STATISTICAL ANALYSIS ON FACTORS INFLUENCINGLIFE EXPECTANCY

Abhinaya.V[1], Dharani.B.C[2], Vandana.A[3], Dr.Velvadivu.P[4], Dr.Sathya.C.[5]

[4,5]Faculty, Department of Computing,
[1-3]Student, Department of Computing, Coimbatore Institute of Technology, Coimbatore.

ABSTRACT: *In this study we have analysed the various factors affecting life expectancy in order to suggest a country which area should be given importance to efficiently improve the life expectancy of its population. The main focus of our study is to determine the predicting factor which is contributing to higher value of life expectancy. The data have been collected from the WHO data repository website and its corresponding economic data was collected from United Nation website for the period 2000 - 2015 for 193 countries. The factors considered in the analysis are immunization, mortality, economic, social and other health related factors. A multiple linear regression model with twenty independent variables and life expectancy as the dependent variable is used to find the relationship between the variables.*

Further stepwise regression and cluster analysis algorithm is used in this study. The final result of the analysis would be the factors with significant influence on life expectancy indifferent countries.

KEYWORDS: Life expectancy, MLR, Stepwise regression, Cluster analysis, Significantfactors, Countries.

INTRODUCTION:

Life expectancy is the key metric for assessing population health and it refers to the number of years a person can expect to live[1]. The study of life expectancy of a population is important for the evaluation of the degree of economic and social development of a country[2]. The residents of a country with high life standards live longer, on average and have a small mortality ratio[2]. In the last decades, life expectancy has increased significantly, at global level. There have been lot of studies undertaken in the past on factors affecting life expectancy considering demographic variables, income composition and mortality rates. It was found that affect of immunization and human developmentindex was not taken into account in the past.

LITERATURE REVIEW:

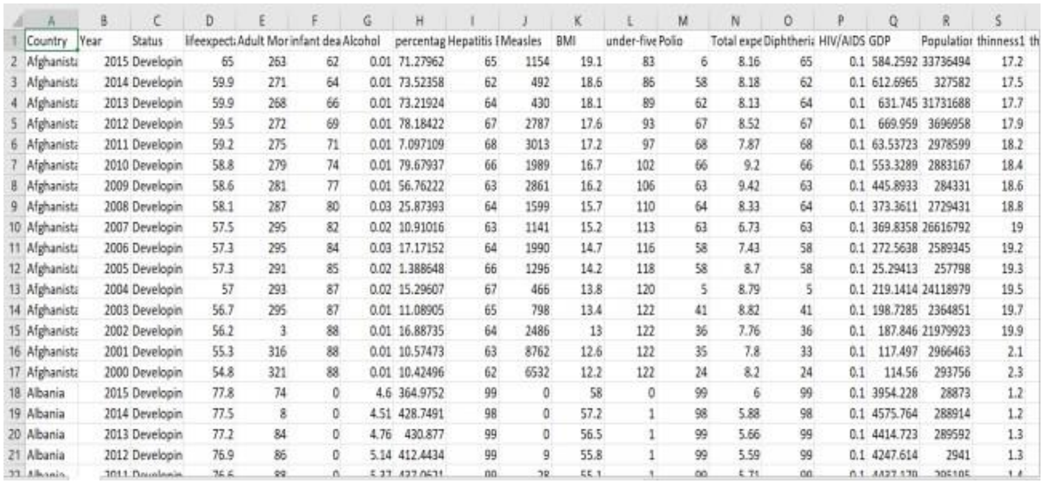
Several studies have been undertaken on factors affecting life expectancy, considering demographic variables, income composition, and mortality rates. Not only by studying cross sectional data but also panel data for a period of a few decades. The research questions wanted to assess thesignificant factors contributing to the improvement of life expectancy in the years and also understand whether life expectancy is influenced by socio economic and demographic inequalities factors. Life expectancy at birth is the average number of years a newborn is expected to live if mortality patterns at the time of its birth remain constant in the future. It summarizes the overall mortality rates for a population and the model that prevails among age groups in a given year.High mortality for young age groupssignificantly lowers the life expectancy at birth. But if people survive their childhood in a country with high child mortality, theymay live much longer. Therefore, in the model, a low life expectancy at birth may also be caused by high childhood mortality.

