MOI UNIVERSITY MAIN CAMPUS

SCHOOL OF SCIENCE AND AEROSPACE STUDIES

STUDY OF FACTORS AFFECTING LIFE EXPECTANCY AND PREDICTION OF LIFE EXPECTANCY FOR KENYAN COUNTIES.

GROUP MEMBERS

NAME	REG NO:
1. OBADIAH KIPTOO	AST/42/19
2. TERRY ACHIENG	AST/02/19
3. SILAS KIBET	AST/15/19
4. CECILIA OSIEKO	AST/04/19

THIS IS A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF APPLIED STATISTICS WITH COMPUTING IN THE DEPARTMENT OF MATHEMATICS, PHYSICS AND COMPUTING IN THE SCHOOL OF SCIENCE AND AEROSPACE STUDIES OF MOI UNIVERSITY.

APRIL, 2023

STUDENTS DECLARATION

This Project is our original work and has not been presented for a degree in any other University.

	NAME	REG:NO	SIGN	DATE
1.	TERRY ACHIENG	AST/02/19		
2.	CECILIA OSIEKO	AST/04/19		
3.	SILAS KIBET	AST/15/19		
4.	OBADIAH KIPTOO	AST/42/19		

SUPERVISORS DECLARATION

This project report has been presented for examination with the approval of the university supervisor.

Mr. Edwin Chirchir.

Signature Date

………………………………… ………………………………………………….

From department of mathematics, physics and computing.

DEDICATION

We dedicate this work to ourselves and our parents. We also dedicate to Dr. Ann and Mr. Charles Mutai, who taught regression analysis using R.

ACKNOWLEDGEMENT

We acknowledge the input and efforts of the group members (Obadiah, Silas, Terry, Cecelia) for their ideas and resilience in coming up with this project and ensuring that we reach the ultimatum for our study. We also appreciate the guidelines and analytical insights from our lead supervisor Mr. Edwin Chirchir. He has been instrumental and played a very critical role in ensuring that we took into consideration all the aspects required for a statistical project. We also acknowledge Mr. Charles Mutai, our lecturer who has always been available for consultation and guidance throughout the research process. We would like to take this opportunity to thank our school and department for providing us with the opportunity to conduct research on relevant societal issues as well as for the assistance throughout this period.

Table of Contents

TABLE OF TABLES

TABLE OF FIGURES

ABBREVIATIONS.

LE	Life expectancy
ML	Machine learning
PrLE	Predicted Life Expectancy Models
KIPPRA	Kenya Institute for Public Policy Research and Analysis
ARV’s	Anti Retro viral
WHO	World Health Organization
KNBS	Kenya National Bureau of Statistics.
HIV	Human Immuno Deficiency Virus
AIDS	Acquired Immuno Deficiency Syndrom
PC	Principal component
PCA	Principal Component analysis.
RF	Random Forest.

ABSTRACT

Life expectancy (LE) models have many effects on the social and financial systems. Many studies have suggested the essential implications of Life expectancy predictions on social aspects and healthcare system management around the globe. These models provide many ways to improve healthcare and advanced care planning mechanism related to society. However, with time, it was observed that many present determinants were not enough to predict the longevity of a set of population. Previous models were generated on mortality-based knowledge of the targeted sampling population. With the advancement in forecasting technologies and rigorous work from the past, researchers have proposed this fact that other than mortality rate, there are other factors needed to be addressed in order to come up with standard Predicted Life Expectancy Models (PrLE). As a result of this, now Life expectancy is studied with some additional set of interests into educational, health, economic, and social aspects. In this study we endeavor to build a model that uses predictor variables to predict life expectancy, study the effects of various predictors under study on life expectancy of a person. The methods that we employed in this study included visualizations, building machine learning (ML) models and the utilization of Pearson correlation coefficient which provided a visual representation of the degree of relationship between our variables, which formed a guiding basis while fitting our linear regression models. Apart from fitting multiple linear regression models we used also Random Forest, principal component regression, ridge regression, decision tree, and support vector machine to fit a predictive model in our data. We found out that life expectancy had a positive relationship with GPD, county expenditure on healthcare, BCG, Diphtheria, Hepatitis. Life expectancy had a positive relationship with adult mortality, under five mortality, HIV rate prevalence and population. We fitted the listed models and the best performing model was Random Forest model. We concluded that increasing health care expenditure and immunization coverage, leads to an increase in Life Expectancy in Kenya. We also concluded that a higher under five and adult mortalities, higher HIV prevalence, high poverty rate leads to a decrease in Life expectancy. We recommend that the Kenyan Government should allocate more resources to the health sector to ensure improved health care service provision to its citizens and ensure a hundred percent immunization coverage.

Chapter 1: INTRODUCTION

This chapter presents an introduction to the whole study which includes a survey of the main concepts in the topic coming as a background to the study, problem statement, general and specific objectives of the study, research questions and assumptions of the study. It will also present the justification and significance of the study, scope and delimitation, definition of special terms, theoretical framework and conceptual framework.

1.1 Background of the study

Life expectancy refers to the number of years a person can expect to live. Life expectancy acts as a measure to gauge the health status of a community. Human Life expectancy has increased during the last century. There is no doubt as this marks a great achievement of our civilization wave. Life expectancy is still affected by the changes occurring in age structures, this has caused a lot of problems for many social institutions. For instance, insurance companies signing agreements with their clients for the retirement rents are interested in correct prediction of such demographic changes (Theory Biosci.2003). Apart from social institutions changes in life expectancy affects fertility behavior, economic growth, intergenerational transfers and incentives for pension benefit claim (Zhang, J., Zhang, J., and Lee, R. D. 2001).

This study deals with the creation of a model that can be used to predict the life expectancy for individuals born in Kenya using specific predictor variables that we dim to be of great influence on the expected average number of years of life of the citizens of Kenya. The study covers the entire population in Kenya with the data segmented in terms of all counties in Kenya. The study is of great importance since life expectancy can be used as a measure that is used to mostly check on the overall health of the various counties. The changes in life expectancy are also used to explain trends in mortality.

The Government of Kenya will therefore be able to plan and provide services and support to its citizens by the use of the model. Improved life expectancy leads to an increase in the number of older members living in the country. The government should be made aware of this trend in order to provide for the dependent elderly members who require aid in undertakings of day to day living.

In the past, the government has been facing challenges in the disbursement of resources to various section of the health sector in order to improve the life expectancy of the citizens.

The model that we build will be used by the government to know the correct proportion of resources to allocate to the various sections for example in immunization so as to prevent some killer diseases that are immunizable.

The expected average number of years of life of Kenyan citizens was 25 years in 1870. The main cause of the low life expectancy was the epidemics and famine that occurred in 1800 through to 1860. There was loss of life of more than a quarter of the population of Kenya that resulted from an epidemic in 1897 and famine during the same period. According to((Theory Biosci.2003). The 1919 Spanish Flu epidemic led to the loss of more than six percent of the country's population hence lowering even further the life expectancy during that period. There was a significant improvement in the life expectancy in Kenya after the end of the Second World War. There was minimal loss of lives after the Second World War attributed by the peaceful coexistence. According to (Cardona C, Bishai D,2018) on the gains of life expectancy since 1950 the Kenyan life expectancy continued on an upward trend for much of the 20th century attributed by the rollout of universal healthcare in 1960s. There was a significant decline in the life expectancy from late 1970s to around 2011 as a result of the HIV/AIDS epidemic which had led to a significant increase in loss of lives in the Kenyan population. The life expectancy began to improve in the early 2010s due to increased funding for research on HIV/AIDS that led to the production of ARVs drugs that aided in management of the disease therefore reducing the loss of lives during that period. The life expectancy in Kenya in 2022 is estimated to be about 67 years.

This study will analyze the effect of various predictor variables on the life expectancy in Kenya. It will also aid in determining which predictor variables greatly affect the life expectancy so that they can be used in building the model to be used by the government in improving the health sector and in planning for the distribution of resources.

According to a report by the Kenya Institute for Public Policy Research and Analysis (KIPPRA), the life expectancy in Kisumu, Siaya and HomaBay is around 40 years under the current economic, social and health conditions. This shows a lower life expectancy as compared to drought and famine prone areas that are also insecure and remote such as wajir, West Pokot and Baringo. The increased number of people contacting HIV/Aids leads to the reduction in the life expectancy in those areas.

In view of the ever-declining life expectancy in Kenya, proportional allocation of resources by the government to various health sectors remains a gap that requires attention. This therefore opens the research gap that this study is interested in studying.

1.2 Problem Statement

The citizens of Kenyan tend to have a shorter life span over the years as a result of factors such as high levels of poverty, increased crime rate, excessive alcohol intake, diseases, bad weather, illiteracy and increased infant mortality rate. The government through the ministry of health and ministry of treasury has been stepping up its effort in bid to improve the life expectancy in Kenya.

The government has been releasing more funds to the counties since the health sector is one of the devolved units of the government. However, this has been an exercise in futility since there is continued steady decline in life expectancy in the country. The government has all along been allocating funds to various heath sectors that have minimal or no effect at all to the life expectancy of its citizens. This leads to wastage of public funds and resources in efforts to curb the factors with less to no effect on life expectancy.

Therefore, there is need for the government to get to know what factors affect the life expectancy in Kenya and to what extent. This will be done by the implementation of the model that we build that will predict the life expectancy based on the factors that we identify to have great effect. This will then aid the government in making proper, proportional and adequate allocation of funds to curb the factors identified to affect the life expectancy according to the degree of its effect. The government will therefore end up saving public money and at the same time improving the life expectancy.

1.3 Research objectives

1.3.1 General objective

Building a model that uses assigned predictor variables to predict life expectancy across Kenyan counties.

1.3.2 Specific objectives.

1. To find out if increasing healthcare expenditure will improve life expectancy.
2. Determining how infant and adult mortality rates affect life expectancy.
3. To study the impacts of HIV prevalence on life expectancy.
4. To find out the impact of immunization coverage on life expectancy.
5. To find out the impact of poverty on life expectancy.

1.4 Research questions.

1. Does improving health care expenditure improve life expectancy?
2. How does infant and mortality rates affect life expectancy?
3. Does HIV prevalence affect life expectancy?
4. Does immunization coverage affect life expectancy?
5. Does poverty affect life expectancy?

1.5 Research Assumptions

The following are assumptions for our study:

1. The Government has been stipulating minimum amount of funds towards healthcare operations during budgeting.
2. Immunization coverage within our counties is still low and needs improvement.
3. Poverty impacts the life span of people in Kenya.
4. Highly populated counties are faced with quite a number of challenges which affect lifespan of people.
5. Lifespan is of humans is affected by a wide range of factors that can be controlled.
6. Reported cases of infant and adult mortalities impacts lifespan of humans.

1.6 Justification of the study.

This study tries to determine the factors affecting the life expectancy with an aim of finding lasting solutions that can be enacted in order to increase the life expectancy of people and also their productivity. From our assumptions it has been argued that, the government has not in its operations been able to allocate and monitor the most critical aspect of its population which is healthcare. Laxity has been reported in this sector considering the minimal support in terms of funding and provision of incentives to health care practitioners in order to improve the health of its growing population. Government intervention would proof to be a most effective and long-lasting solution to this major problem, this is simply because the vast majority of people living in the country cannot quiet afford the better but expensive health care services which are offered by the private health care facilities.

This study also tries to provide solutions by assessing whether the existing strategies on vaccination, HIV/AIDS cases management, provision of affordable education for all, if they need to be strengthened or improved in a number of ways so that it can, in the long run ensure that the people live in an environment that enables them to grow and meet their goals.

The study is going to be important in a number of ways. The outcome from the study will always be available to the general public, NGO's and any other body that tries to study and build on this study on the life expectancy of the people of Kenya and factors that affects it. The study hopes to be a supplementary if not base for other studies that are to be conducted. Secondly, it is anticipated that the findings from the study is going to influence decision making efforts by policy makers, government while providing for its citizens, stakeholders and the general public consumption.

1.7 Scope and delimitation of the study.

The scope of this study is about determining factors affecting the life expectancy of people across Kenyan counties. This study will be limited to answering the earlier stated research questions. It will analyse the existing measures like existing health care spending strategies, immunization coverage, lifestyle of people and finding solutions to each of these problems is the ultimate goal of the study.

