Predicting Life Expectancy at Birth based on Socioeconomic Factors

Project report for DS203 Programming for Data Science 2021, IIT Bombay

Aaditya Sakrikar
Dept. of Mechanical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
200100002@iitb.ac.in

Amritaansh Narain
Dept. of Mechanical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
200100022@iitb.ac.in

Sneha Kulkarni
Dept. of Mechanical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
200100090@iitb.ac.in

***Abstract*—Life expectancy at birth is the number of years a person who is just born is expected to live. The trend of its value reflects the trend of the factors related to mortality, the state of a country's health infrastructure and the overall prosperity of a country. This makes Life Expectancy a very important factor to study and try to predict its value in the future. We collected World Bank's data for various related variables, then we did EDA of all variables we thought are related to Life Expectancy and modelled our target variable based on these selected variables. We found that Neural Networks gave us the best model. Also, we saw which variables affect the Life Expectancy the most and whether that effect is directly or inversely proportional. Also, we compared the values of variables over time and it's correlation.**

I. Introduction

Life expectancy at birth is one of the most important indicators used to assess the health of a country's population. It is a statistical measure of the number of years a person can be expected to live after birth depending on the nation's socioeconomic and geographical characteristics and reflects the overall mortality level of a particular country or region. A nation's overall income, the standard of living, vaccinations, density of population, pollution levels, health infrastructure, food availability, people with different diseases, etc play a crucial role in determining life expectancy at birth.

A country with higher population below the poverty line can be expected to have a lower life expectancy because of its inability to afford electric power, top-quality medical care, and other facilities necessary for living. At the same time, a country with a high density of population, but with less industrialization can be expected to have lower life expectancy because of a shortage of supply of resources. Another example can be of a country with higher literacy, in this case as the people of that country are more educated and aware, hence there is higher probability that they would be vaccinated in comparison to a country with lower literacy levels. This in turn will lead to higher Life Expectancy in the country with higher Literacy levels. Similarly, factors like the inflation rate, unemployment of labour, pollution, and the literacy rate of a country will significantly affect life expectancy.

We have done a thorough analysis of how each factor affects our target variable, the life expectancy at birth, and also observed the mutual trends among the above variables. Apart from this, we have selected four countries in each income category, i.e., high, low and middle, and studied each parameter for these groups of countries. Along with this, we have interpreted the results for an overview of high, middle, and low-income countries on the basis of the above parameters. After Exploratory Data Analysis, we have trained machine learning models including Linear Regression, Decision Tree, Support Vector Machine, and Neural Networks, wherein our aim was to minimize the error parameter.

II. Background and Prior Work

The data that we used for our project has been obtained from 'World Bank Data' which is an open-source website containing data about World Development Indicators, Statistical Capacity Indicators, Gender Statistics, Health and Nutrition Statistics for several countries. Out of these, we shortlisted the data relevant for determining life expectancy. Dataset for each variable was available for each country and some groups of countries from year 1960-2019. We first looked at all possible variables whose datasets were available and logically short-listed the ones we thought were the most relevant for predicting the value of Life Expectancy at Birth. The variables we selected were as follows:

1) Life Expectancy
2) Access to Electricity (%)
3) Electric Power Consumption (kWh/capita)
4) Percentage of Rural Population
5) Year
6) Food Production Index
7) GDP per capita ($)
8) Immunization (DPT)
9) Immunization (Measles)
10) Road Accident Mortality
11) Population Density
12) Total Greenhouse Gas Emissions
13) Unemployed Labour (%)
14) Literacy rate (%)

We created two datasets, a smaller one for analysis (EDA) and a more comprehensive dataset with more data for mod-

elling. For the EDA part we chose 12 countries, 4 each from 3 income categories, low income, middle income and high income. The timeline we chose for the analysis was 2000-2014 as most of the parameters had good amount of data in this time frame. The one variable which was an exception to this was adult literacy rate data. So, we had to manually fill adult literacy rate using arithmetic progression and the additional values we found on the internet for the years where the data was missing, as the literacy rate in the dataset was available only for intervals of 5 years. We used a shorter dataset to keep accuracy in actual analysis since for a large number of countries it was harder to fill the literacy rate data which was mostly blank. We organized the collected data grouped by country name, and columns were the variables. Also, we added additional 3 countries, namely, High Income, Middle Income and Low Income, which basically have the mean value of all the parameters for all the countries in the particular Income bracket for the complete dataset. This was done in order to do an overall comparison of the 3 income categories during EDA.

For the larger data to be used for analysis, we collected data in a similar fashion but again the issue we faced was the mostly blank literacy rate dataset. We collected data for 134 countries for each year from 2000-2014. We grouped the data from separate variable datasets to a single dataset grouped by country and columns being the variable values. This gave us a total of 2010 records to use for modelling. The issue of hardly filled adult literacy data was solved by using mean of available data for a country with some added noise to fill the blank spots. This was done because for the countries which had completely filled literacy rate, there was hardly any shift in percentage among just 15 years.

This way we had created 2 datasets, one for analysis which was more accurate while the other one was much larger however we lost little bit of accuracy in adult literacy rate data. Note that we could not leave literacy rate data despite being hard to collect because adult literacy, as we read, plays a strong role in life expectancy prediction. [**?**]

III. Data and Methodology

A. *Life Expectancy at Birth*

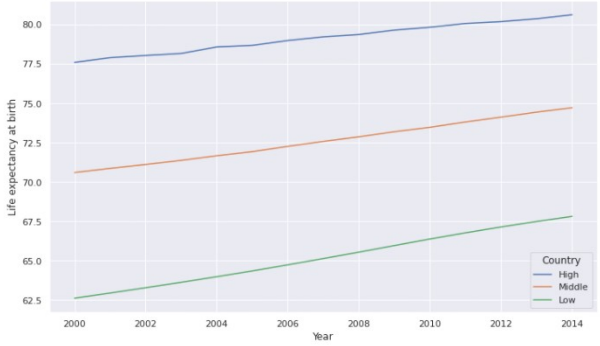

Fig. 1. Life Expectancy at Birth for countries distributed by income

Over the years we can see a steady increase in Life Expectancy for all the countries combined. There is a significant difference between the Life Expectancy of the three income categories overall years but the rate of increase of life expectancy is much faster for poorer countries as compared to the higher ones. On an average Japan has had the highest Life expectancy and Zimbabwe has had the lowest value.

For India we see a continuous and steady increase in life expectancy over the years.