RESEARCH METHODOLOGY :

The data for the analysis was collected from the WHO[3] ,United Nations[4] and other authentic websites[5]. Data preprocessing was done using various methods. We started with the basic descriptive statistics and the count of null values in each column.The null values were filled with the mean values of their column. After computing null values ,a MLR model is built to find therelationship between the variables .We usedstepwise regression approach to fit data intoa model to achieve the desired result and cluster analysis to identify the homogeneous group of countries.Finally,we found the variables which are statistically significant.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

DATA ANALYSIS AND INTERPERTATIONS:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Country	Year	Status	lifeexpect	Adult Mor	infant dea	Alcohol	percentag	Hepatitis I	Measles	BMI	under-five	Polio	Total expe	Diphtheri	HIV/AIDS	GDP	Populatior	thinness1
2	Afghanista	2015	Developin	65	263	62	0.01	71.27962	65	1154	19.1	83	6	8.16	65	0.1	584.2592	33736494	17.2
3	Afghanista	2014	Developin	59.9	271	64	0.01	73.52358	62	492	18.6	86	58	8.18	62	0.1	612.6965	327582	17.5
4	Afghanista	2013	Developin	59.9	268	66	0.01	73.21924	64	430	18.1	89	62	8.13	64	0.1	631.745	31731688	17.7
5	Afghanista	2012	Developin	59.5	272	69	0.01	78.18422	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9
6	Afghanista	2011	Developin	59.2	275	71	0.01	7.097109	68	3013	17.2	97	68	7.87	68	0.1	63.53723	2978599	18.2
7	Afghanista	2010	Developin	58.8	279	74	0.01	79.67937	66	1989	16.7	102	66	9.2	66	0.1	553.3289	2883167	18.4
8	Afghanista	2009	Developin	58.6	281	77	0.01	56.76222	63	2861	16.2	106	63	9.42	63	0.1	445.8933	284331	18.6
9	Afghanista	2008	Developin	58.1	287	80	0.03	25.87393	64	1599	15.7	110	64	8.33	64	0.1	373.3611	2729431	18.8
10	Afghanista	2007	Developin	57.5	295	82	0.02	10.91016	63	1141	15.2	113	63	6.73	63	0.1	369.8358	26616792	19
11	Afghanista	2006	Developin	57.3	295	84	0.03	17.17152	64	1990	14.7	116	58	7.43	58	0.1	272.5638	2589345	19.2
12	Afghanista	2005	Developin	57.3	291	85	0.02	1.388648	66	1296	14.2	118	58	8.7	58	0.1	25.29413	257798	19.3
13	Afghanista	2004	Developin	57	293	87	0.02	15.29607	67	466	13.8	120	5	8.79	5	0.1	219.1414	24118979	19.5
14	Afghanista	2003	Developin	56.7	295	87	0.01	11.08905	65	798	13.4	122	41	8.82	41	0.1	198.7285	2364851	19.7
15	Afghanista	2002	Developin	56.2	3	88	0.01	16.88735	64	2486	13	122	36	7.76	36	0.1	187.846	21979923	19.9
16	Afghanista	2001	Developin	55.3	316	88	0.01	10.57473	63	8762	12.6	122	35	7.8	33	0.1	117.497	2966463	2.1
17	Afghanista	2000	Developin	54.8	321	88	0.01	10.42496	62	6532	12.2	122	24	8.2	24	0.1	114.56	293756	2.3
18	Albania	2015	Developin	77.8	74	0	4.6	364.9752	99	0	58	0	99	6	99	0.1	3954.228	28873	1.2
19	Albania	2014	Developin	77.5	8	0	4.51	428.7491	98	0	57.2	1	98	5.88	98	0.1	4575.764	288914	1.2
20	Albania	2013	Developin	77.2	84	0	4.76	430.877	99	0	56.5	1	99	5.66	99	0.1	4414.723	289592	1.3
21	Albania	2012	Developin	76.9	86	0	5.14	412.4434	99	9	55.8	1	99	5.59	99	0.1	4247.614	2941	1.3

Fig-1: Life expectancy dataset(WHO) [3] [4] [5]

The above table contains values of various factors which involves in life expectancy for different countries. It involves Mortality, infant death, Heaptitis, Total Expenditure, Schooling,,etc.

