

Indian Institute of Technology, Kanpur

Department of Industrial and Management Engineering

**Project Report on Prediction of Life Expectancy using Machine Learning Algorithm**

*Submitted By*:

Group 04

Aniruddha Tondare – 20114004

Harsh Jain - 20114007

Mansi Varshney- 20114010

Mohit Gupta - 20114012

Vishesh Jain - 20114023

***Contents:***

1. Problem Description
2. Data Understanding
3. Data Preprocessing
4. Modelling
5. Result and Interpretation

**1. Problem Description:**

To predicts the life expectancy of various countries using machine learning algorithm. Through this project we are trying to predict the life expectancy of various countries based on the data collected from the data repository website of WHO and its corresponding economic data was collected from the united nation website. During data pre-processing, we used various techniques to treat outliers and null values. After data pre-processing, we tried several machine learning models to correctly predict our target attribute. Best model was decided using r-square and root-mean-square as parameters.

**2. Data Understanding:**

We are having data of 193 countries, each country was having data from year 2000 to 2015 with attributes as population, BMI, Infant Death, HIV etc. with life expectancy as target attribute.

In terms of type of dataset, it is having the following characteristics:

* No. of rows: 2938
* No. of predicting variables: 21
* 1 ordinal variable, 2 nominal variables, 18 continuous numeric variables.
* Following is a subset of our dataset.

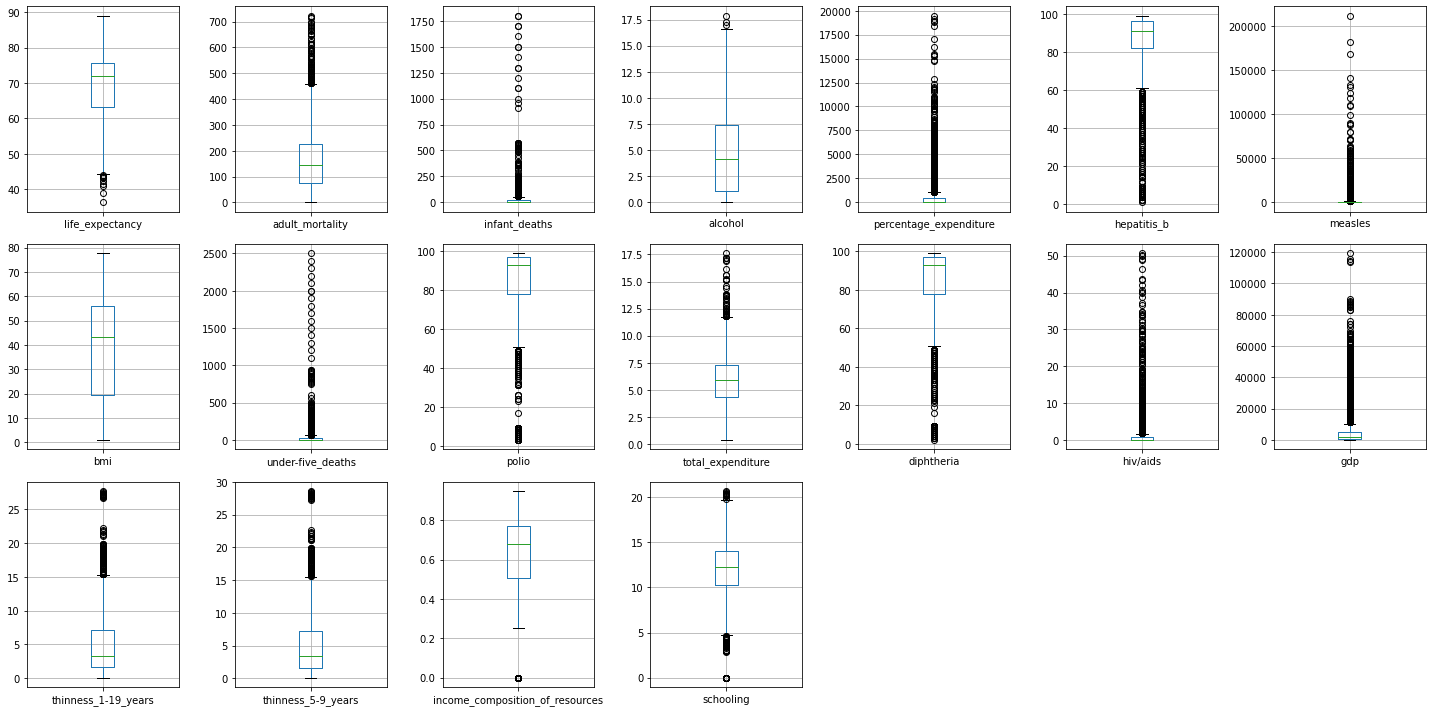
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | Year | Status | Life expectancy | Adult Mortality | infant deaths | Alcohol | percentage expenditure |
| Afghanistan | 2015 | Developing | 65 | 263 | 62 | 0.01 | 71.27962 |
| Burundi | 2004 | Developing | 52.6 | 378 | 24 | 5.72 | 11.22655 |
| Chad | 2013 | Developing | 52.2 | 366 | 46 | 0.64 | 76.52383 |
| Djibouti | 2002 | Developing | 57.9 | 322 | 2 | 1.03 | 13.46662 |