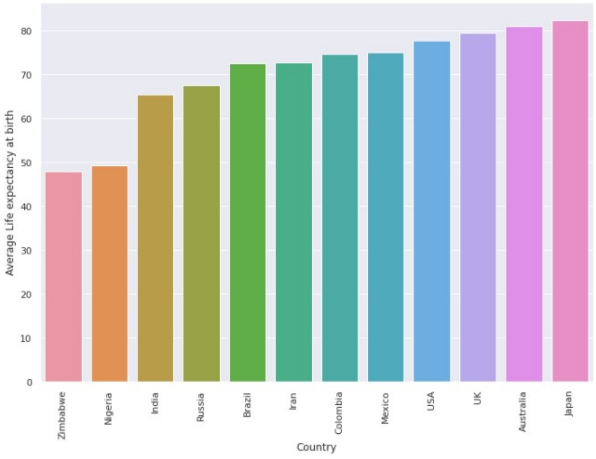

Fig. 2. Average Life Expectancy at Birth of different countries

B. *Access to Electricity (%)*

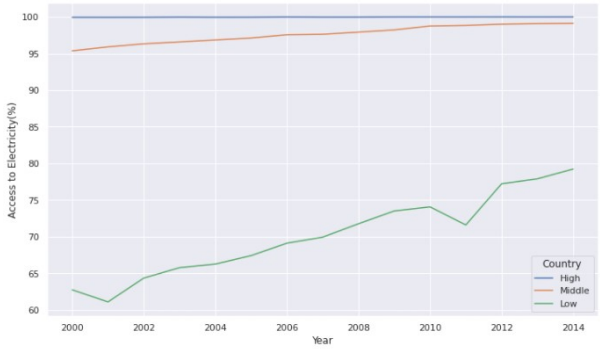

Fig. 3. Access to Electricity (%) for countries distributed by income

We see that most of countries today have very high electric coverage but there are a few poor countries like Zimbabwe, Nigeria, India which lack behind by a huge margin. Also, we can see that there is insignificant difference in access to electricity for high and middle-income countries but a huge difference between these two categories and poor income countries. From the scatter plot we can see that there is direct correlation between access to electricity and life expectancy and life expectancy increases significantly with better access to electricity. For India the overall trend over the years has been an increase in access to electricity, with a slight dip on 2-3 instances.

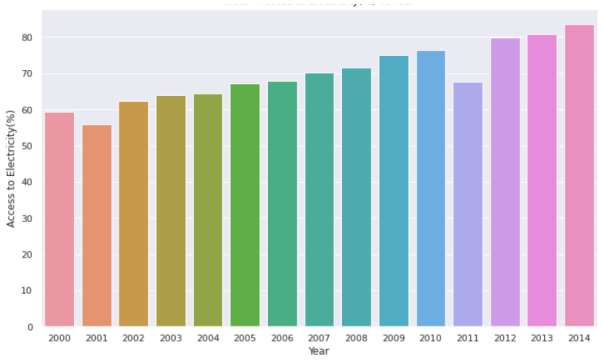

Fig. 4. Access to Electricity (%) in India over the years

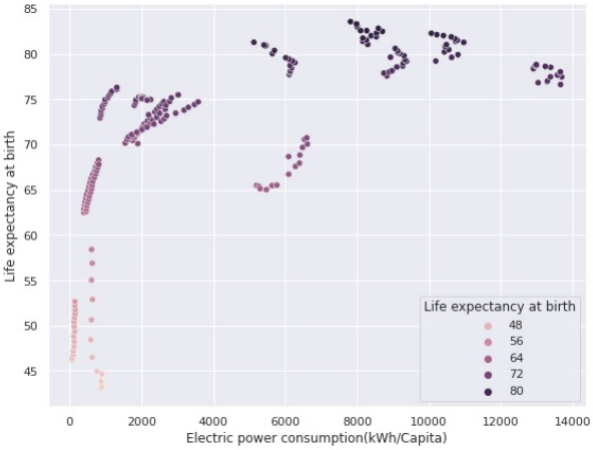

Fig. 5. Scatter Plot of Life Expectancy at Birth vs Electric Power Consumption (kWh/capita)

C. *Electric Power Consumption*

We see a huge difference in power consumption between the high-income countries with respect to the other income countries, also there has been a reasonable increase in power consumption for middle income countries but negligible change for poor countries over the time period. USA has the highest power consumption and Nigeria the lowest with huge differences between the rich and poor countries. Also, as the power consumption increases there is a clear increase in life expectancy. For India power consumption has been on a significantly rapid increase over the years.

D. *Rural Population*

We see big differences between rural population percentages of high, middle and low income with a continuous decrease across income categories but the decrease is the most rapid for middle income countries. Poor countries like India, Zimbabwe and Nigeria have very high rural populations whereas Japan and other rich countries have very low percentages. From the scatter plot we see that life expectancy decreases significantly with an increase in rural population.

For India the rural population has been on a slow but steady decline with the rate of decline increasing over the years.

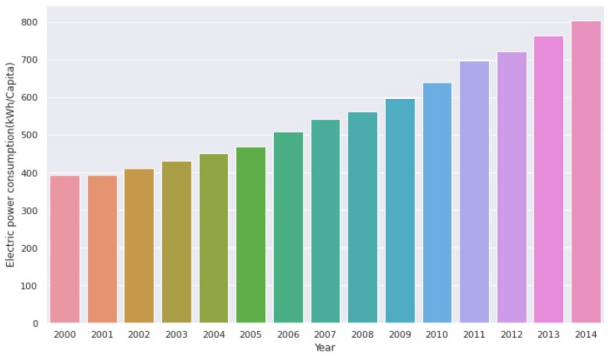

Fig. 6. Electric Power Consumption (%) in India over the years

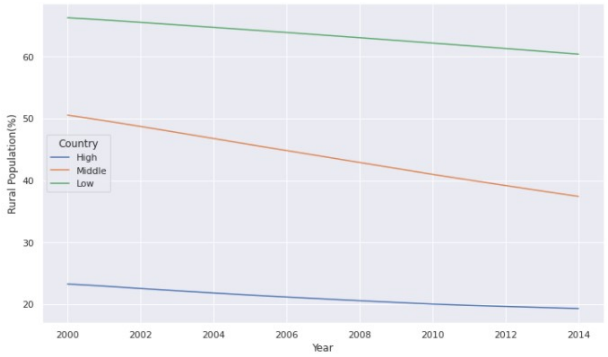

Fig. 7. Rural Population (%) for countries distributed by income

E. *Road Accident Mortality*

Studies show that road accident injuries are one of the major causes of death among the young population. The scatter plot shows that countries having low road accident mortality (5-10 per 1,00,000 people) have a higher high Life expectancy (as high as 80). Further, Road Accident Mortality has been the highest for Middle-Income countries and the lowest for High-Income countries. Also, high-income countries show gradually decreasing mortality over the years as opposed to the other

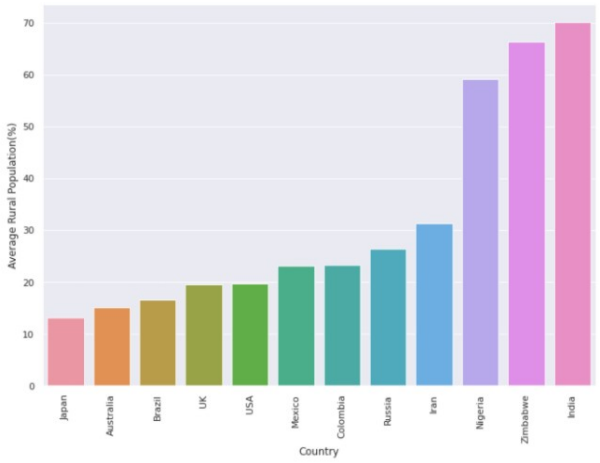

Fig. 8. Average Rural Population (%) for different countries

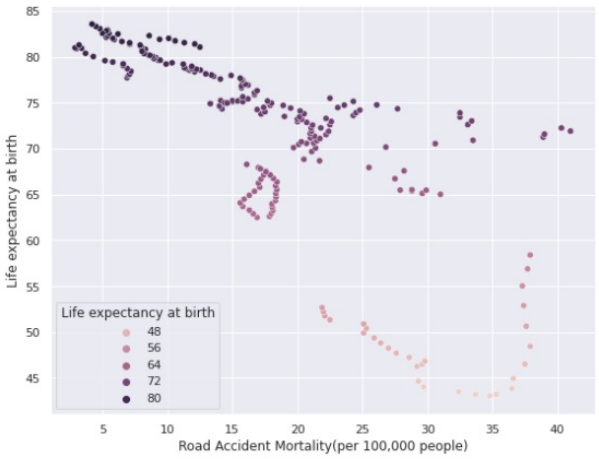

Fig. 9. Scatter Plot of Life Expectancy at Birth vs Road Accident Mortality (per 100,000 people)

two.