Geographically the study is going to cover counties which make up the country Kenya.

Chapter 2: LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of literature and information on previous studies on factors affecting life expectancy. Key theories relating to the study variables are also introduced. Health care expenditure, mortality rate, population and immunization coverage are among the factors to be discussed.

According to (Roser, Ortiz-Ospina, Ritche,2019) Life Expectancy refers to the number of years a person can live . Many studies have been conducted in the recent study with an aim of studying life expectancy. The study of life expectancy is important for the evaluation of degree oof development that a county has attained (Balan, Jaba,2011). Taking into consideration the demographic factors, income composition in the past and mortality rates recorded in the population, many researchers have been able to make projections about the life expectancy for a given population according to (Abhinaya, Dharani, Vandana, Dr. Velvadivu, Dr. Sathya,2021).

The stated research questions in our study wanted to assess the and draw an understanding on what factors under assumption affect life expectancy for a given population, is it the socio economic activities together with demographic inequalities. According to (Abhinaya, Dharani, Vandana, Dr. Velvadivu, Dr. Sathya,2021) life expectancy at birth for a given population is the measure of the average number of years that a newborn is expected to live if mortality patterns at their birth remain constant throughout their growth cycle. It in the long run summarizes the overall mortality that prevails for a population and the model that will suffice amongst the age groups in a given year. High mortality for the young ones significantly leads to a decline in life expectancy at birth. So if people survive their childhood it does not necessarily mean that they will not have a long life expectancy but they may live longer than expected. Model performance for a model that has low life expectancy at birth may be caused solely by high childhood deaths (Abhinaya, Dharani, Vandana, Dr. Velvadivu, Dr. Sathya,2021).

In our study we included Poverty rate across the counties with an assumption that it is one of the factors that impacts the metric life expectancy. According to World Bank (2006) there are three predominates of poverty in Indonesia namely those in vulnerability of falling poor, poverty incomebased om income only poverty characterized by the region of origin. According to Sholch (Khomsan, Dharmawan, Sukandar, & Syarief, H. 2015), poverty is the inability to meet basic consumption needs and improve living conditions, lack of business opportunities, to a broader understanding that covers moral and social aspects. According to (Morduch,1994) poverty is the people's inability to fulfill the living standard. According to the study done by Widodo (Widodo, Dewata, Umar, Putra, 2021) on data collected from the Indonesian population, after fitting a linear regression model to assess the impact of poverty (independent variable) on life expectancy (dependent variable) they were able to come up with a model Y=73.995 – 0.764(Poverty). Poverty impacted negatively on life expectancy for this particular set up. They also concluded that Poverty cannot be attributed specifically to unemployment as streams of income vary according to the social classes. Some regions recorded a lower life expectancy but still with a high stream of income per household.

In our study we also factored in vaccination coverage as one of the factors affecting life expectancy. According to (Rappuoli,2014) Vaccination since its inception has been the most effective medical intervention, and has alleviated millions of deaths that would have otherwise occurred. According to (Rappuoli, 2014) a recent study points out to 40 million cases of diptheria, 35 million cases of measles, and a total of 103 million cases of childhood diseases that have been prevented since 1924 in United States of America (USA). According to (World Health Organization, 2013), 2.5 million deaths per year are prevented through vaccination. These are positive results and figures that indicate the positive impact vaccination has had in reducing mortality rates that would have led to decline in life expectancy during childhood stages (Rappuoli 2014).

In the recent past many techniques have been employed in estimating life expectancy. Most common and adopted methods have been Machine learning. According to (Pisal,2022) studies on life expectancy ae mainly based on labelled data in which supervised approaches are of more relevance in model building and prediction. Further studies on tree-based classification approaches have been used in many life expectancy studies (Karacan *et al*, 2020;Meshram,2020;Vydehi *et al.* (2020). A most recent study conducted by (WHO) dataset using a decision three revealed that 9 out of 25 attributes significantly influence life expectancy. The study revealed using three regression models, Linear Regression, Decision Tree Regressor and Random Forest, found the Random Forest Regressor was the best model R2=0.99 (training) and 0.95 (testing), with 4.43 and 1.58 as the mean squared error and mean absolute error (Rahman *et. al* ;2022). For the Tree based models, Random Forest and Random Tree the performance analysis is conducted against their Roots Mean Squared Error, Relative Absolute Error and Receiver Operation Characteristic (ROC). ROC is the ability of the test to perform the classification correctly. The values of ROC area between 0.7-0.8 are acceptable, 0.8-0.9 excellent and above 0.9 is outstanding (Hosmer & Lemeshow, 2000).

Chapter 3: METHODOLOGY.

3.1 Introduction

This chapter describes the research design and methodology that was used in the study. The aim of the study was to develop an explanatory model to account for the factors that contribute to the life expectancy for people living in Kenya counties.

3.2 Methodology

For this particular study it is necessary to ensure that the objectives and essential theory behind the problem are clearly defined. Every person needs to know the practical aspect of our problem and how it will be implemented in real time.

Research design

This part gives the conceptual structure within which the research is conducted; it comprises of the overall plan for the collection, measurement and analysis of data. For our study we adopted an analytical research design. Analytical research involves use of secondary data to analyze the information in order to draw inferences.

Data

Our study research utilizes secondary source of data. This secondary data is obtained from the Kenya National Bureau of Statistics (KNBS), KDHIS. R, tableau and EXCEL were the main statistical software's employed in the analysis of this data. Spatial files for the map of Kenya by counties was also sourced from KNBS.

With the vast growth in data science and ML methods, there quite a number of Integrated Development Environments that can be used, but for our study we chose to utilise R language using R studio.

The following factors given below are some of the parameters we will use in estimating life expectancy.

Attribute	**Description**
County Expenditure	Revenue allocated by the county Governments to the health sector.
Population	Population of people in the respective counties.
GDP	It is an estimation of price of all products and services created in a particular tear in the country
Adult Mortality rate	Depicts the death of adults per one thousand adults.
Poverty rate	Represents the poverty rates at various counties,
Life Expectancy	Represents the average life expectancy of individuals in various counties
Under-five mortality	Represents the number of total deaths of children after birth till 5 years per 1000 number of deaths.
Vaccination (Diptheria, Measles, Polio, BCG)	Represents the vaccination coverage for the included vaccinations.

Table 3.1;Table of attributes.

The objective of this study is to predict the result of the number of dependent variables in comparison with the independent variable. After comparing various research methods, we intend to use the Graphical data analysis, Pearson Correlation, multiple linear regression and Random Forest modelling technique to analyze the various influencing factors on life expectancy. The Pearson Correlation Coefficient provides a visual comparison of the degree of correlation between our factor under investigation (life expectancy) and other variables which will provide a basis for the development of predictive models.

3.2.1 Graphical data analysis

This involves the use of visual representations (graphs) to uncover insights in our data. Graphical data representation can include the use of histograms, Pie charts, bar charts, line graphs scatter plots etc.

According to Graphical data representation is powerful in the sense that it complements with other traditional data analysis methods. For example, for regression we can display the regression line in order to understand the nature of relationship between the independent and dependent variable, we can also understand some data properties such as normality of our data.

3.2.2 Correlation

Correlation refers to a statistical measure that shows the extent to which two variables are linearly related (This means they change together at a constant rate). For our study we employed the use of Pearson Correlation coefficient because the variables are normally distributed so a parametric test works well. Pearson correlation is the most common correlation method it corresponds to the two covariances normalized (i.e., divided by their standard deviations) (Makowski,2020; Martan,2020; Patil, 2020; Ludecke:2020).

$$rxy = \frac{cov\ (x,\ y)}{SD_x \times SD_y}$$

For our case we will get our correlations using R. It will not be a single correlation for Life Expectancy and other single variables, but we will get a dataframe of correlations of other variables with Life Expectancy. The function cor () in R outputs all the correlations in a data frame for a provided dataset. For an enhanced understanding we will use the package corrplot (). This will help us output a correlation plot containing correlations between the variables under study.

3.2.3 Modelling.

3.2.3.1 Multiple Linear Regression

According to (Kassambara,2018) Multiple regression models are widely applicable and commonly used in life expectancy research to extract important information from a large amount of raw information and to mathematically model the relationship between variables so that the value of the dependent variable can be determined from the value of the independent variable. As life expectancy is influenced by a number of factors, the multiple regression model is of great practical significance and is more suitable for exploring the specific relationship and the degree of influence between Multiple linear regression: $Y=B_0 + B_1 * X_1 + B_2 * X_2$ where:

- Where B_0 is the intercept
- B_1 is the coefficient associated with the predictor variable X1
- B_2 is the coefficient associated with the predictor variable X2

For a linear regression model to suffice it must meet the following assumptions:

1. Normality
2. Homoscedasticity of errors
3. Independence of observations
4. Linearity- The true relationship is linear

3.2.3.2 Random forest (RF)

This is an ensemble method, which consists of decision trees for classification and prediction problems. The final output is the average of the prediction values of the individual trees. This method gives a better and accurate result which is caused by overfitting the training set. Correlation between the individual models is key in this method. This is because decision trees are extremely responsive and competent to the data. The random forest allows individual trees to randomly sample from the dataset with replacement. For classification models we cannot be able to get the model coefficients but rather, we can get to know the variable importance by getting the importance in data frame or even using carets powerful dependencies to get the importance plots.

The random forest classifier.

It consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the RF spits out a class prediction and the class with the most votes become our model's prediction (Yiu,2019). With the below model the prediction is 1 as it is the dominant model.

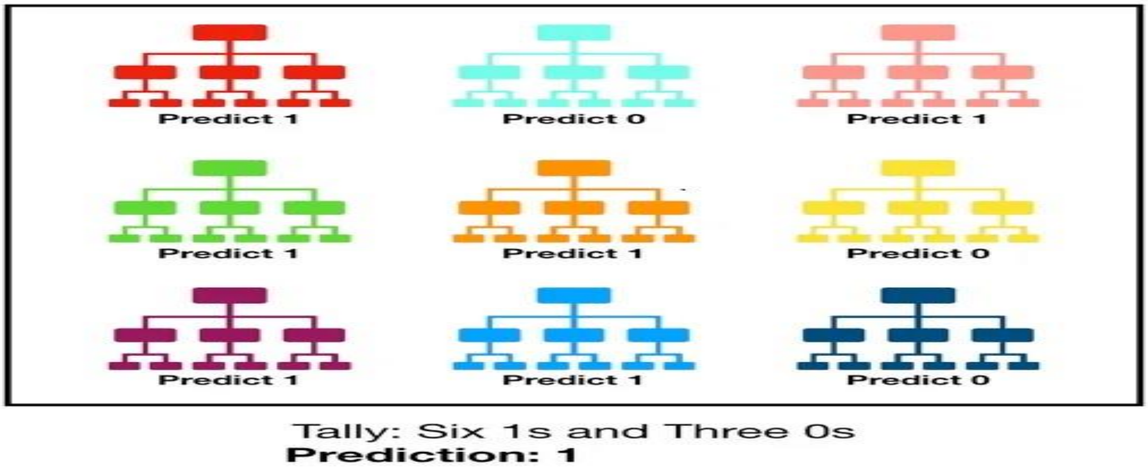


Figure 3.1; Random forest classifier sourced from Towards data science web.

In data science, the reason behind the powerfulness of RF is that *"A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models." Wisdom of crowds.*

3.2.3.3 Principal component regression

It is a regression technique that serves the same purpose as standard linear regression model. The difference is that PCR uses the principal components as the predictor variables for regression analysis instead of the original features. It works in three steps:

1. Apply PCA to generate principal components (PCs) from the predictor variables, with the number of principal components matching the number of original features P
2. Keep the first K PC's that explain most of the variation (where K<P) and K is being determined by cross-validation.
3. Fit a linear regression model by using ordinary least squares on the generated K PC's.

The idea behind it is the smallest numbers of PC's represent most of the variability in our data and the relationship with the target variable (Leung,2022). Therefore, instead of using the whole features of regression we utilize only the subset of the PC's.

Importance

1. It helps reduce overfitting by fitting a linear regression model on k components rather than all the components.
2. It helps in elimination of multicollinearity by removing PCs associated with small eigen values.
3. The performance improvement is more prominent in data with many features (high cardinality) and significant multicollinearity.