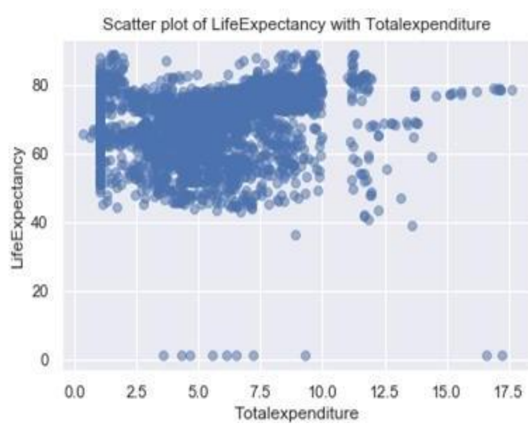
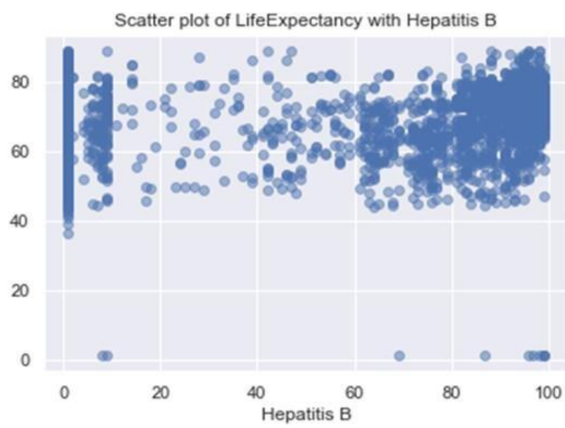


Chart- 1: Scatter plot of life expectancy with total expenditure

The above plot shows that the total expenditure has a positive relationship withlife expectancy. Growth in expenditure shows an increase in life expectancy.


Scatter plot of LifeExpectancy with infantdeaths
0
250
500
750
1000
1250
1500
1750
infantdeaths


Scatter plot of LifeExpectancy with Hepatitis B
80
60
40
20
0
0
20
40
60
80
100
Hepatitis B

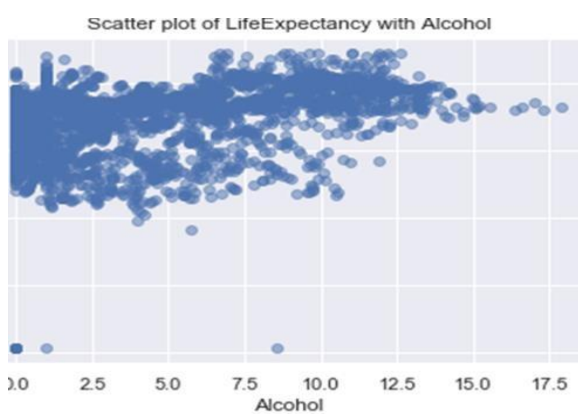

Scatter plot of LifeExpectancy with Alcohol
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Alcohol

Chart-3 : Scatter plot of life expectancy with infant deaths ,Alcohol ,Hepatits B , Measles

The scatter diagram of the observations between infant death rate and life expectancy shows that the two variables are inversely related.The plot between alcohol consumption and life expectancy shows that the two variables are not tightly associated. Also the plot between Hepatitis B and life expectancy , Measles and life expectancy shows that they are inversely related.

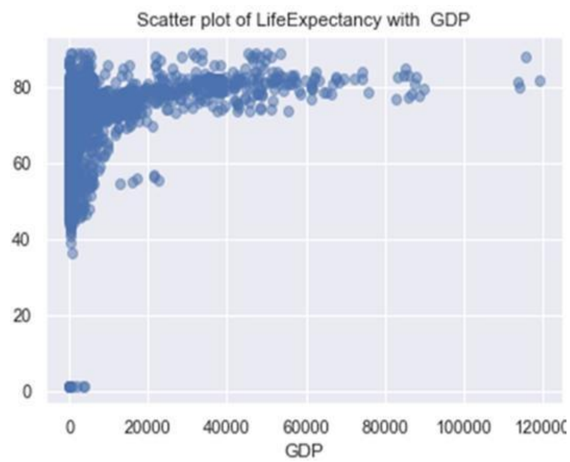


Chart- 4 : Scatter plot of life expectancywith GDP

This plot displays the correlation between life expectancy and gross domestic product (GDP) per capita. It shows that in general, countries with higher GDP tend to have a higher life expectancy.