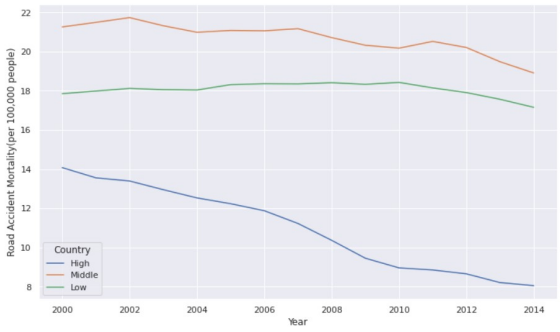

Fig. 10. Road Accident Mortality for countries distributed by income

F. Population Density

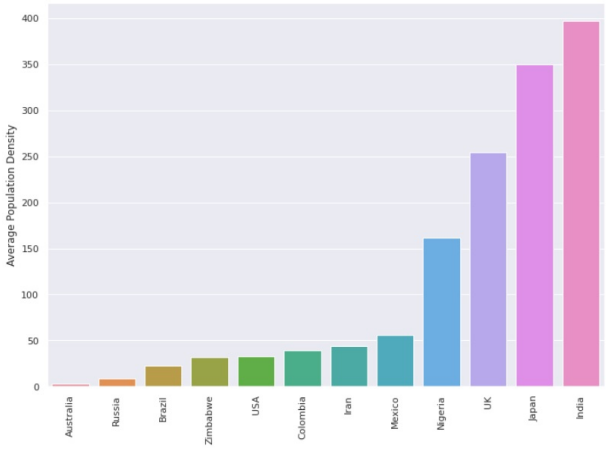

Fig. 11. Average Population Density for different countries

It is the ratio of the total population to the geographical land area of a country. The low-income countries, apart from having a higher population density, have a much higher rate of increase in population density compared to that of high and middle-income countries. Amongst the countries that we analysed; India has the highest population density while Australia has the lowest. Moreover, the population density for India has been continuously increasing over the years at a faster rate compared to the other low-income countries.

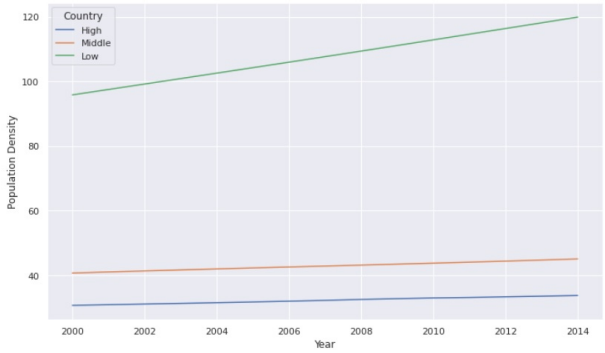

Fig. 12. Population Density for countries distributed by income

G. Total Greenhouse gas emissions

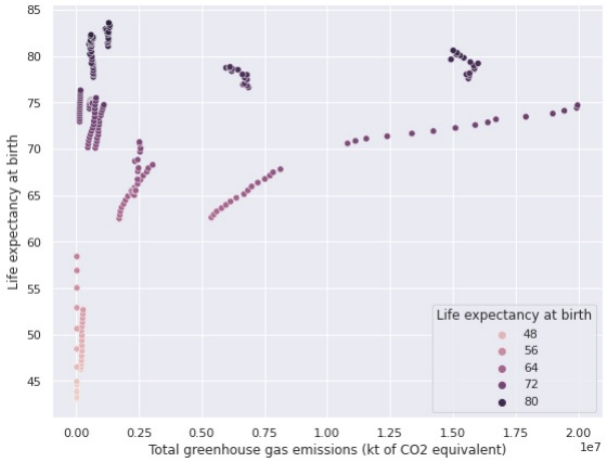

Fig. 13. Scatter plot of Life Expectancy at Birth vs Total Greenhouse Gas emissions

The USA has the highest emission of greenhouse gases $6.5 \cdot 10^5$, which is more than twice that of Russia $2.5 \cdot 10^5$, which is the second-highest in this category. We expected countries having low carbon emissions to achieve high life expectancy, but low income. However, vice versa may not be true, which can be seen from the scatter plot, where there are several cases where high life expectancy does not imply low emission of greenhouse gases. As far as India is concerned, greenhouse gas emission has risen by almost twice in a span of 15 years.

H. Unemployed Labour

It is obtained by dividing the number of unemployed people by the total number in the labor force. A very noticeable

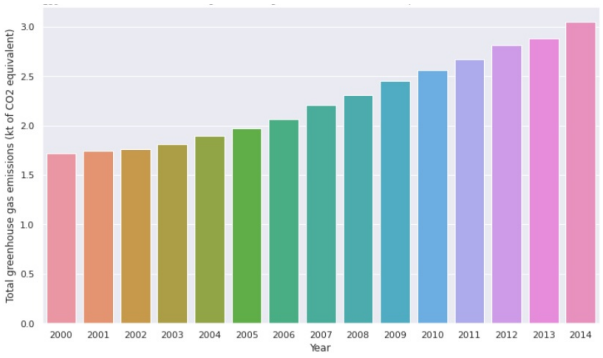

Fig. 14. Total Greenhouse gas emissions of India from 2000-2014

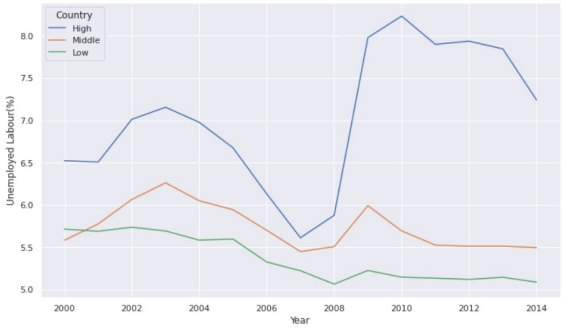

Fig. 15. Unemployed Labor among groups of countries divided by Income

change is seen from the years 2004 to 2009 where the unemployment has initially decreased and then has risen to a peak for the countries in all the income categories, with the most drastic increase being for the High-income countries. This peak is due to the Great Recession of 2008, which affected the high-income countries the most, whereas its effect on low-Income countries is negligible. In India, the analysis shows that the unemployed labour percent decreased in the period 2006-2008 and has maintained a nearly constant value before and after this dip in unemployed labour percentage.