**3. Data Preprocessing:**

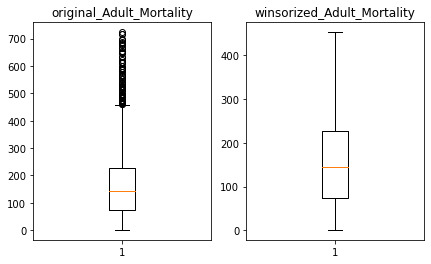
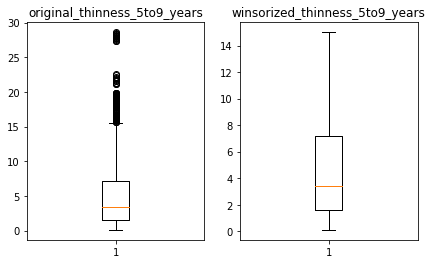
We started our data exploration to check for the consistency of our data. To avoid spaces and semicolon in our attribute name, we tried using underscore between each words to make it look more consistent. During exploration, we found few countries that were having data of only one year, we dropped those countries for further analysis. As per the data of descriptive statistics, we found a lot of variation in data values of other attribute as well. We were having outliers and missing values in most of the attributes.

***Treating Missing values***: We further analysed missing values and found that attributes like population and hepatitis are having maximum null values with 21% and 18% respectively. We used median imputation to treat null values. We used median imputation in such a way that each missing data for a particular year is filled by median of that particular year so that any specific value of a attribute gets normalized with other countries.

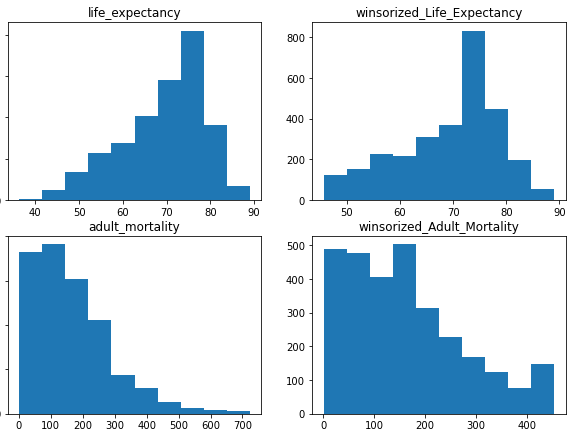
***Treating Outliers:*** On analysis on outliers from box plots, we found the following result,