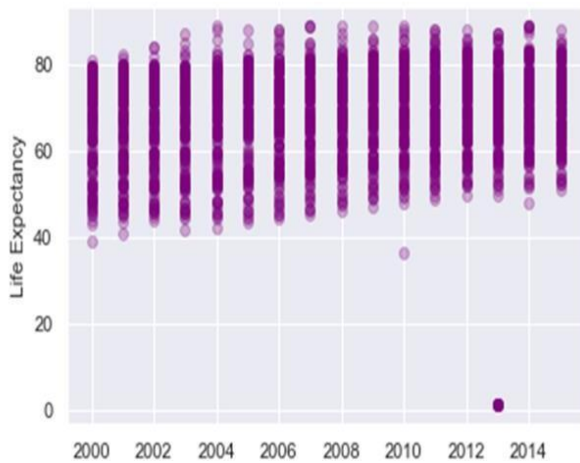


Chart- 5 : Plot between years and lifeexpectancy

The above plot shows that over the year, life expectancy has been increased slightly but it had a sudden downfall during the year 2012 and again it gradually increases in the year 2014.

```
[-2.99000851e-02 -3.38744739e-03  1.07570707e-02  6.55693667e-01
  9.29716767e+00  5.78701250e-05 -4.77933546e-02  4.47764798e-02
  2.28107566e-01]
```

Fig- 1 : Estimation of regression coefficient - Multiple linear regression

From this Multiple Linear Regression Model we infer that the independent variables like adult mortality has the negative value of -2.99000081e-02. As this value of adult mortality decreases the life expectancy rate increases. Similarly all the independent variables which has negative values is inversely related to the life expectancy. And the independent variable like Schooling which has a positive value of 6.55693667e-01, the life expectancy rate increases, that is they are directly related.

	Country	lifeexpectancy
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

TABLE- 1 : The above table represents thetop 10 country with the highest life

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        lifeexpectancy   R-squared (uncentered):          0.974
Model:                           OLS   Adj. R-squared (uncentered):     0.973
Method:                Least Squares   F-statistic:                 1.347e+04
Date:               Mon, 12 Apr 2021   Prob (F-statistic):               0.00
Time:                       08:41:44   Log-Likelihood:                -11305.
No. Observations:               2938   AIC:                         2.263e+04
Df Residuals:                   2930   BIC:                         2.267e+04
Df Model:                          8
Covariance Type:           nonrobust
==================================================================================
                                   coef    std err          t      P>|t|      [0.025
----------------------------------------------------------------------------------
Adult Mortality                  0.0217      0.002     14.168      0.000       0.019
Alcohol                         -0.3241      0.062     -5.221      0.000      -0.446
percentage expenditure          -0.0003      0.000     -1.267      0.205      -0.001
Polio                            0.2192      0.009     24.444      0.000       0.202
Total expenditure                1.1089      0.080     13.828      0.000       0.952
GDP                          -2.646e-05   3.83e-05     -0.690      0.490      -0.000    4.
Income composition of resources 33.1015      1.029     32.173      0.000      31.084
Schooling                        1.7613      0.057     31.015      0.000       1.650
==============================================================================
Omnibus:                     439.613   Durbin-Watson:                   0.812
Prob(Omnibus):                 0.000   Jarque-Bera (JB):             6883.065
Skew:                         -0.038   Prob(JB):                         0.00
Kurtosis:                     10.498   Cond. No.                     7.34e+04
------------------------------------------------------------------------------
```

Fig-8 :Step down regression - OLSregression results

From the above statistical summary we infer that finally there are eight variables left namely Adult Mortality, Alocohol, Percentage expenditure, Polio, Total expenditure, GDP, Income composition ofresources, Schooling and all of these variables are significant.

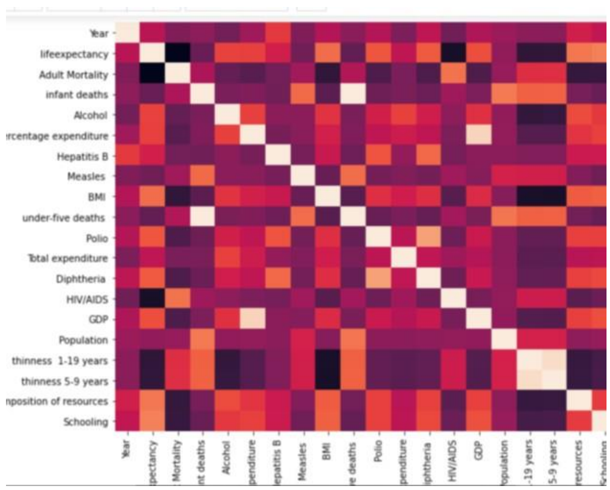


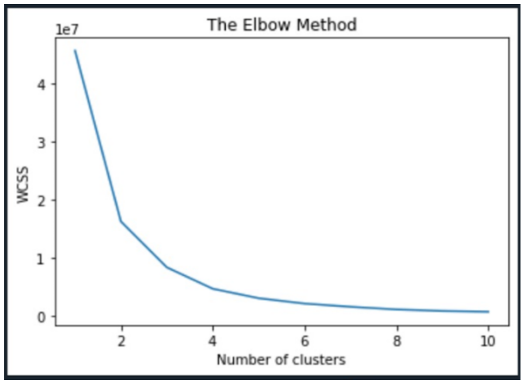
Fig-9 : Heatmap - plot for correlation

Evidence of negative correlation between Life expectancy and Diseases can be seen in the above heatmap. Whereas, Income and Schooling has a strong positive correlation with Life expectancy. Life expectancy has a positive correlation with vaccines of Hepatitis B, Polio, and Diphtheria which explains a higher life span for those who are vaccinated. Both infant and adult mortality have a negativecorrelation with Life expectancy which isalso very intuitive.