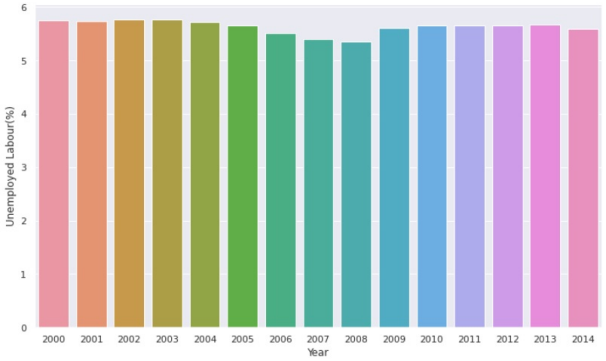

Fig. 16. Unemployed Labour for India from 2000-2014

I. Food Production Index

Food Production Index is a measure of the rise in prices of agricultural goods in a given year from base year. Calculated using the Laspeyres Formula.

$$\text{Laspeyeres Index} = \frac{\sum \text{Observation Price} \cdot \text{Base Quantity}}{\sum \text{Base Price} \cdot \text{Base Quantity}}$$

The quantities it covers are all crops and livestock products originating in the country. As expected because of net inflation positive for High, Middle and Low income countries in the list except Japan (High Income country), the FPI also shows an increase. Middle Income countries show lower FPI than

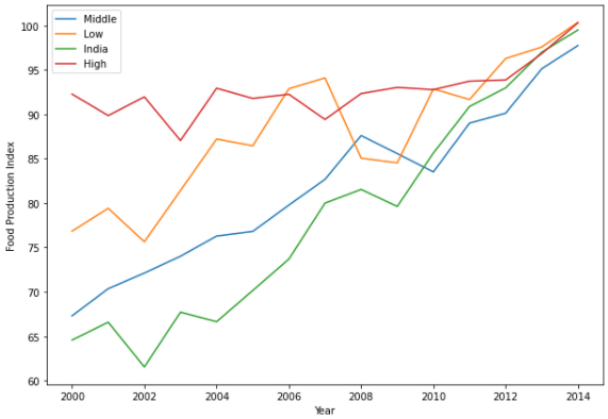

Fig. 17. Food Production Index of India with country groups based on income

both low and middle income countries throughout except a peak in 2008 (Year of global recession). By 2014 both low and high income countries reach the same FPI. India's FPI is worse among all the averages, however we see sharp increase in FPI after 2010. In the above scatter plot we can see our

Fig. 18. Scatter Plot of Life Expectancy at Birth vs Food Production Index

hypothesis that FPI has approximately zero relation with the life expectancy of the country.

We conclude FPI to not be an important variable in predicting the life expectancy of a country.

J. GDP per Capita

Measures the average economic contribution of each individual inside the country towards GDP. Measured by dividing

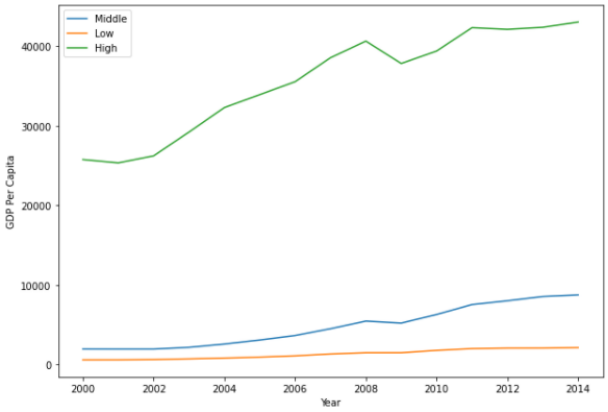

Fig. 19. GDP per capita of different country groups based on income

the GDP by the population of a country. Since we essentially defined countries by their income, which is a component in GDP calculation results of graph are as expected. Low income countries show stagnant GDP per capita which says they might be mostly underdeveloped countries or the GDP is increasing at the cost of increasing population. Which middle income countries are showing an increasing GDP per capita but at much slower rate than the high income countries. We

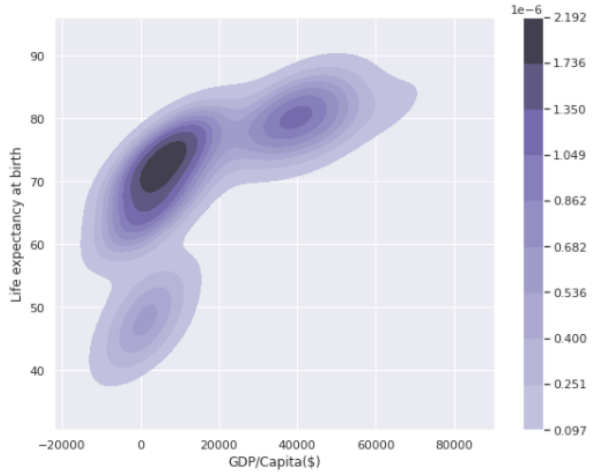

Fig. 20. Plot of Life Expectancy at Birth vs GDP per capita

can notice a linear relation between GDP per capita and life expectancy however, we see higher density in a region of low GDP per capita and high life expectancy at birth. This may seem to suggest less of a linear relation between the two but rather a graph which starts saturating after a given GDP per capita.

K. *Immunization*

We are measuring immunization to be the percentage of people immunized with either DPT vaccine or Measles vaccine. We took mean of the data for these two vaccines for rough estimate of percentage of people vaccinated. We get quite counter intuitive results for the middle income countries.

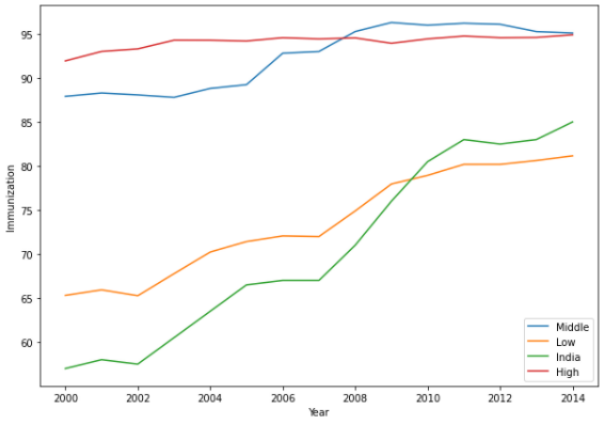

Fig. 21. Immunization of India and different country groups based on income

We see that in years following 2008, percentage of people vaccinated in middle income countries is more than high income countries though both converge to same values at 2014. Low income countries are still at the bottom with low percentage of people vaccinated. India however is showing very steep increase in percentage of population vaccinated.

L. *GDP Deflator*

The GDP price deflator is the implicit price deflator or Inflation, measures the changes in prices for all goods and services produced in an economy with respect to a base year.

$$\text{GDP Deflator} = \frac{\sum \text{Observation Price} \cdot \text{Current Quantity}}{\sum \text{Base Price} \cdot \text{Current Quantity}}$$

Developed countries tend to have low inflation rate and also

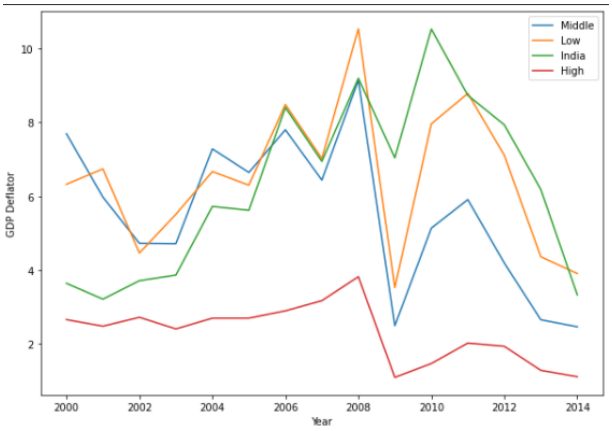

Fig. 22. GDP Deflator of India and different country groups based on income

quite stable ones, which can be seen in the data. Except we see a massing drop in inflation in all groups of countries and India, after 2008, this was because of a housing market crisis in USA, 2008. Because of the concept of sticky prices, instead of decrease in salary of people, people started losing jobs and high unemployment. The scatter plot suggests high density of records with low Inflation rates and high life expectancy. In low life expectancy region we see sparsely distributed points

Fig. 23. Scatter Plot of Life Expectancy at Birth vs GDP Deflator

in both high and low inflation rates. Low and stable inflation rates are generally for the developed/high income countries which tend to have high life expectancy.