3.2.3.4 Support Vector Machine (S.V.M)

SVM is a supervised ML technique whose concept revolves around finding a marginal hyperplane that best separates the dataset in random space into classes i.e., the hyperplane having the possible maximum margin between the support vectors of the given dataset. The hyperplane is found in n-dimension space (n is the feature numbers) distinctly classifying the data points. SVM can be utilized for problems of both regression and classification (Suthaharan, 2016). For linearly separable dataset, a linear SVM classifier can be used while for the nonlinear data a nonlinear classifier has to be used to find the hyperplane.

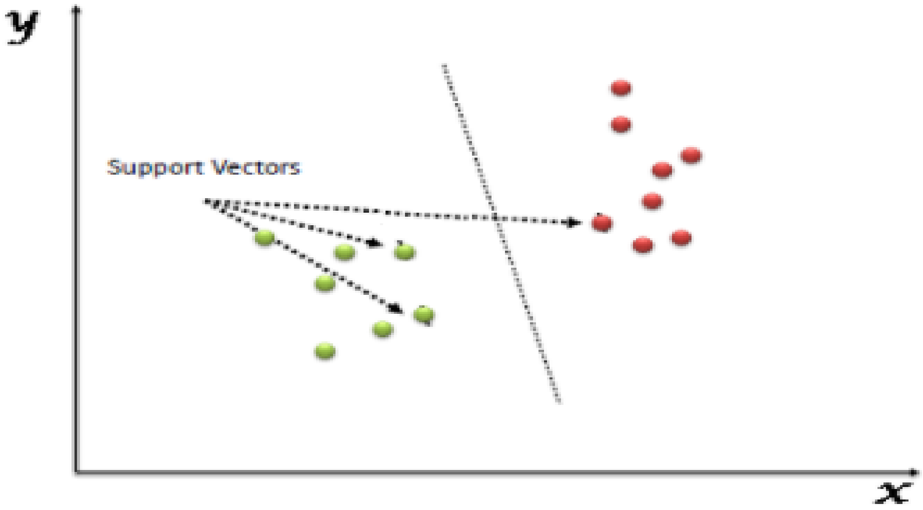


Figure 3.2; SVM linear kernel sourced from Towards Data Science web.

In practice, the S.V.M algorithm is implemented using kernel. A kernel function maps data to a higher dimension space making non-separable problem separable and it also maps data into better representation space (Suthaharan, 2016).

Kernel functions can be classified as [3]:

Linear Kernel

The Linear kernel is the kernel's simplest function. It is provided by the inner product <x,y> together with an optional constant c. Algorithms utilizing a linear kernel version are usually equivalent to the non-kernel ones. For instance, KPCA with linear kernel is similar to the standard PCA.

$$k(x,y) = x^T y + c$$

Linear kernel was selected for this project because it takes less training time and is more efficient than other kernels when the dataset contains a large number of features.

Examples of other kernels used in SVM are as follows:

- ✓ Polynomial Kernel
- ✓ Gaussian Kernel
- ✓ Exponential Kernel

3.2.3.5 Ridge Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated(Hilt, Donald E.; Seegrist, Donald W., 1977). .It has been used in many fields including econometrics, chemistry, and engineering(Gruber, Marvin ,1998). .Also known as Tikhonov regularization, named for Andrey Tikhonov, it is a method of regularization of ill-posed problems. It is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters(Kennedy, Peter ,2003). .In general, the method provides improved efficiency in parameter estimation problems in exchange for a tolerable amount of bias (see bias–variance tradeoff) (Gruber ,1998). .

The theory was first introduced by Hoerl and Kennard in 1970 in their *Technometrics* papers "RIDGE regressions: biased estimation of nonorthogonal problems" and "RIDGE regressions:

applications in nonorthogonal problems”. This was the result of ten years of research into the field of ridge analysis (Hoerl.; Kennard, Robert W. ,1970).

Ridge regression was developed as a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear (highly correlated) independent variables—by creating a ridge regression estimator (RR). This provides a more precise ridge parameters estimate, as its variance and mean square estimator are often smaller than the least square estimators previously derived (Hoerl.; Kennard, Robert W, 1970).

3.2.3.6 Decision Trees

A decision tree is a flowchart that starts with one main idea and then branches out based on the consequences of your decisions. It’s called a “decision tree” because the model typically looks like a tree with branches (Chang and Pavlidis,1977).

These trees are used for decision tree analysis, which involves visually outlining the potential outcomes, costs, and consequences of a complex decision (Janickow,1998). You can use a decision tree to calculate the expected value of each outcome based on the decisions and consequences that led to it. Then, by comparing the outcomes to one another, you can quickly assess the best course of action (Quinlan,1987). You can also use a decision tree to solve problems, manage costs, and reveal opportunities.

A decision tree includes the following symbols:

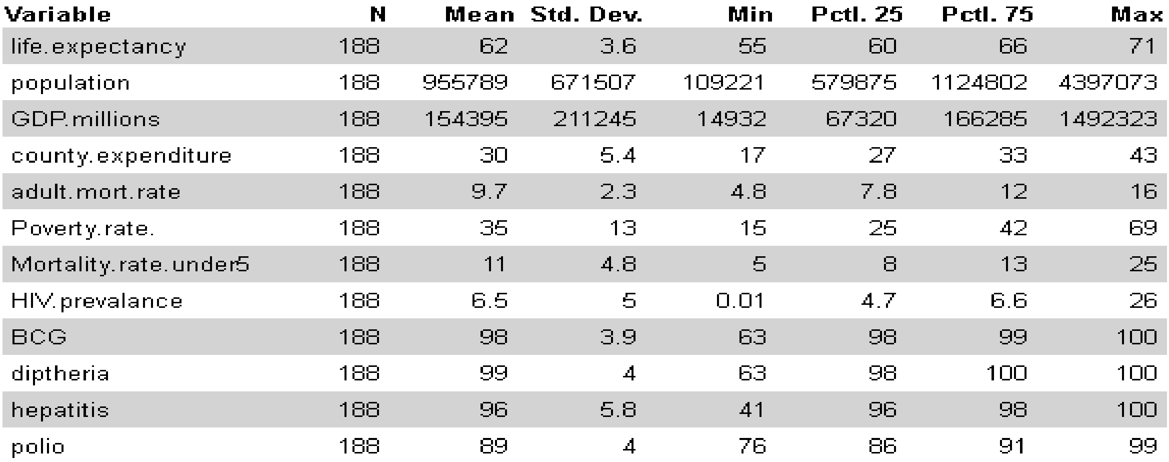
- Alternative branches: Alternative branches are two lines that branch out from one decision on your decision tree. These branches show two outcomes or decisions that stem from the initial decision on your tree (Quinlan,1986).
- Decision nodes: Decision nodes are squares and represent a decision being made on your tree. Every decision tree starts with a decision node.
- Chance nodes: Chance nodes are circles that show multiple possible outcomes.
- End nodes: End nodes are triangles that show a final outcome (Chang and Pavlidis,1977).

A decision tree analysis combines these symbols with notes explaining your decisions and outcomes, and any relevant values to explain your profits or losses. You can manually draw your decision tree or use a flowchart tool to map out your tree digitally.

Chapter 4: DATA ANALYSIS AND RESULTS.

4.1 introduction

This section presents Data analysis and results on the study of life expectancy, and factors affecting life expectancy. Research questions will be answered in-depth by the data analysis methods employed to the data.



Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
life.expectancy	188	62	3.6	55	60	66	71
population	188	955789	671507	109221	579875	1124802	4397073
GDP.millions	188	154395	211245	14932	67320	166285	1492323
county.expenditure	188	30	5.4	17	27	33	43
adult.mort.rate	188	9.7	2.3	4.8	7.8	12	16
Poverty.rate.	188	35	13	15	25	42	69
Mortality.rate.under5	188	11	4.8	5	8	13	25
HIV.prevalance	188	6.5	5	0.01	4.7	6.6	26
BCG	188	98	3.9	63	98	99	100
diptheria	188	99	4	63	98	100	100
hepatitis	188	96	5.8	41	96	98	100
polio	188	89	4	76	86	91	99

Table 4.1;Summary statistics for the data.

According to table 4.1 inn a sample size of 188 the mean life expectancy was reported to be 62 years while the minimum and maximum number of years a person is expected to live was estimated to be 55 and 71 years respectively. The minimum county expenditure on healthcare was 17 percent and maximum 43 percent.

CORRELATION PLOT.

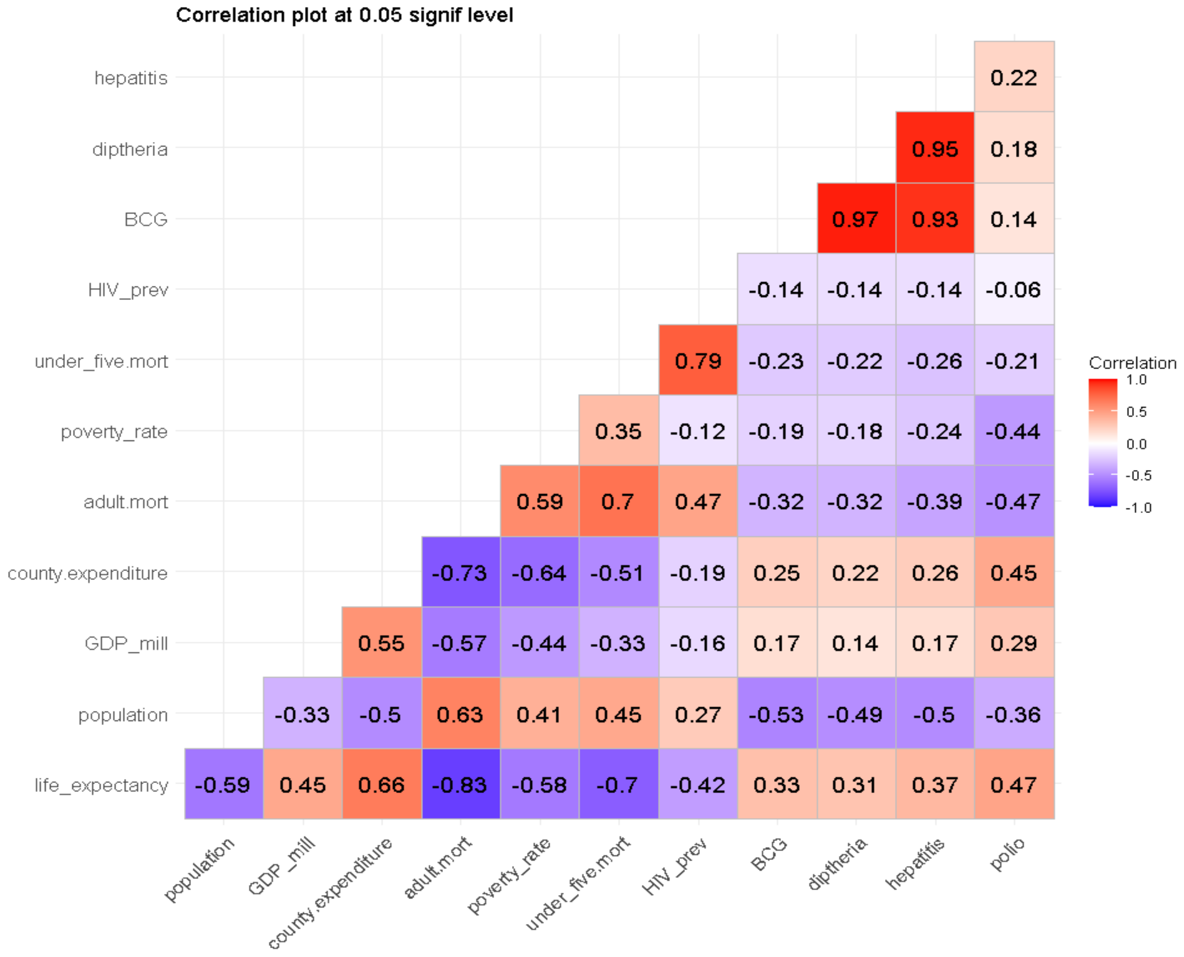


Figure 4.1; Correlation plot.

