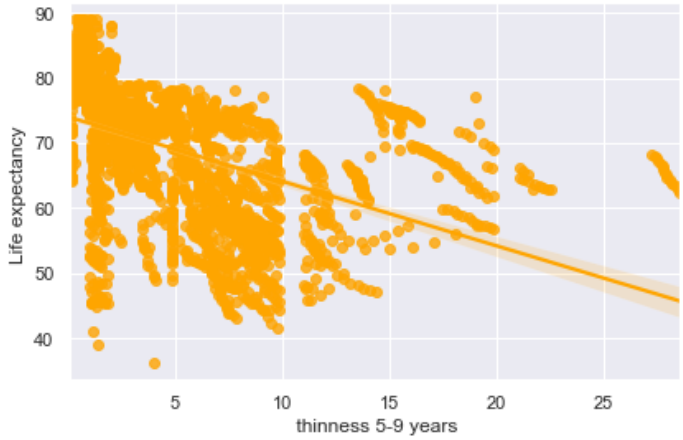
Here, we can see clearly that Life expectancy is negatively correlated to Adult Mortality rate.



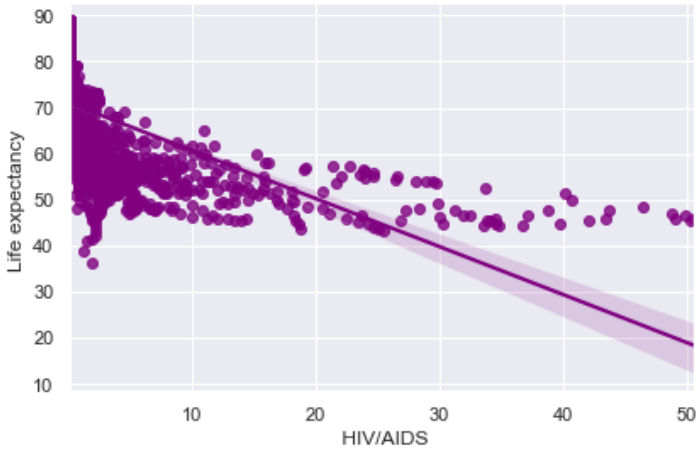
Next, we see that Polio immunization coverage % is positively related to Life expectancy.



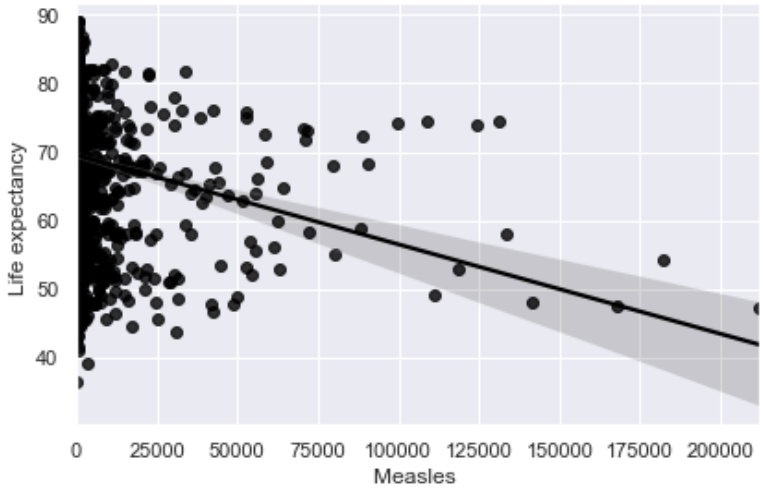
We also see that Prevalence of thinness among children for Age 5 to 9(%) is negatively correlated to Life expectancy.



There is a negative relationship between number of deaths due to AIDS/HIV and Life expectancy.



As seen below, number of reported Measles cases is inversely related to Life expectancy.



5. Data Preprocessing

Since we have already checked for missing values, we only need to change the categorical features (Status) to binary.

6. Splitting the data into train and test and scaling the data

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test= train_test_split(X, y, test_size= 0.30, random_state=9)
```

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

7. Training and Testing the models using Regression

Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x). Briefly, the goal of regression model is to build a mathematical equation that defines y as a function of the x variables.

To predict Life expectancy, classification or regression can be done. Here, I have used five regressors namely Random Forest, AdaBoost, GradientBoost, Decsion Tree and Linear regressors to predict the life expectancy of the test data.