M. Adult Literacy Rate

The literacy rate is defined by the percentage of the population of ages 15 and above that can read and write. The adult literacy rate corresponds to ages 15 and above. Literacy rates

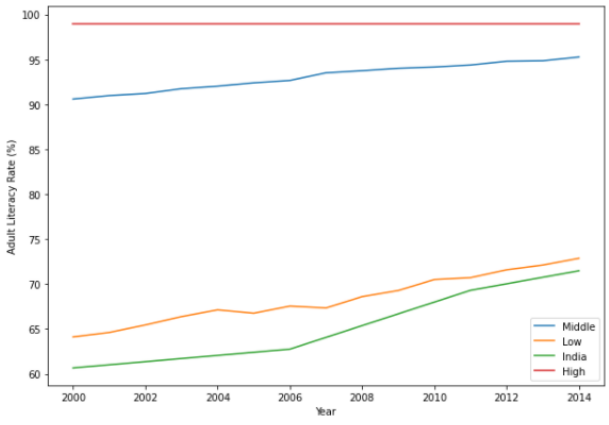

Fig. 24. Adult Literacy Rate of India and different country groups based on income

show near 100 values for the high income countries which are mostly developed. India however shows below average literacy rates even among low income countries. In above scatter plot we notice a very crude linear relation between literacy rates and life expectancy. At high literacy rates, there is high life expectancy but there are some instances where at about 83% literacy rate, set of instances of low expectancy. Plot also shows a saturation in literacy rates at high life expectancy. We also see, there can be high life expectancy even at low literacy rates.

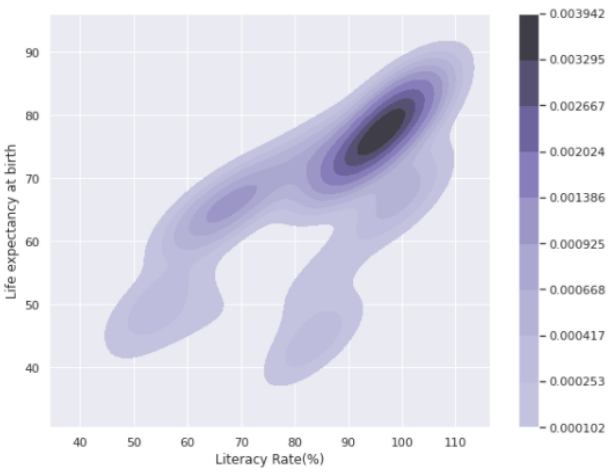

Fig. 25. Plot of Life Expectancy at Birth vs Literacy Rate (Adult)

N. Miscellaneous

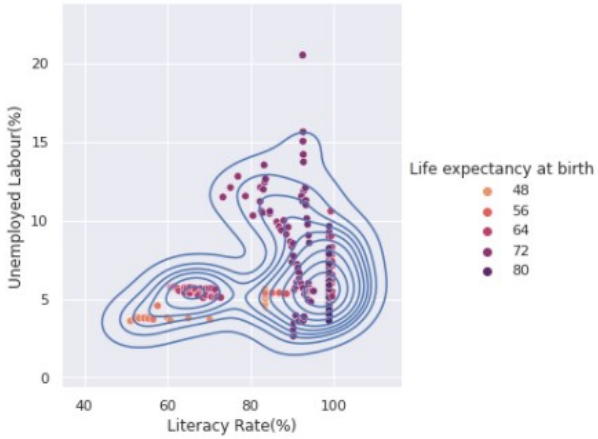

Fig. 26. Scatter Plot of Unemployed Labour vs Literacy Rate

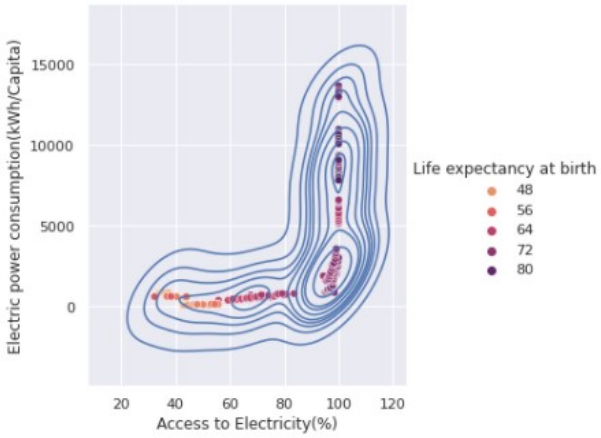

Fig. 27. Scatter Plot of Electric Power Consumption vs Access to Electricity

*1) **Scatter Plots**:* We can conclude from the scatter plots that interestingly as literacy rate increases life expectancy

increases but also the unemployment increases. We can see that as access to electricity increases, the power consumption increases almost exponentially and the life expectancy also increases quite rapidly.

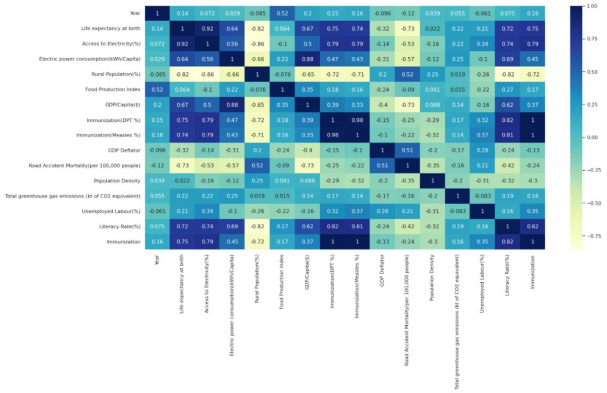

Fig. 28. Heatmap of variables

*2) **Heat Map**:* We observe from the heatmap that the variables 'Immunization (Measles)' and 'Immunization (DPT) are very highly correlated with a coefficient of 0.98, so we ignore these two and instead logically take a new variable, 'Immunization' and define it as mean of the other two variables to use in our modelling part.