From the figure 4.1, correlation plot we can obtain correlations between the variables in the plot. We will only consider the correlations between our variable of interest (Life Expectancy) and other variables. Life Expectancy has a positive correlation with; GDP in millions, County Expenditure on Healthcare, BCG immunization coverage, Diphtheria immunization coverage, Hepatitis immunization coverage, and polio immunization coverage, and is negatively correlated with population, adult mortality rate, poverty rate, under five mortality rate and HIV prevalence.

Life expectancy has a strong correlation with county expenditure on healthcare, adult mortality rate and under five mortality rates.

4.2 Visualizations.

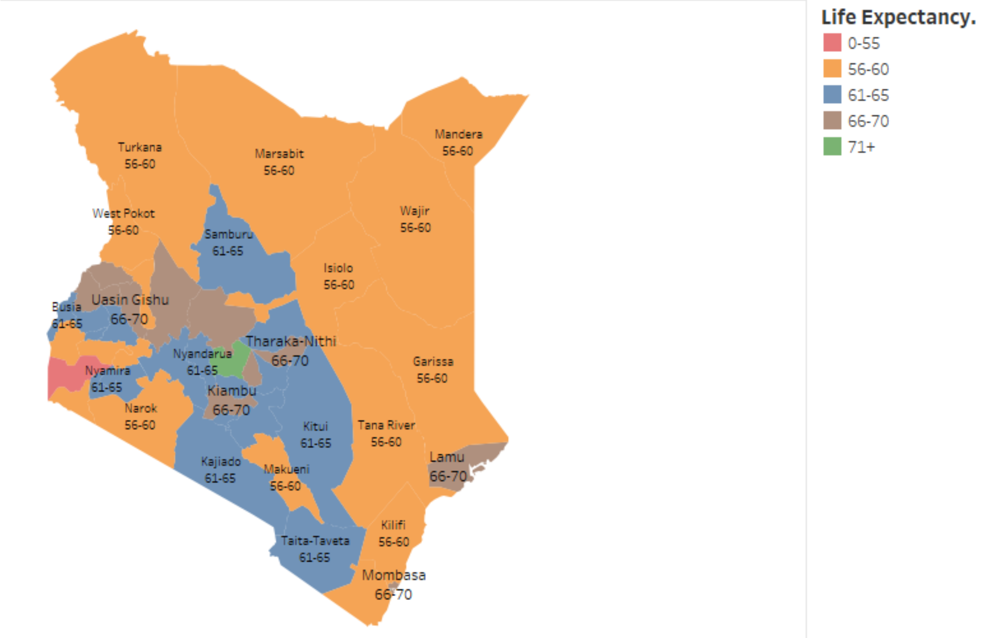


Figure 4.2; A map of Kenya showing distribution of life expectancy by counties.

As per map 4.2 above, majority of the regions had a life expectancy of between 56 and 60 years. On close observation, these regions were mostly arid and semiarid areas which make about 80 percent of Kenya’s land surface. These regions tend to experience severe drought, food insecurity, poverty and generally lack access to proper healthcare which are the major contributing factors to its lower life expectancy.

Nyanza region, specifically Homabay county had the lowest life expectancy of between 0 and 55 years. Despite having better access to health care services and availability of food and education the region has the highest prevalence of HIV/AIDS. The national AIDS Control Council report of 2022 showed Homabay county with a leading HIV/AIDS prevalence of 19.6 percent followed by Kisumu17.55percent, Siaya 15.3 percent and Migori 13.3 percent. This has had a huge impact on the region hence an all-time low life expectancy.

Central region had the highest life expectancy with the highest life expectancy of more than 71 years recorded in Nyeri county. The county has been identified as one with an exemplary performance in health service delivery as they have the lowest maternal mortality rate and infant mortality rate nationally as per the national census of 2019. The region is located in the Kenyan highlands hence agriculture is the main economic activity thus has access to food.

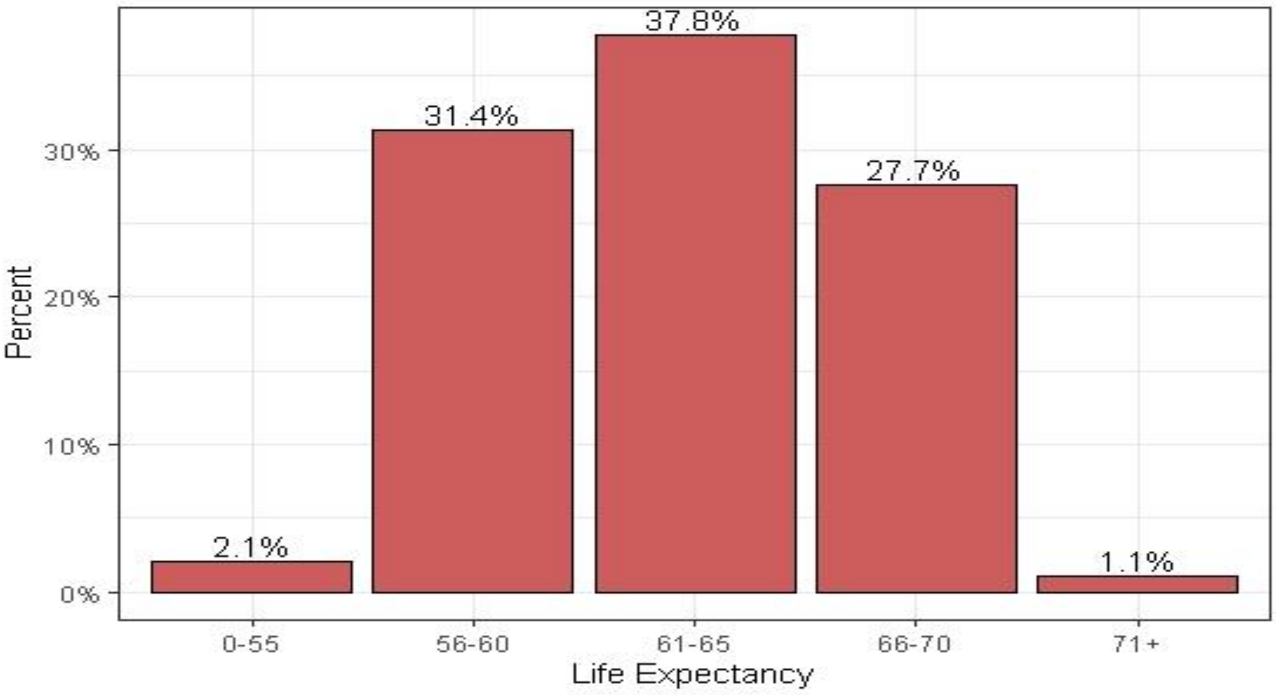


Figure 4.3; A barchart of distribution of life expectancies in percent per counties

From figure 4.3 the life expectancy bar graph follows a normal distribution with majority of the counties having a life expectancy between 61 and 65 years recording the highest percentage at 37.8 percent. Life expectancies between 56-60 years and 66-70 years recorded fairly low life expectancies at 31.4 percent and 27.7 percent respectively. The extreme ends recorded a minimum percentage with 2.1 percent of the counties having the lowest life expectancy of between 0 to 55 years and 1.1 percent of the counties having a life expectancy of more than 71 years.

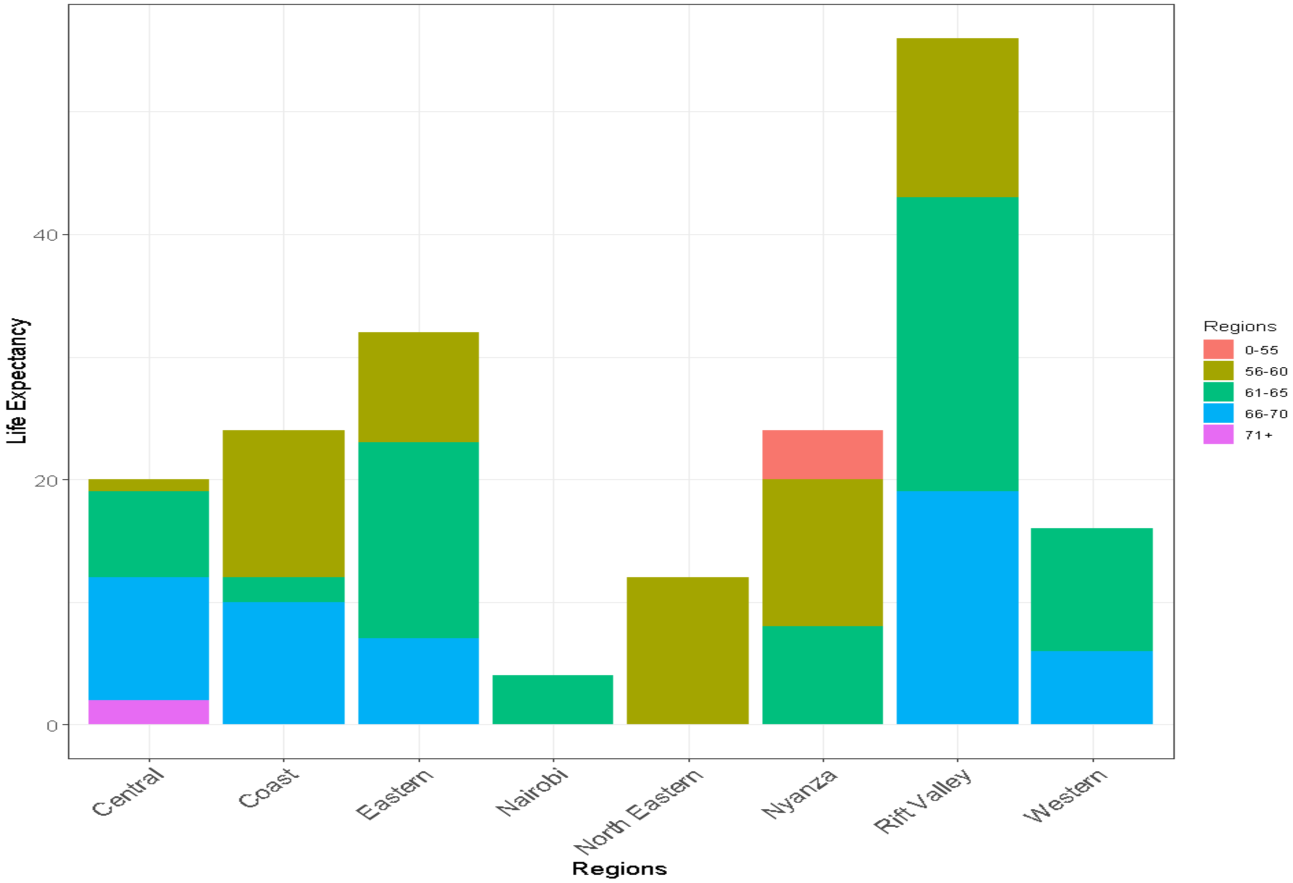


Figure 4.4; A graph of Life Expectancy distribution per regions.

From figure 4.4 it is revealed that only Central region had a life expectancy of more than 71 years which happens to be the highest life expectancy reported. The region had a fairly high life expectancy as none of the counties reported the lowest life expectancy that is between 0 and 55 years. The region is agriculturally rich hence has constant access to food and has invested more in its healthcare services compared to other counties, hence a record high life expectancy of its population.

Majority of Nairobi's population reported a life expectancy of between 61 and 65 years.

The population in North Eastern region reported a fairly low life expectancy of between 56 and 60 years. Despite having the lowest HIV/AIDS prevalence in Kenya, the region is majorly arid and semiarid hence constantly hit by climatic shocks like floods and droughts. They also lack access to proper health care and proper education hence lowering the life expectancy of its population.

Out of all regions Nyanza region is the only region that recorded the lowest life expectancy of between 0 and 55 years. The region has fewer people living below the poverty line compared with the national average, more people with access to improved health care services but its HIV/aids prevalence remains at an all-time high nationally hence largely affecting the life expectancy of its population.