```
from sklearn.ensemble import RandomForestRegressor, AdaBoostRegressor, GradientBoostingRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_absolute_error as mape
from sklearn.metrics import mean_squared_error as mse
from sklearn.metrics import r2_score as r2

alg = [RandomForestRegressor(), AdaBoostRegressor(), GradientBoostingRegressor(), DecisionTreeRegressor(), LinearRegression()]
arr =[]
```

8. Use the performance metrics- R squared score, Mean Absolute Error and Mean Squared Error to find the model which was most accurate.

VI. Tools Used

Libraries Used:

- pandas
- Seaborn
- Matplotlib
- NumPy
- SciKit-Learn

VII. Other Tools

- Keras
- SciPy
- TensorFlow
- PyTorch
- Scrappy
- PyCaret
- Plotly

VIII. Performance Metrics

Three metrics have been used for model evaluation in regression:

1. R Square/Adjusted R Square

2. Mean Square Error(MSE)/Root Mean Square Error(RMSE)

3. Mean Absolute Error(MAE)

1. *R Square* measures how much of variability in dependent variable can be explained by the model. It is square of Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

2. While R Square is a relative measure of how well the model fits dependent variables, *Mean Square Error* is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

3. *Mean Absolute Error(MAE)* is similar to Mean Square Error(MSE). However, instead of the sum of square of error in MSE, MAE is taking the sum of absolute value of error. Compare to MSE or RMSE, MAE is a more direct representation of sum of error terms. MSE gives larger penalisation to big prediction error by square it while MAE treats all errors the same.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

IX. Results

After the training and testing, each model predicted the life expectancy based on the same features.