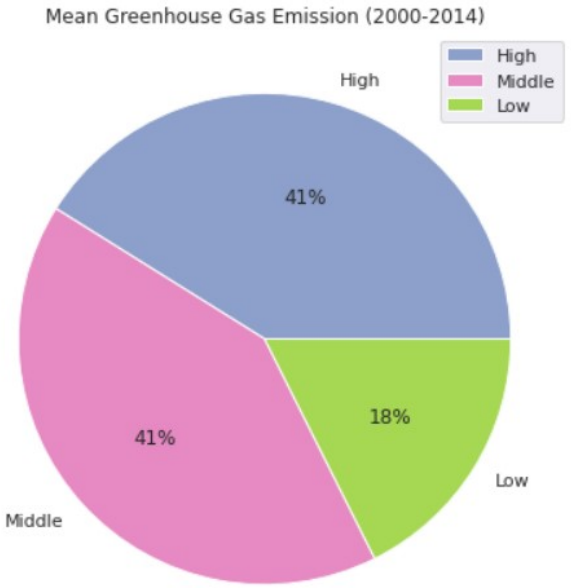

Fig. 29. Mean Greenhouse Gas Emission (2000-2014)

*3) **Pie Charts**:* A comparison of the mean greenhouse gas emission for the countries over a period of 15 years shows that the contribution of low-income countries is the least (18%), whereas the countries in the middle and high-income category contribute almost equally (41%) to the greenhouse gas emission. Energy consumption is by far the biggest source of human-caused greenhouse gas emissions and since this consumption is much more in the high and middle-income category compared to low-income countries, the greenhouse emission gets affected accordingly. Comparing the net electric

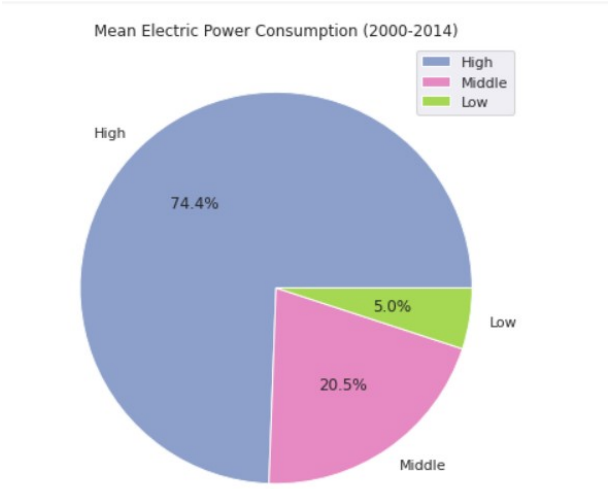

Fig. 30. Mean Electric Power Consumption (2000-2014)

power consumption for the countries in different income categories shows that the countries in the high-income category consume the most (74.4%), followed by Middle-income countries (20.5%), while the low-income countries have noticeably low consumption of electric power (5%). The reason behind this significant difference in the electric power consumption for the countries in different categories is the access and affordability of electric power.

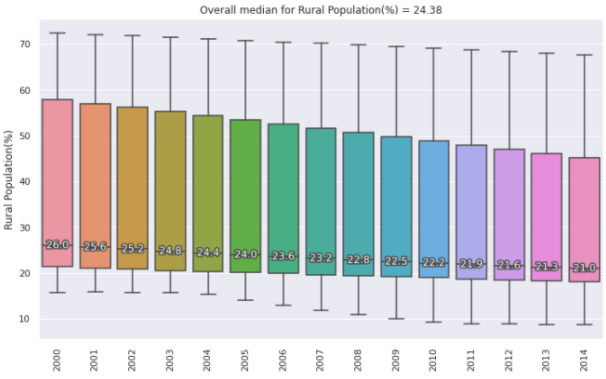

Fig. 31. Boxplot of Rural Population (%) (2000-2014)

*4) **Boxplots**:* We can see over the years average rural population population has been showing a decrease, which could not only mean more influx of people into urban areas but also that more and more rural regions have started to urbanize. Region beyond 3rd quartile of rural population does show a decrease but a decrease but is lesser than the decrease in median rural population percentage. Lower end of rural

population shows a massive decrease between 2004 to 2010, after 2010, there is a sort of stagnation on lower outliers. Notice that region between 2nd and 3rd quartile has a far larger spread than the region between 1st and second quartile.

Median life expectancy shows an increase over the years.

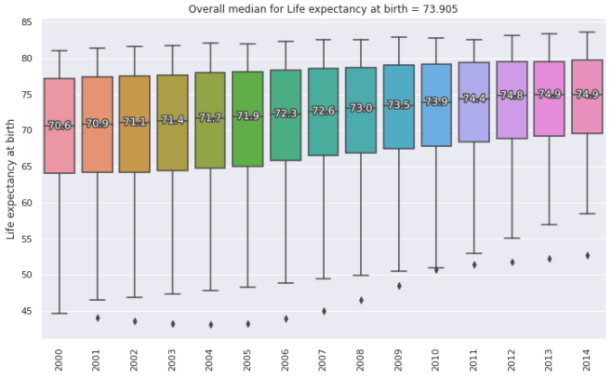

Fig. 32. Boxplot of Life Expectancy at Birth (2000-2014)

Notice however that end of region beyond third quartile shows local maxima in year 2009 after which we see a short period of decrease followed by a rise to a global maxima. Median is at an even location with the region between first and second quartile of approx same width as region between second and third quartile. Even lower end outliers have shown increase in life expectancy. The region between first and third quartile has shown decrease overall which means more and more people are going to die at similar ages. This increase in life expectancy could be attributed mostly to better health care facilities and more investment in healthcare RnD among most countries. We shouldn't attribute this increase to betterment of natural immunity of humans as 15 years is a very short period of time for increase in natural immunity among human species. There has been a very slight increase in median greenhouse gas

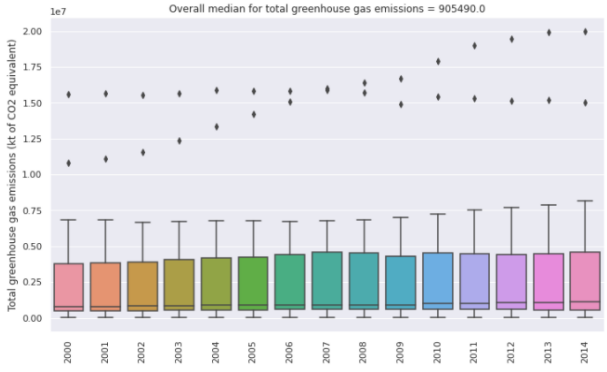

Fig. 33. Boxplot of Total Greenhouse Gas Emissions (2000-2014)

emission measured in term of kT of CO2 gas equivalent, each year. Despite slight rise in median greenhouse gas emission there are significant number of countries with almost null gas emission and they haven't shifted at all over the years. The region beyond third quartile being high, yet is nowhere close to some outlier countries which seem to produce enormous amount of greenhouse gas. Also, the average increase in greenhouse gas emission just among the outliers is higher than all other countries together. Same number of countries with gas emission above the median value are lot more spread out in their emissions than same number of countries with emissions below the median value. Graph is really pointing out the large disparity in greenhouse emission, despite this the effects of this will be mostly evenly spread across globe.

IV. Experiments and Results

A. *Linear Regression*

Linear Regression is the simplest model we use for predicting our target variable. Here we use Multiple Linear Regression because we have to predict a target variable using multiple explanatory variables. The basic equation for a MLR is as below:

$$y_i = \beta_0 + \beta_1 \cdot x_{i_1} + \beta_2 \cdot x_{i_2} + ... + \beta_p \cdot x_{i_p} + \epsilon$$

Here our target variable is y_i *(Life Expectancy)*, β_0 is *intercept*, x_i's are the different explanatory variables, β_is are weights for different explanatory variables and ϵ is the error term. The Loss parameter we used for this model were *MAE (Mean Absolute Error)*, *MSE (Mean Squared Error)*, *R2 Score (Coefficient of Determination)* and *RMSE (Root Mean Squared Error)*. By comparing these parameters for different split ratios of test-train we find the best split ratio and then predict the life expectancy using its coefficient.