```
RMSE: 5.27
R_squared: 0.736
```

Fig- 10 : Statistical data of the Regression model fitted with the selected features.

From the above statistical data we infer that the Root Mean Squared Error indicates the absolute fit of the model to the data and the R squared value,0.736 illustrates the regression model explain almost all the variances of the variable and therefore there is very close relation between life expectancy and all other dependent variables.



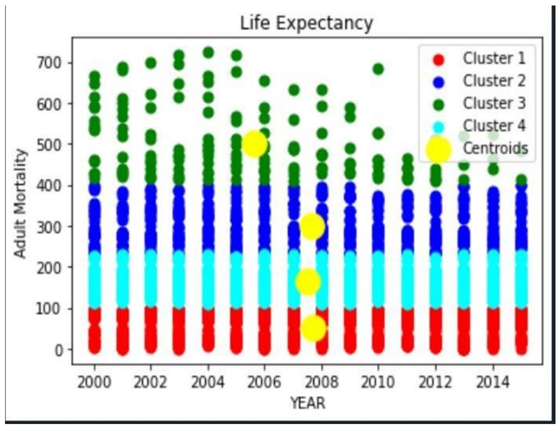


Chart-11 : Cluster analysis – plot

The elbow method is used to find the optimal number of clusters.The clusters give us the relationship between the life expectancy and adult mortality in each year.

CONCLUSION:

In conclusion we can say that the factors with significant influence on life expectancy in different countries are AdultMortality, Population, Percentage expenditure, Polio, Total expenditure, GDP, Income composition of resources and Schooling.So a country should give importance to these areas to efficiently improve the life expectancy of itspopulation.

INFERENCES:

- Over the year, life expectancy hasbeen increased slightly.
- Life expectancy decreases withincrease in infant's death.
- Countries having chronic diseaseshas lower life expectancies.
- GDP and Life expectancy has astrong linear relationship.
- Countries with high population hasslightly low life expectancy.
- As total income composition of a country increases, life expectancyalso increases.
- Schooling impacts the lifeexpectancy as expected.

REFERENCES:

[1] Max Roser, Esteben Ortiz-Ospina, Hannah Ritche , 2019, Life Expectancy,ourworldindata.org.

[2] Christiana balan, Elisabeta jaba, 2011,Statistical Analysis of the Determinants of Life Expectancy in Romania,RJRS Vol 5 No 2 Winter 2011

[3] World Health Organization data repository https://www.who.int/data/gho

[4] United Nations website https://data.un.org/

[5] Authentic website https://www.kaggle.com/kumarajarshi/life-expectancy-who.