```
RandomForestRegressor()
Test Root Mean Squared error: 1.9402415001033093
R Square Score: 0.9590031503453206
Mean Absolute Error: 1.2141014475917877
---------------------------------------------------------------------
AdaBoostRegressor()
Test Root Mean Squared error: 2.9870762956374843
R Square Score: 0.9028301489305206
Mean Absolute Error: 2.3364483557015734
---------------------------------------------------------------------
GradientBoostingRegressor()
Test Root Mean Squared error: 2.2837172039881923
R Square Score: 0.9432032411130977
Mean Absolute Error: 1.6395904841277227
---------------------------------------------------------------------
DecisionTreeRegressor()
Test Root Mean Squared error: 2.6992173777464967
R Square Score: 0.9206559017499877
Mean Absolute Error: 1.6577380177936387
---------------------------------------------------------------------
LinearRegression()
Test Root Mean Squared error: 4.1621753919264215
R Square Score: 0.8113400736351479
Mean Absolute Error: 3.0740104519327307
---------------------------------------------------------------------
```

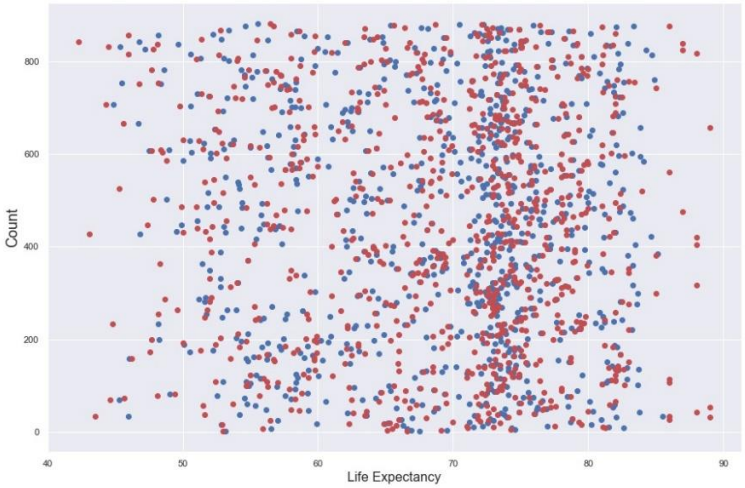
As we can see from the output, Random Forest Regressor has the least error and a reasonable R squared value.

The following plots were made with the y axis being an iterable count which denotes each row in the dataset, the x axis being Life expectancy.

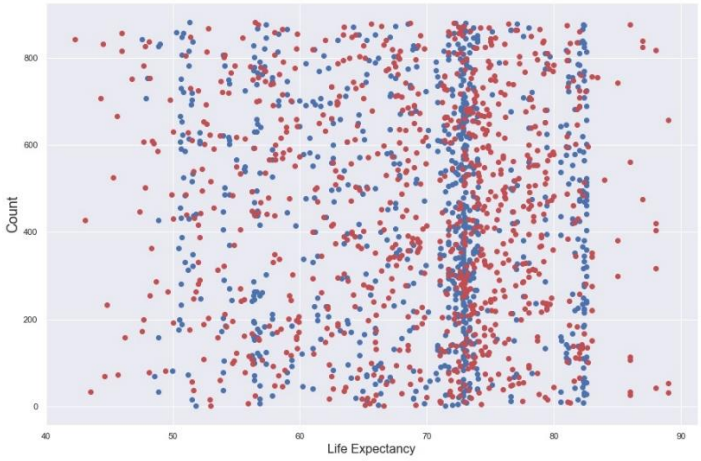
Red dots—denote the target Life expectancy values

Blue dots—denote the predicted expectancy values

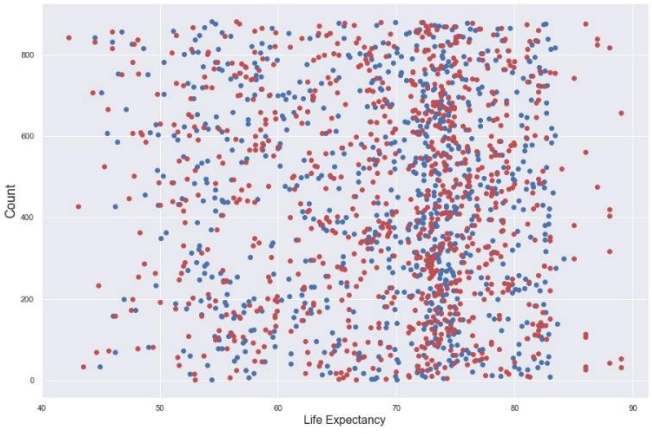
Random Forest Regression

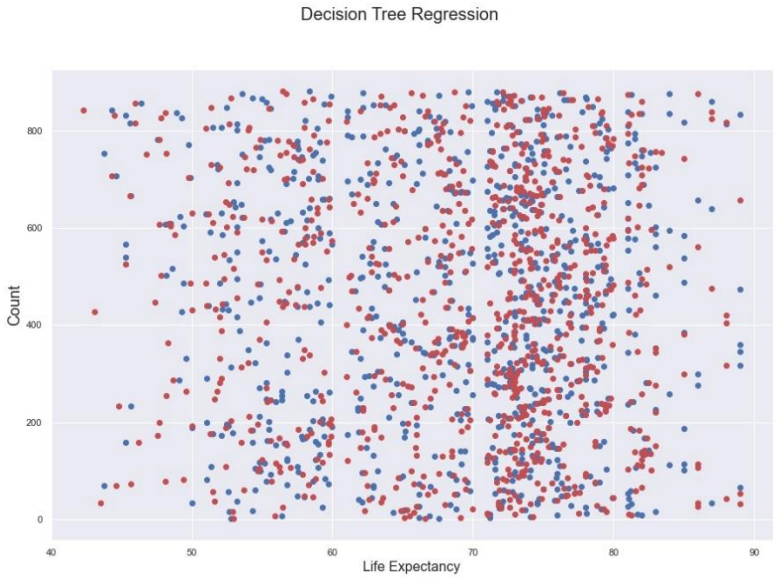


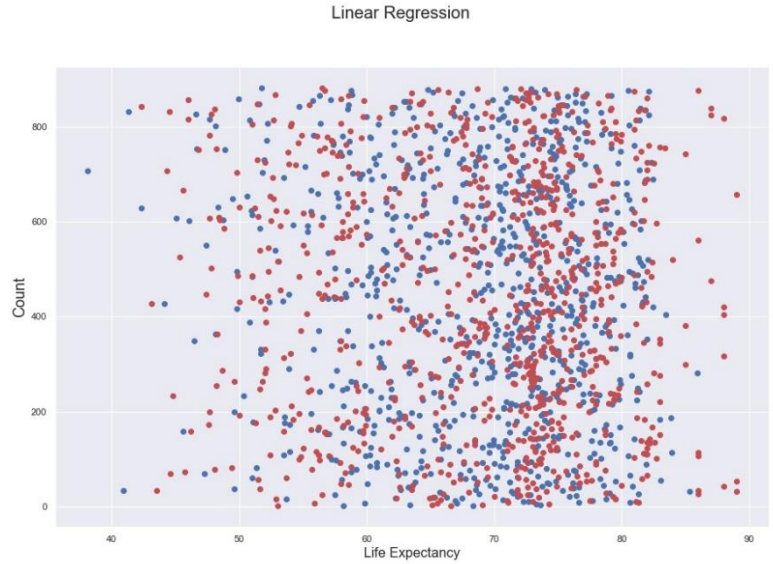
Ada Boost Regression



Gradient Boosting Regression







Random Forest Regressor is the best model in our study as it has the least error and an acceptable R squared score. It can also be seen that the plot with random forest regressor has the most overlap with the target data.

X. Conclusion

From our EDA as well as our prediction of life expectancy, we can conclude that life expectancy majorly depends on Adult Mortality, Polio and Hepatitis immunizations, Measles patients, HIV/AIDS deaths, thinness in the first 5-9 years and thinness in the first 1-19 years.

The prediction done using Random Forest Regressor was observed to be the best amongst the four other regressors. The data predicated seemed to have a relatively good accuracy.

XI. References

[1]. Mhd. Khafiroh Zamzamy Sormin *et al* 2019 *J. Phys.: Conf. Ser.* **1255** 012017

[2]. Hendricks, Audrey and Graves, Philip E., Predicting Life Expectancy: A Cross-Country Empirical Analysis (September 23, 2009) doi: 10.2139/ssrn.1477594

[3]. Kyle J Foreman, Neal Marquez, Andrew Dolgert, Kai Fukutaki, Nancy Fullman, Madeline McGaughey, et al doi: https://doi.org/10.1016/S0140-6736(18)31694-5

[4]. https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b

[5]. https://medium.com/@nirankari.naveen.13et1136/how-to-predict-life-expectancy-using-machine-learning-5c253ab25125

[6]. https://github.com/SmartPracticeschool/llSPS-INT-2166-Predicting-Life-Expectancy-using-Machine-Learning