We implement this model and find the best train-test split

	Fraction	Intercept	MAE	MSE	R2 Score	RMSE
0	90:10	68.551351	2.246135	9.376017	0.836655	3.062028
1	80:20	68.626982	2.283324	10.048181	0.840939	3.169887
2	70:30	68.631378	2.472308	40.846597	0.333171	6.391134
3	60:40	68.525274	2.382400	29.873105	0.510407	5.465629

Fig. 34. Scores of each train-test split for Linear Regression

to be 90:10. Also this is not a very accurate fit and other models were found to be much better, the possible reason for this would be that such complex dependencies can't be simply modelled as linear, which gives rise to errors.

We can see the coefficients of the linear regression to see on which variables the life expectancy has the highest dependence and if this dependence is directly or inversely proportional to life expectancy. From our model we can see that FPI has the highest dependency whereas Population Density has the least dependence on Life Expectancy.

B. *Decision Tree*

A Decision Tree is a supervised machine learning algorithm. They are algorithms that can be used for both continuous/non-continuous input variables. Here in our problem, we have the input as well the output variables are continuous, so we a Decision Regression Tree. In a Decision Tree the dataset is broken down into smaller and smaller subsets and a tree

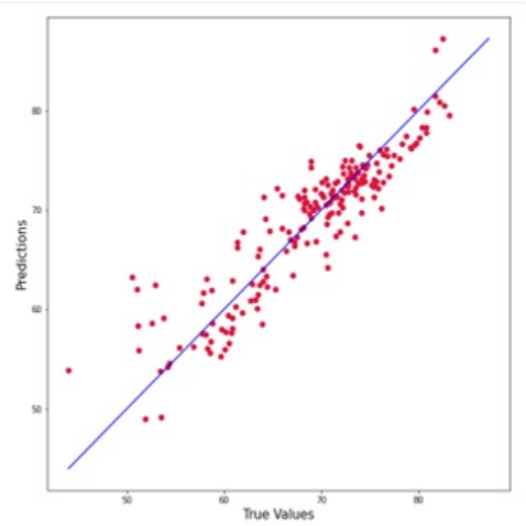

Fig. 35. Predicted vs Actual Value with reference 45 degree line for Linear Regression

Variable	Coefficient
Access to Electricity(%)	0.186622
Food Production Index	4.909636
Population Density	-0.008902
Unemployed Labour(%)	0.066662
Road Accident Mortality(per 100,000 people)	-0.478863
Electric power consumption(kWh/Capita)	-1.601477
GDP Deflator	-1.089480
GDP/Capita($)	0.069838
Literacy Rate Adult	1.673666
Total greenhouse gas emissions (kt of CO2 equi...	-1.026722
Rural Population(%)	0.283877
Immunization	-0.612290

Fig. 36. Weights for the Linear Regression model

like structure is developed with nodes and leaf nodes. For continuous data, the split is done for elements higher than a particular threshold. Also, if there could be non-linearity with respect to explanatory variables, Decision Tree would perform better than linear regression. We split the data in different test-train ratios and observe the best one and see its predictions. We find that the best split in this case is also 90:10 and the Decision Tree performs much better linear regression. We can conclude from this that the dependence of variables is non-linear and hence Decision Tree performs better than Linear Regression.

	Fraction	MAE	MSE	R2 Score	RMSE
0	90:10	0.677059	1.611242	0.971930	1.269347
1	80:20	0.783975	2.523343	0.960056	1.588503
2	70:30	0.923538	3.304409	0.946055	1.817803
3	60:40	0.954674	3.220411	0.947220	1.794550

Fig. 37. Scores for splits of Decision Tree model

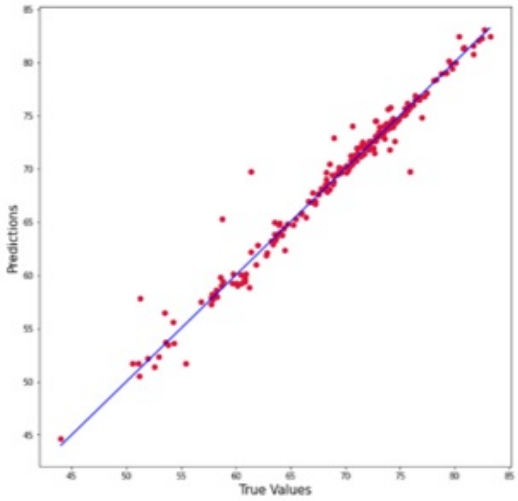

Fig. 38. Predicted vs True value for Decision Tree

C. *Support Vector Machine*

Regression is a version of Support Vector Machines. The model produced by support-vector classification depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Similarly, the model produced by SVR depends only on a subset of training data, because the cost function of building model ignores any training data close to model prediction. Model is trained by solving :

$$\text{minimize } \frac{1}{2}||w||^2$$

$$\text{subject to } |y_i - < w, x_i > -b| < \epsilon$$

where x_i is a training sample with target value y_i. The inner product plus intercept $< w, x_i > +b$ is the prediction for that sample and ϵ is a free parameter that serves as a threshold, all predictions have to be within an ϵ range of true predictions.

Original data has 2010 instances of organized by country and year. Before training we shuffle the data. Different data splits have been used among $60 : 40, 70 : 30, 80 : 20$ and $90 : 10$. We used grid search over a 5 fold cross validation across training dataset to find the best hyper parameters to train the *Support Vector Regressor (SVR)* over. We used multiple scoring criteria which included *Mean Squared Error (MSE)*, *R2 Scoring* and *Mean Absolute Error (MAE)*. Fortunately, over all scoring methods and all train-test splits there was a unique winner for hyper parameters. We made sure by multiple grid searches that the unique winning hyper-parameters are not

on edge of parameter grid. We got these as the best hyper parameters:

Kernel : Radial Bias Function (RBF)

C : 100

Gamma : 0.1

Epsilon : 0.1

We can see in below table which shows the different scores on test data for each split. We can see that on each scoring

Split	Mean Absolute Error	Mean Squared Error	R2 Score
60:40	0.607	0.913	0.986
70:30	0.491	0.653	0.989
80:20	0.504	0.628	0.990
90:10	0.519	0.736	0.986

Fig. 39. Scores of train-test split for Support Vector Machine

criteria except Mean Absolute Error, 80:20 split is giving best results, so we will be using 80:20 split for further discussion about model.