Rift valley region is the largest consisting of 14 counties hence with the largest population. The region consists of most regions with fairly high life expectancy of between 61 years and 70 years while the remaining have a fairly low life expectancy of between 56 and 60 years. This may be attributed by counties like Uasin-Gishu being the country's bread basket because of its large-scale maize and wheat production. Majority of the counties are generally agriculturally rich hence have access to adequate food and have access to health care services and education. However, for the counties with pastoralist communities like Laikipia, Samburu and Narok the life expectancy tends to be lower due to inaccessibility of education and health services. Other counties like Baringo and Turkana are constantly hit by extreme weather conditions.

Western region reported life expectancy of between 61 years and 70 years. The region consists of Kakamega, Vihiga, Bungoma and Busia counties which all share farming as the main economic activities. Despite Kakamega and Bungoma being ranked top five contributors to national poverty the region still recorded a fairly high life expectancy. This may be contributed to the fact that as much as most of the population is poor, they still have access to adequate food and can access some form of healthcare service and education.

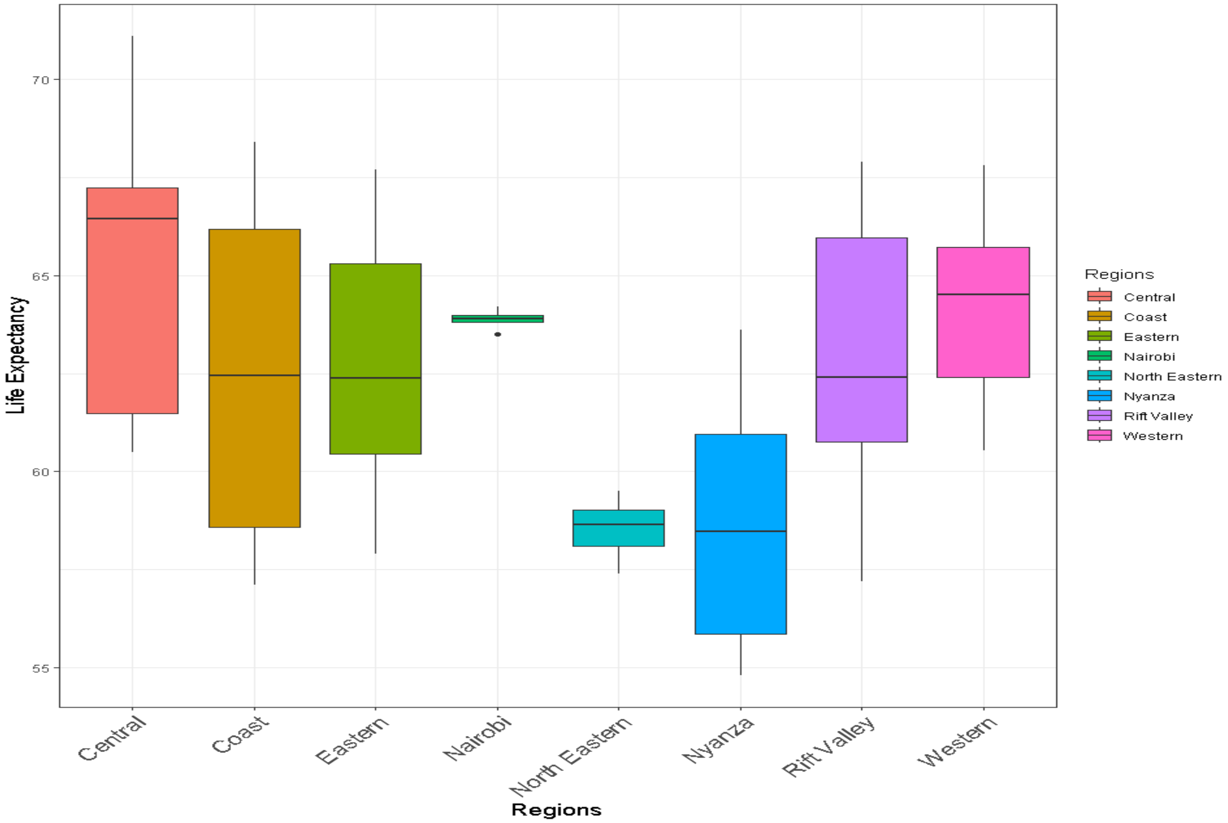
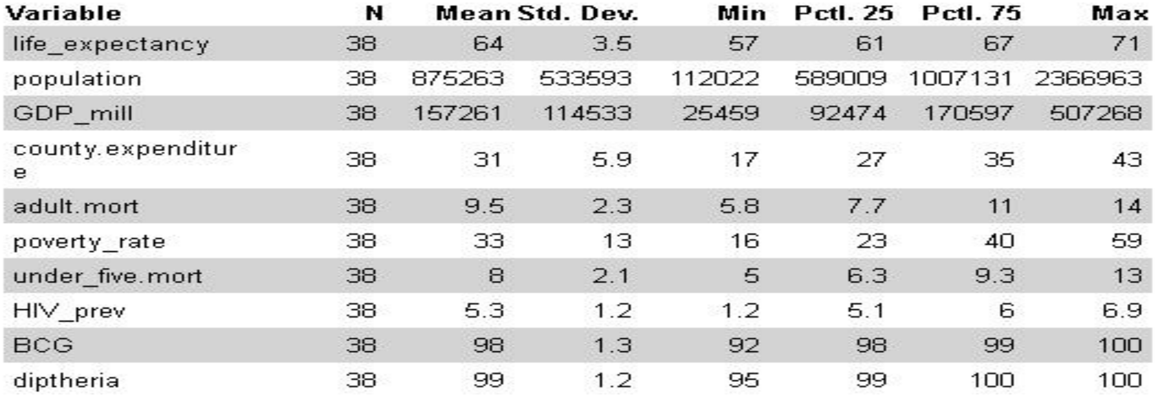


Figure 4.5; A boxplot of Life expectancy distribution per regions.

From box plot 4.5 shown are the lower quartiles medians and the upper quartiles for the life expectancies for various regions. The regions that posted the highest for the above metrics being Central and Rift Valley. Regions that posted the lowest were Nyanza and North Eastern.

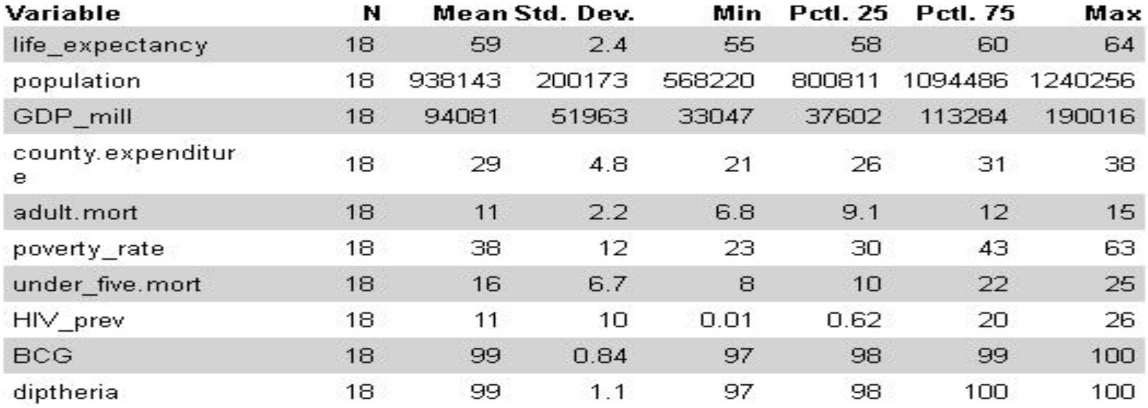
CENTRAL AND RIFT VALLEY(CLASS 1)



Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
life_expectancy	38	64	3.5	57	61	67	71
population	38	875263	533593	112022	589009	1007131	2366963
GDP_mill	38	157261	114533	25459	92474	170597	507268
county.expenditure	38	31	5.9	17	27	35	43
adult.mort	38	9.5	2.3	5.8	7.7	11	14
poverty_rate	38	33	13	16	23	40	59
under_five.mort	38	8	2.1	5	6.3	9.3	13
HIV_prev	38	5.3	1.2	1.2	5.1	6	6.9
BCG	38	98	1.3	92	98	99	100
diptheria	38	99	1.2	95	99	100	100

Table 4.2; Summary statistics of Central and Rift Valley

NYANZA AND NORTH EASTERN (CLASS 2).



Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
life_expectancy	18	59	2.4	55	58	60	64
population	18	938143	200173	568220	800811	1094486	1240256
GDP_mill	18	94081	51963	33047	37602	113284	190016
county.expenditure	18	29	4.8	21	26	31	38
adult.mort	18	11	2.2	6.8	9.1	12	15
poverty_rate	18	38	12	23	30	43	63
under_five.mort	18	16	6.7	8	10	22	25
HIV_prev	18	11	10	0.01	0.62	20	26
BCG	18	99	0.84	97	98	99	100
diptheria	18	99	1.1	97	98	100	100

Table 4.3; Summary statistics of Nyanza and North Eastern

Figure 4.2 and figure 4.3 show the summary statistics for the regions that posted the highest and lowest of life expectancies. Getting into the details of the supposed factors affecting life expectancy. For Central and Rift Valley which had the highest on average of life expectancies (64) and for Nyanza and North Easter n which showed the minimum of life expectancies (59) on average, there are quite a number of differences that separate the two classes. They included:

- **Mortalities**: Checking at their mortalities (adult, under five), class 1 posted a mean of (9.5,5.3) respectively while class2 posted on average (11,16) which was higher than class 1.
- **County expenditure on healthcare**. The county expenditures on healthcare for the two regions were as follows class 1(31) and class 2(29). Class 1 had the highest budget as compared to class 2.
- **HIV prevalence**: Taking a closer look at the rates of HIV prevalence for class 1 they reported a mean of 5.3 while class 2 reported a mean of 11. Class 2 posted a higher rate of HIV prevalence as compared to class 2.

<u>ANSWERS TO THE RESEARCH QUESTIONS</u>.

1. To find out if increasing healthcare expenditure will improve life expectancy.

<u>A GRAPH OF LIFE EXPECTANCY AGAINST COUNTY EXPENDITURE ON HEALTHCARE</u>

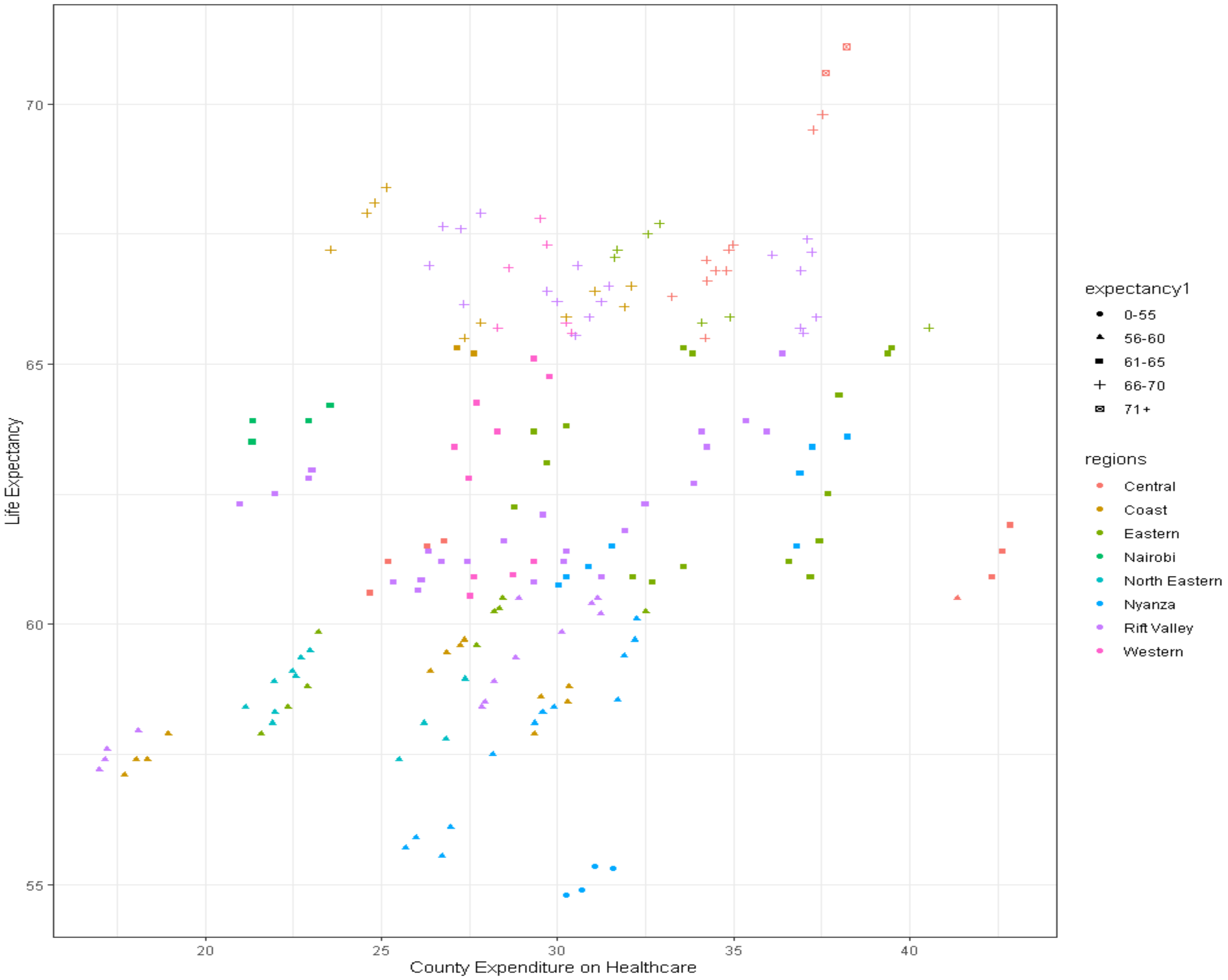


Figure 4.6; A scatter plot of life expectancy against County Expenditure on Health care.