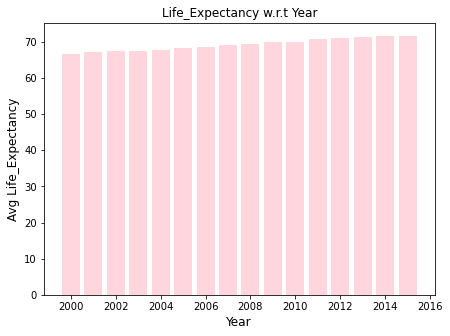
Outliers were treated using winsorization technique based on 1.5 IQR, a few example of this implementation is shown below,

In feature engineering, while observing distributions of attributes like total expenditure, income composition of resources and schooling are normally distributed. Alcohol, infant. Deaths, under five deaths, percentage. expenditure, Measles, Adult, Mortality, thinness 1-19 years, thinness 5-9 years, and HIV AIDS are distributed with skewness at right. Hepatitis, Polio, Diphtheria are distributed with skewness at left. BMI has bimodal distribution. Then, we tried to analyse the behaviour of our initial and winsorized data and we found more uniform distribution of winsorized in comparison to original ones. Here are some examples,



After taking care of outliers and null values we tried to analyse if there is any relation between life expectancy with the year and surprisingly there is a relation. Average life expectancy was found to be increasing with year.

****

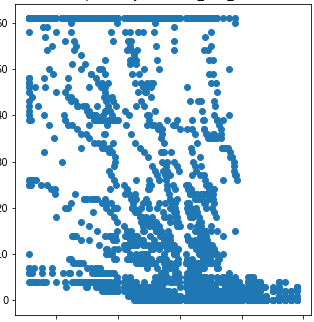
We also tried to see if there is any relationship between our target attribute and status of a country and it was found that life expectancy is higher for countries with developed in comparison to developing countries.

****

In developed countries adult mortality rate, thinness between 1 to 19 years, infant deaths and deaths from HIV are less in comparison to developing countries which ultimately leads higher life expectancy in developed countries. On the other hand, the other variables such as Income composition of resources, years of schooling, percentage expenditure and Total expenditure on health are more in developed countries.

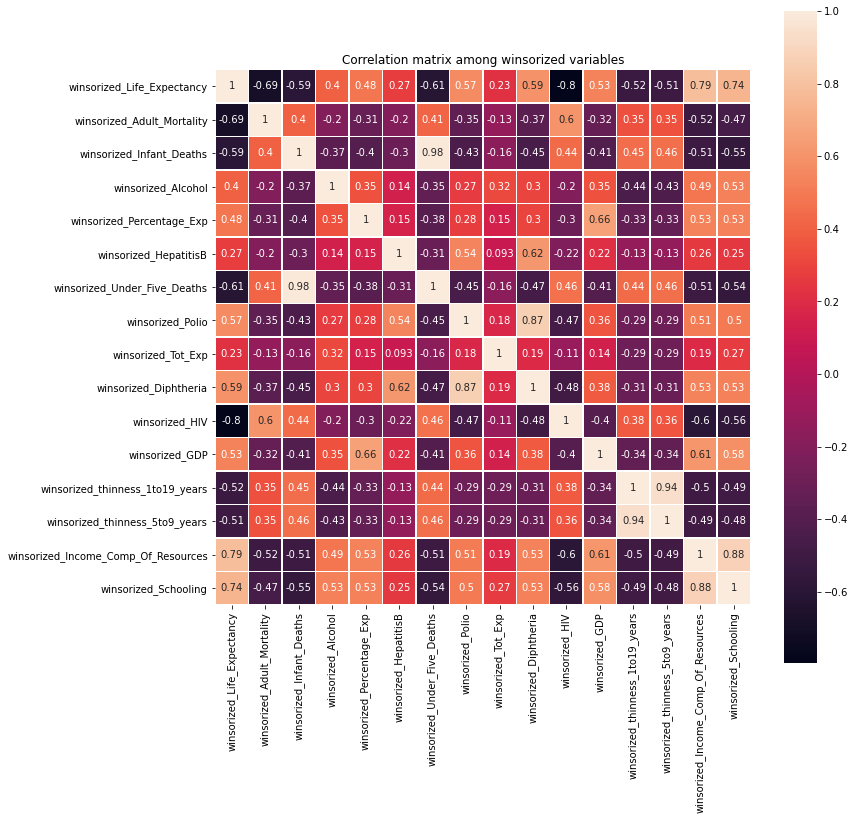
Further analysis was done with scatter plot between life expectancy and all other attributes. Some very exciting results of that is as follows:

1. income\_comp\_of\_resources and schooling are having almost same behavior towards life expectancy.
2. Same observation is observed between thinneess\_5to9\_years and thinness\_1to19\_years and Under\_five\_death and infant\_death

Life Expectancy vs Infants Death Life Expectancy vs under five death

With this behaviour, we looked for correlation factor between attributed using heatmap.



Inferences from Heatmap:

1. Based on the heat map above, we found correlation factor of 0.98, 0.94 and 0.88 for under five death vs. infant death, thinness\_5to9\_deaths vs. thinness\_1to19\_deaths and income comp of resources vs. schooling respectively.

2. Same behaviour was observed in scatter plot also.

3. We decided to drop one of the two attributes (based on more null values observed) that are having same behaviour towards target attribute,

* under five death among under five death or infant death
* thinness\_5to9\_deaths among thinness\_5to9\_deaths or thinness\_1to19\_deaths
* income comp of resources among income comp of resources or schooling

**4. Modelling:**

We divided data into two parts as training and test data in 7:3 ratio. Model was trained on 70% of the data randomly chosen and testing on remaining 30%. In each of the model we also tried modelling with and without the attributes we were planned to drop. No significant change in R2 value was observed in doing, so all our models are now trained without those three attributes. We started with linear regression to check if our data can be fitted in it as it is the simplest and can easily be interpreted and we had to predict a continuous value as target variable. With increasing complexity of model, we then chose decision tree regression as it gives better result with large dataset. We got quite satisfied r2 value in decision tree itself but we moved to more robust model, random forest regression. We chose it as it reduces the chances of overfitting and leads to more accurate results.

Model 01: Linear Regression

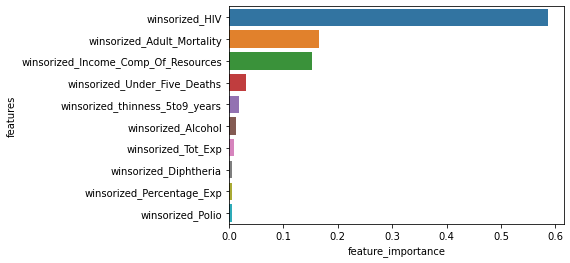
After decision tree regression, we tried linear regression where we find its r2 and MSE value as 0.84 and 13.71 respectively.

Model 02: Decision Tree Regression

We started with fitting our data to decision tree regression at first and observed their r2 as well as MSE value which was found to be 0.912 and 7.9 respectively

Model 03: Random Forest Regressor

In this model we found r2 and MSE values as 0.95 and 4.12 respectively. Feature Importance explains predictive power of features in dataset. Just showing that mining algorithms predicts well is not enough. So there must be some contributions of input data features as well.Therefore after evaluating these models we calculate Feature Importance score for all these attributes.



As can be seen, HIV has highest Feature Importance score of around 0.6.

**5. Result and Interpretation:**

All of three models were compared based on their R2 and MSE values, table of which is attached below. As can be inferred, random forest regression is giving best result among all three, so we finalize our model as random forest regression.

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No. | Model | R Squared | MSE |
| 1 | Decision Tree | 0.91 | 7.9 |
| 2 | Linear Regression | 0.84 | 13.7 |
| **3** | **Random Forest** | **0.95** | **4.1** |