We can see in the below plot of predicted values vs actual values for the 80:20 train-test split that our model has a reasonable high accuracy as we see high density of points along the reference 45 degree line. Despite the high accuracy

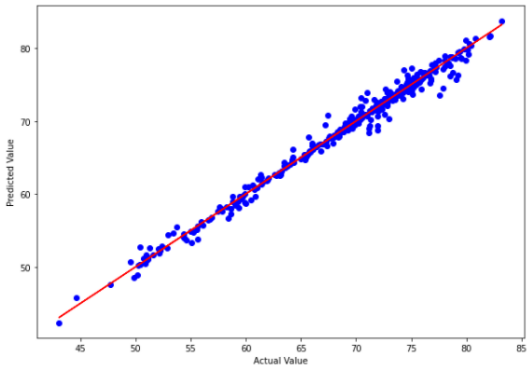

Fig. 40. Predicted vs Actual Value with 45 degree reference

which it seems from the scatter plot, we can take a look at a bar plot identifying the difference between predicted and actual value. This graph helps to magnify the shortcomings of Support Vector Regressor. We see a max offset of around 4 units from actual. Larger density of predicted values is in 65 to 80 range. On getting a mean of difference we see that an approx offset is just above 0.5, as calculated from mean absolute error which is at 0.504. Support Vector Machines work out considerably good, but as we see in neural networks model that we can further improve our accuracy.

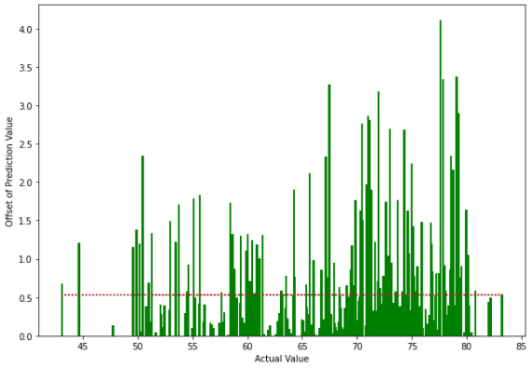

Fig. 41. Offset from Actual Value vs Actual Value

D. Neural Networks

We implemented Neural Network Regression using a *single hidden layer* and the *'Adam' optimizer* since it can handle sparse gradients on noisy problems. Since the number of important parameters in determining life expectancy is 13, the input layer consists of 13 neurons, while the output layer consists of a single neuron. Since the problem that we are solving is a regression problem, the output layer uses *'linear' activation*. The *'ReLu' activation* was used for the input and hidden layer mainly because of the non-saturation of its gradient. The learning rate was kept at its default value, i.e, 0.001. Since our aim was to minimize the mean square error, we split the data into various test and validation sizes and chose the one which showed the least MSE (Split 80 : 20) and avoided over fitting, by varying the number of neurons in the hidden layer, number of epochs, batch size and implementing early stopping. The graph below shows the variation of the

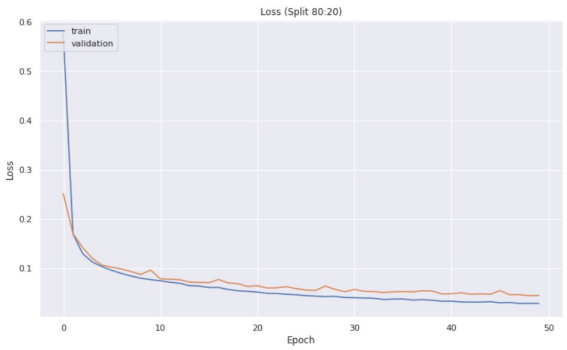

Fig. 42. 80 : 20 Split Loss vs Epoch

model loss (MSE) for training and validation set as a function of the number of epochs for the train-validation split as 80 : 20. It shows the least over fitting as opposed to the 70 : 30 split. The neural network model is as follows:

1) Train-validation split - 80 : 20
2) Loss function - MSE
3) Optimizer - Adam
4) Number of epochs - 50
5) Batch size - 25

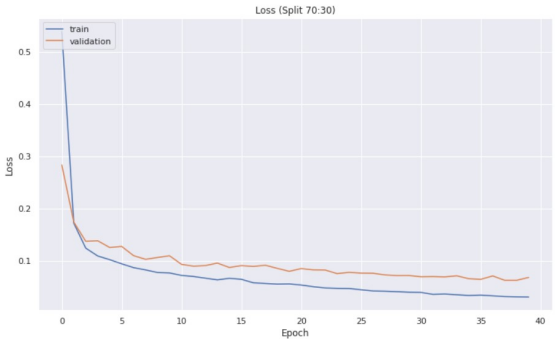

Fig. 43. 70 : 30 Split Loss vs Epoch

6) Learning rate - 0.001
7) Number of neurons in hidden layer - 200

```
Model: "sequential_13"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_39 (Dense)             (None, 13)                182
_________________________________________________________________
dense_40 (Dense)             (None, 200)               2800
_________________________________________________________________
dense_41 (Dense)             (None, 1)                 201
=================================================================
Total params: 3,183
Trainable params: 3,183
Non-trainable params: 0
```

Fig. 44. Neural Network Model

The MSE data for various splits can be seen in the stated table.

Split	**MSE (Training Data)**	**MSE (Validation Data)**
60:40	0.028	0.066
70:30	0.03	0.068
80:20	**0.028**	**0.045**
90:10	0.03	0.0493

V. Learning and Conclusions

The models that we used are Linear Regression, Decision Tree, Support Vector Machine and Neural Networks. We trained the respective models for different train-validation splits. Following is the summary of the best results for each of the above models:

Model	**Split**	**Error**
Linear Regression	90 : 10	3.06
Decision Tree	90 : 10	1.27
Support Vector Machine	80 : 20	0.63
Neural Network	80 : 20	0.04

The error as we see is least for Neural Network with the ratio of training to validation data as 80:20. Hence, we concluded that this model is the best for predicting life expectancy based on the different parameters that we considered.

We saw the dependence of life expectancy on various parameters and also compared the variation of all these variables with time over 2000-2014. Some of the variables have very high contribution in determining the value of Life Expectancy of a country like Food Production Index, Electric Power Consumption, GDP deflator and Greenhouse Emissions. This can be explained by the logic that these are the most important variables that reflect the country's socioeconomic status and standard of living. Similarly, we saw some interesting trends over time and for different countries for the value of variables. We saw that overall, Life Expectancy has been increasing, similarly more people have access to electricity, there has been a continuous decline in rural population across countries, the immunizations have been on a rise. We can observe global event impact through this analysis, for example, due to the Economic Recession across the world in 2008, we saw a huge rise in the value of GDP Deflator as well as unemployment during this period. We also saw that India was an exception to the phenomenon of increase in unemployment during these years of crisis. We observed a huge increase in greenhouse emissions over the years which explained the rapid rise in problems related to Global Warming.

We see a huge disparity of resources consumed and standard of living between the rich and the poor countries. This is cause for concern and needs to be looked into. The richer countries need to cut down on the quantity of resources they are consuming and start looking for options which have lesser negative impact on environment. Secondly, there is an urgent need to start funding, awareness campaigns and taking other necessary steps to increase the Life Expectancy of the poorer countries by enhancing their socioeconomic as well as healthcare infrastructure status.

VI. Contribution of Team Members

We split up that work between ourselves such that everyone gets to work on a variety of different topics. So, the EDA was split up such that each one of us takes up different variables and different types of required graphs for their analysis. Similarly, we used 4 models for predictions and each of us worked on the models they had chosen and later on we compared the models and saw which one gave the best results. Similar splitting of work was followed during report writing where each of us picked the equal number of variables and models and analysed them which were later merged. The work of making the video was done together. Formatting of report in Latex was done by Amritaansh. The writing of Abstract, Introduction and Conclusions was done by Aaditya and Sneha.

VII. References and Data

Data Repository : Github Repository

1) For collecting raw datasets : World Bank Data
2) General Python and ML Related Doubts : GeeksForGeeks and StackOverflow