Interpretation: counties that spend less on healthcare experience the lowest rate of life expectancy. Some counties in rift valley spend the least in healthcare, majority of the regions spending between 24-35 percent and central and coast being the highest in expenses in healthcare.

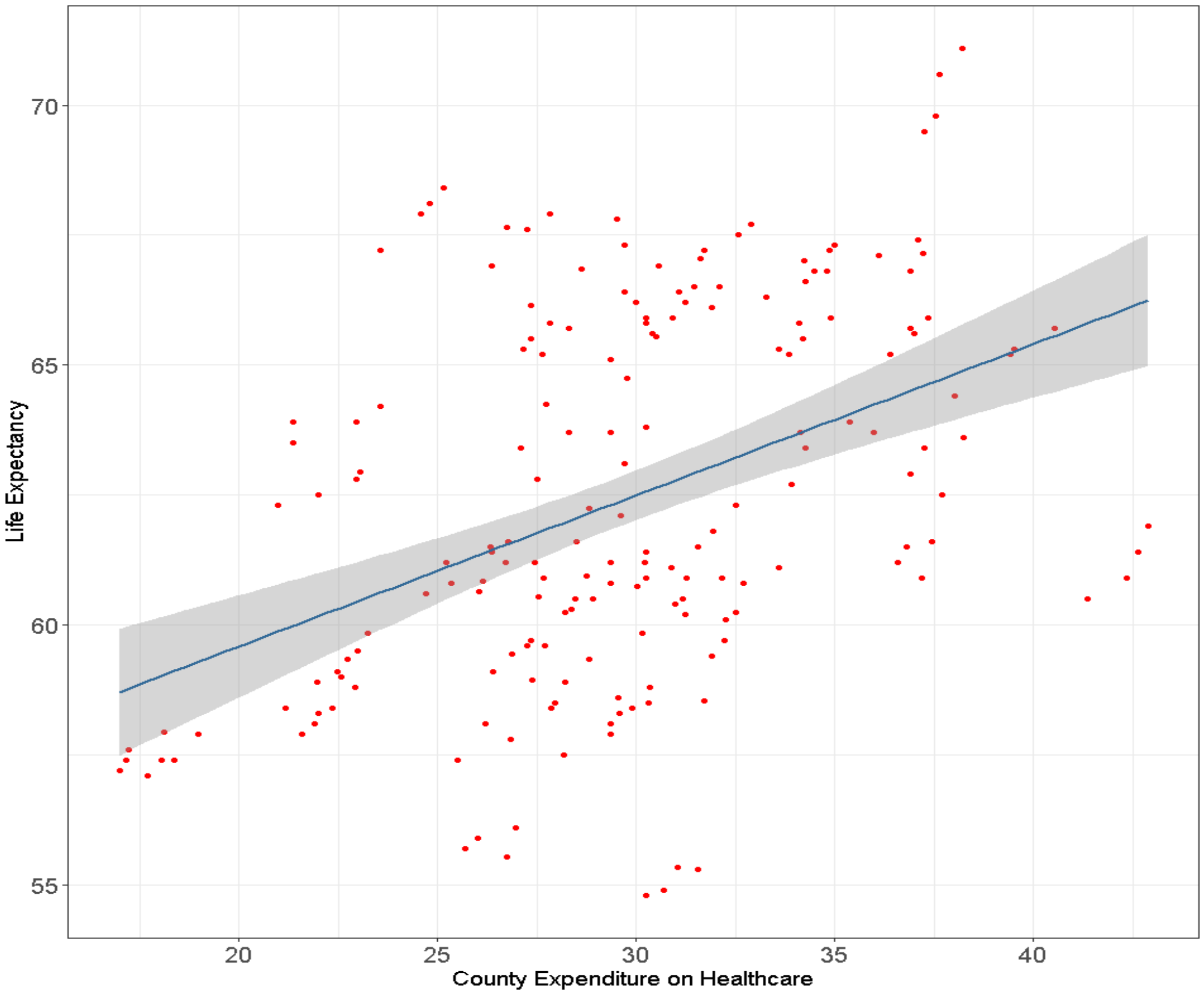


Figure 4.7; A scatter plot of life expectancy against County Expenditure on Health care

Figure 4.7 above shows a linear relationship between life expectancy and County Expenditure on healthcare. This is due to a positive slope depicted by the linear relationship. This implies that increasing expenditure improves life expectancy.

2. Determining how infant and adult mortality rates affect life expectancy.

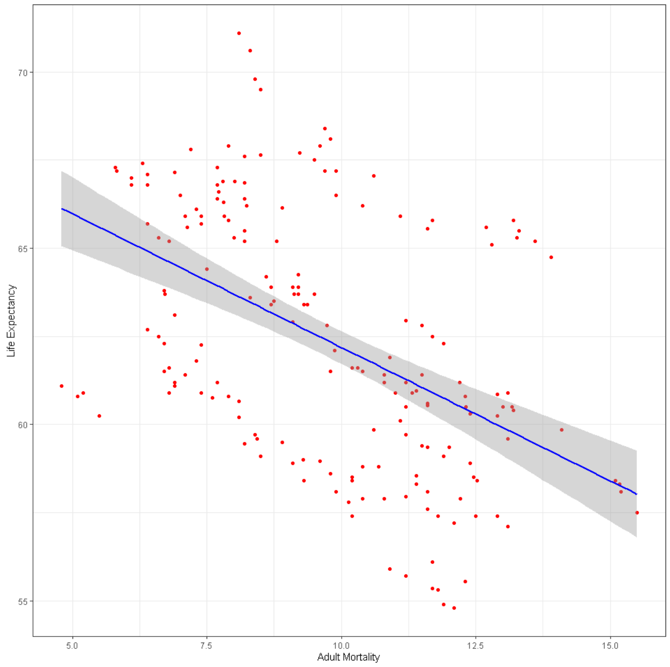


Figure 4.8; a graph of life expectancy against adult mortality

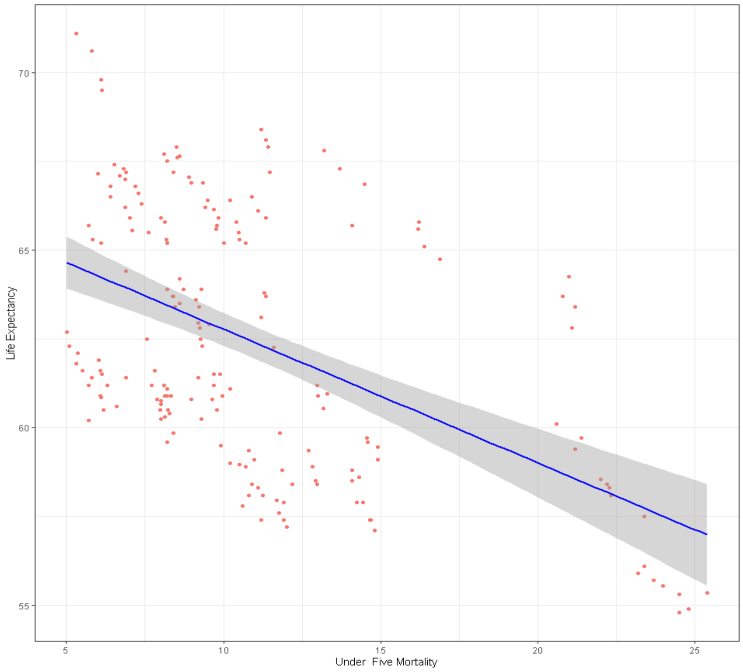


Figure 4.9; a graph of life expectancy against under-five mortality

Figure 4.8 and figure 4.9 graphs explain the relationship between life expectancy, adult mortality and under five mortalities. It is quite clear from the graphs that mortality for the adults and under- five negatively affects life expectancy because of the negative slope.

3. Does HIV prevalence affect life expectancy?

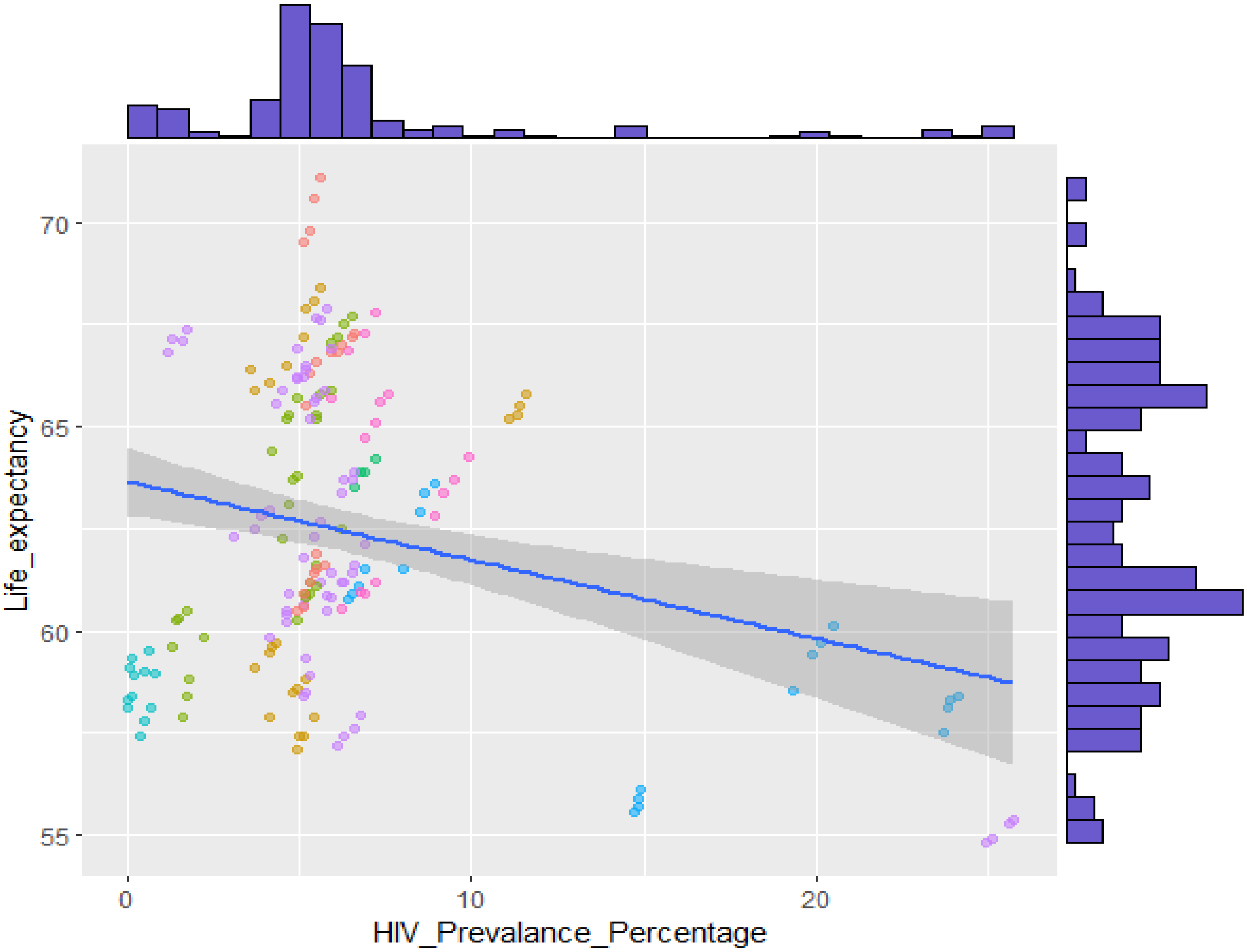


Figure 4.10; a graph of life expectancy against HIV prevalence.

Figure 4.10 above shows a strong negative relationship between life expectancy and HIV Prevalence. This is as a result of a negative slope between life expectancy and HIV prevalence.

4. Does immunization coverage affect life expectancy?

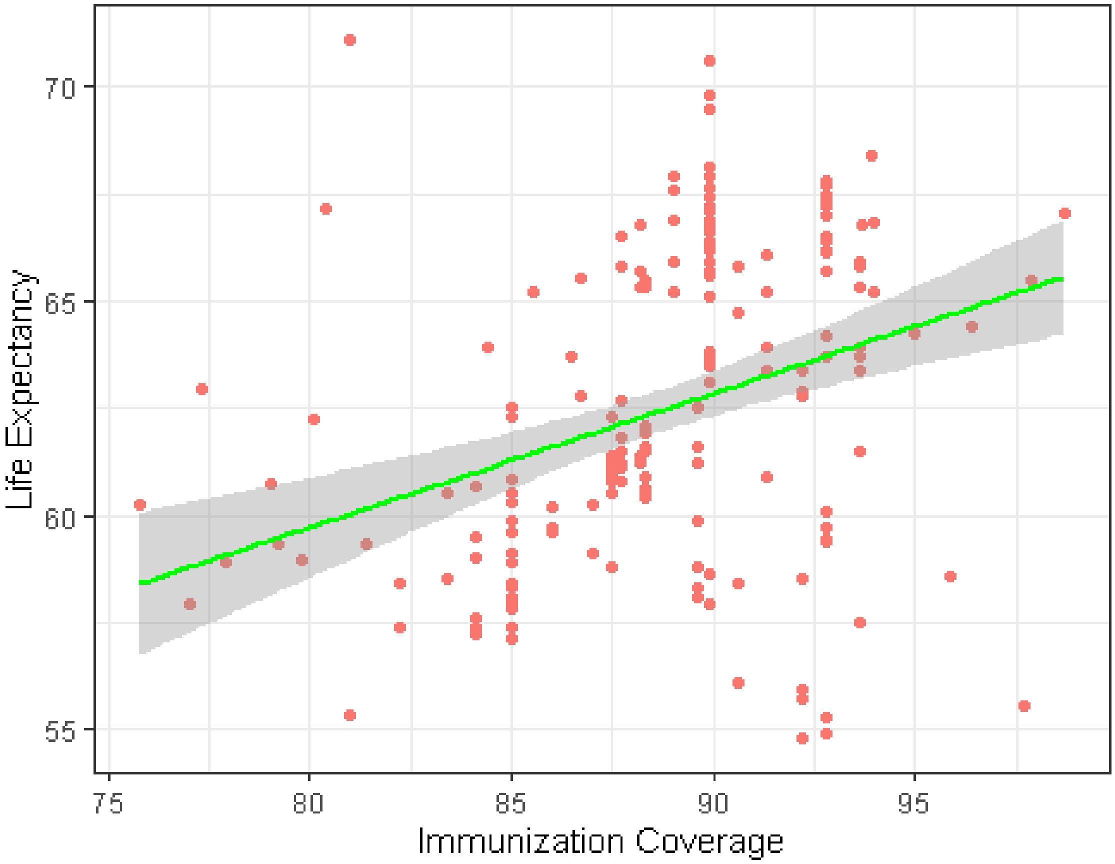


Figure 4.11; A graph showing linear relationship between life expectancy and immunization coverage.

Figure 4.11 above shows that there exists a positive linear relationship between life expectancy and immunization coverage. This is because of the positive slope that is depicted by the data. The relationship is not perfect as the regression line does not pass through the origin.

5. Does poverty rate affect life expectancy?

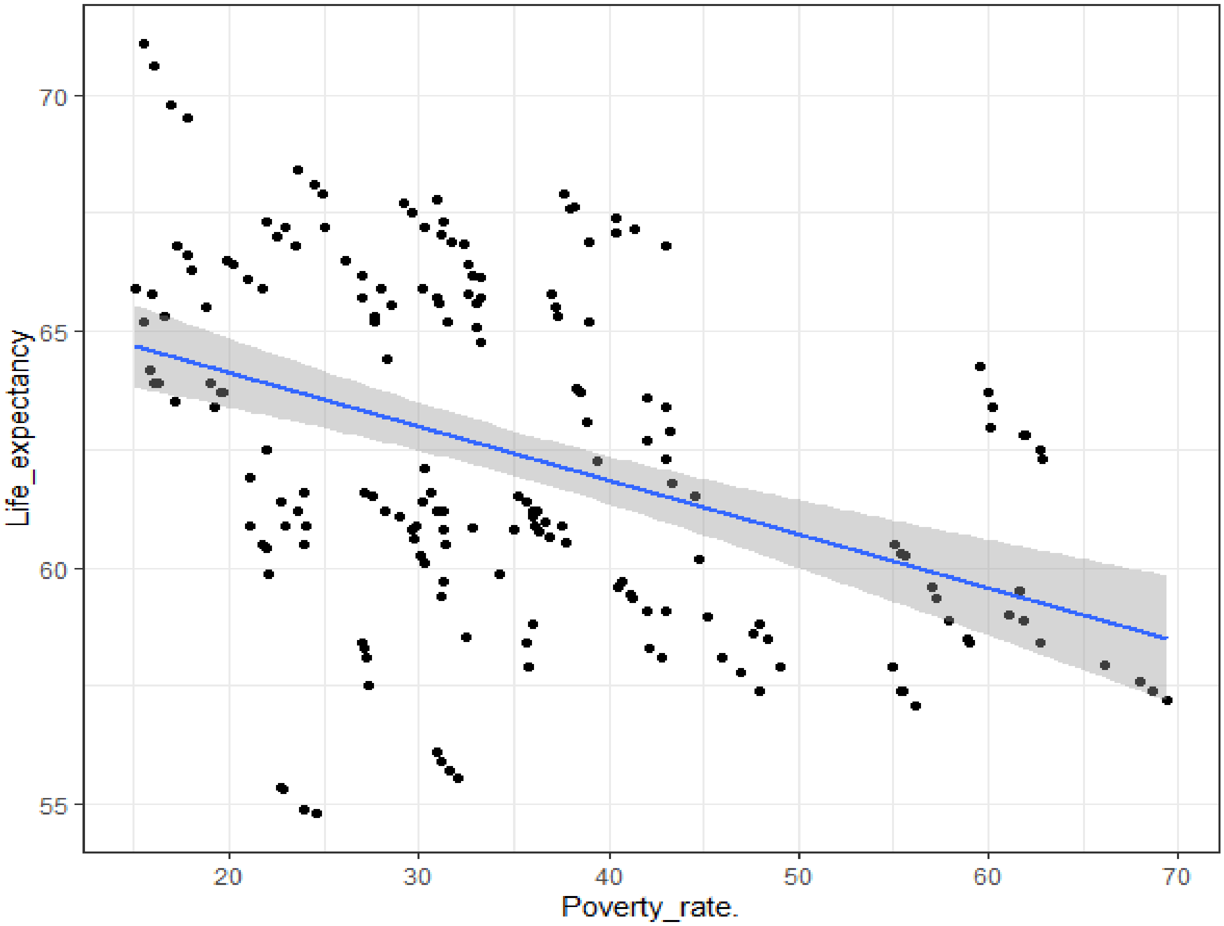


Figure 4.12; A graph showing linear relationship between life expectancy and poverty rate.

Figure 4.12 above shows a negative relationship between life expectancy and poverty rate. This is because of the negative slope exhibited by the data.

4.3 modelling.

4.3.1 Linear regression model diagnostics.

LINEARITY

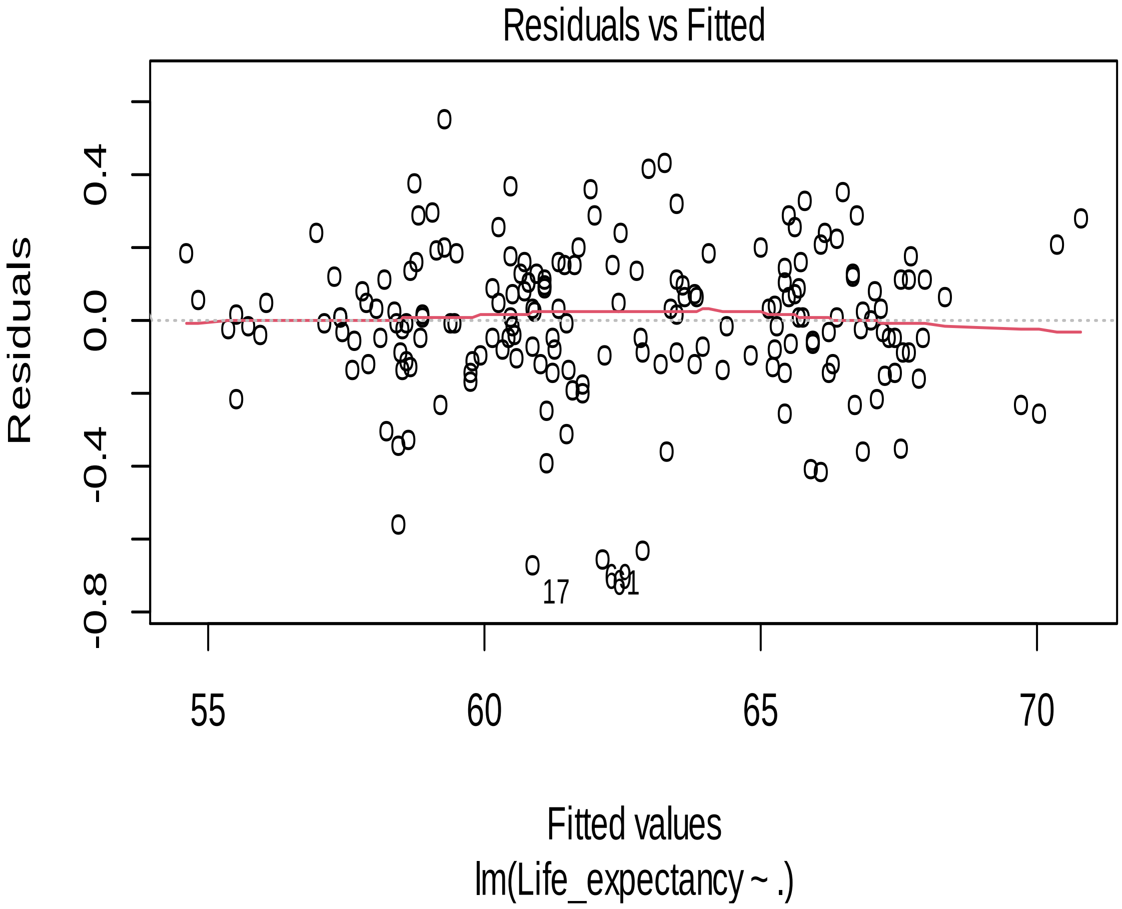


Figure 4.13; Residuals vs fitted plot

From figure 4.13 since the residuals are equally spread around the horizontal line and without any distinct pattern then the linearity assumption has been met.

NORMALITY.

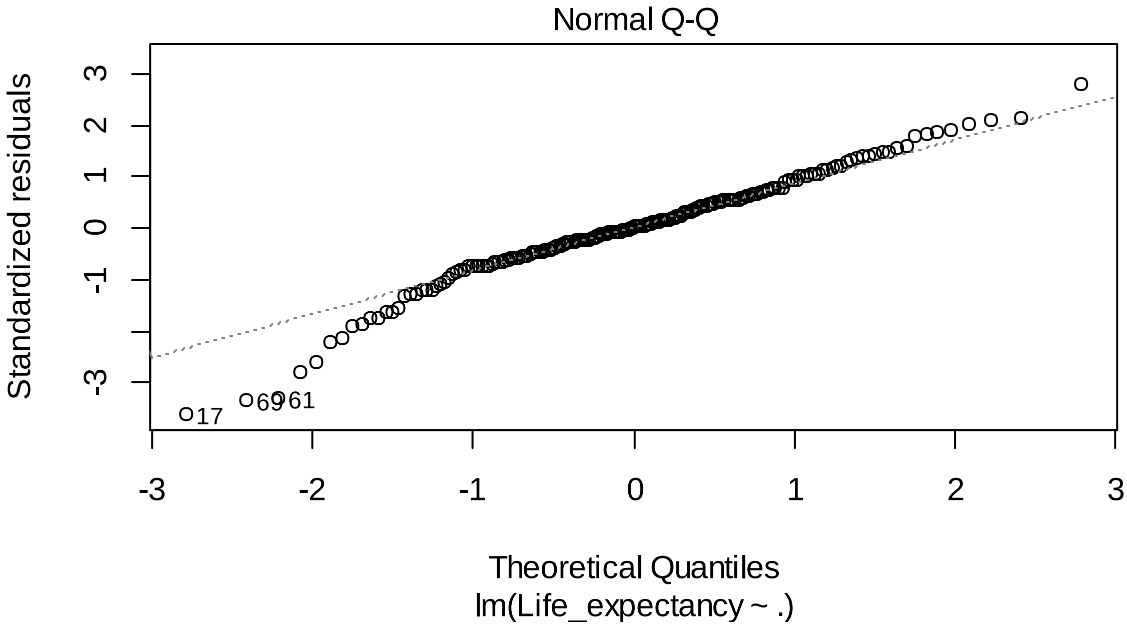


Figure 4.14; Normal Q-Q plot.

From figure 4.14 the quantile-quantile plot for the residuals shows a straight line connecting the first and third quartiles thus confirming that the normality assumption has been met and that the residuals come from a normal population. If data come from a normal distribution, then the quantiles of their standardized values should be approximately equivalent to the known quantiles of the standard normal distribution.

HOMOSCEDASTICITY.

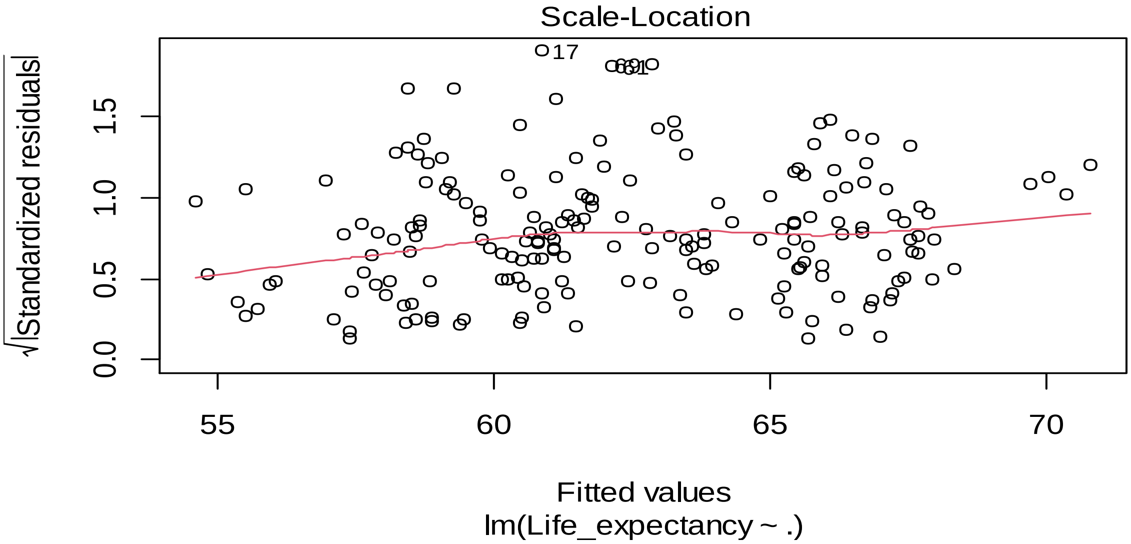


Figure 4.15; Scale location plot.

From figure 4.15 Since there is no pattern in the above scatterplot, then we say that the homoscedasticity (constant variance) assumption has been met. The residuals are symmetric about zero. If the constant variance assumption is violated, a transformation on the response that stabilizes the variance on the response is recommended. This transformation may be a log transformation of the response variable so that you model the logarithm of the response as a function of the predictors. Alternatively, a square root transformation of the response variable could be used so that you model the square root of the response as a function of the predictors.

INDEPENDENCE OF OBSERVATIONS.

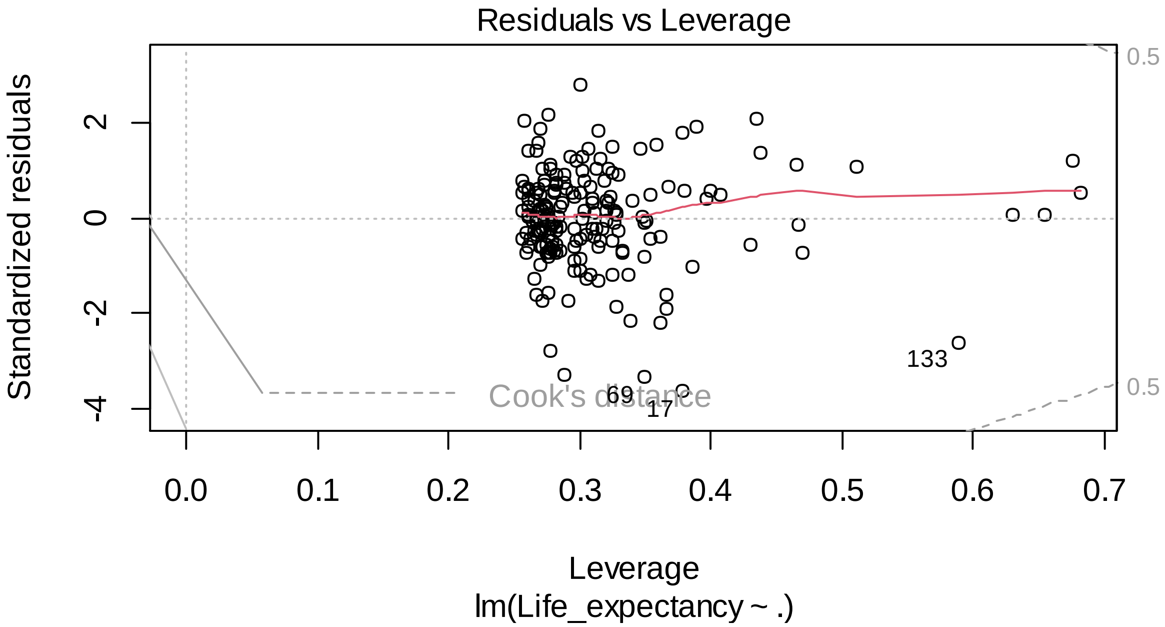


Figure 4.16; Residuals vs Leverage plot.

Cook’s distance -this quantity measures how much the entire regression function changes when the i-th variable is deleted. From figure 4.16 since there is no influential case beyond the cook’s distance dotted line then there will be no significant change in the r squared value if any case is removed.

4.3.2 model performance summaries.

		Initial Model		After Prediction	
	Algorithm	**R squared**	**MSE**	**R squared**	**MSE**
1	Decision Tree	0.8492	1.0232	0.8591	0.9764
2	Support Vector Machine	0.8828	0.9721	0.8743	0.9728
3	Ridge Regression	0.9310	0.8628	0.9234	0.8317
4	Multiple Linear Regression	0.7544	1.8942	0.7485	1.7485
5	Principal Component Regression	0.7514	1.7341	0.7321	1.8762
6	Random Forest	0.9233	1.0316	0.9801	0.7391

Table 4.4; Model performance summaries.

Table 4.4 above shows summary of models fitted, Random Forest was the best performing model as it had the minimum mean squared error and it explained majority of the variation in the data at 98.01%.

For our random forest model, the following plot can explain to us the variables that were of much importance while predicting life expectancy.

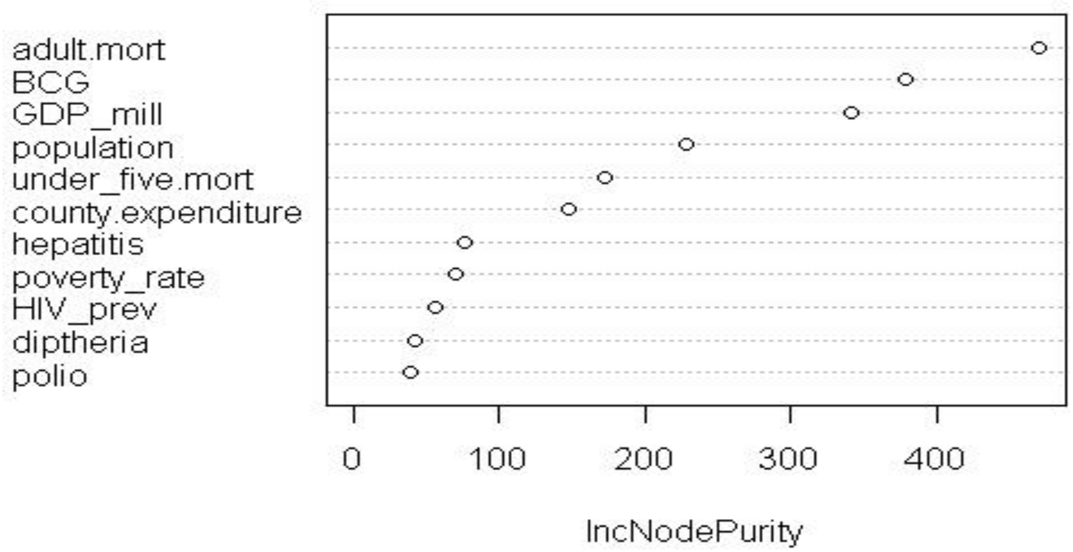


Figure 4.17; A graph showing importance of variables in Random Forest Model

From figure 4.17, adult mortality rate was more important in predicting life expectancy followed by BCG and the least important was Polio.

Chapter 5: CONCLUSIONS AND RECOMMENDATIONS.

Introduction

This chapter includes the conclusions and recommendations for the study outcomes.

5.1 Summary

We found out that life expectancy had a positive relationship with GPD, county expenditure on healthcare, BCG, Diphtheria, Hepatitis. Life expectancy had a positive relationship with adult mortality, under five mortality, HIV rate prevalence and population. We fitted the listed models and the best performing model was Random Forest model.

5.2 Conclusions.

1. Increasing health care expenditure leads to an increase in life expectancy. The more people in various regions spend on healthcare, the more they are likely to live longer. Factors like employment, balanced diet food which are healthy habits lead to long life.0-55 spend average on healthcare hence average life expectancy.56-60 spend relatively high on healthcare hence relatively high rate of expectancy.
2. A higher adult and under five mortalities negatively affect the life expectancy of people living in Kenyan counties. An increase in either of these mortality rates leads to a decline in life expectancy.
3. Life Expectancy and HIV Prevalence have a negative relationship. An increase in HIV prevalence leads to a decline in the expected number of years a person will live.
4. Life expectancy and immunization coverage has a positive relationship. Increasing the immunization coverage leads to an increase in life expectancy.
5. Poverty rate and life expectancy has a negative relationship. When poverty rate increases the life expectancy of people tends to decline.

5.3 Recommendations for policy

This study reveals to policy makes the importance of immunization. Immunization saves lives. Immunization helps protect future generations by eradicating diseases thus making it essential towards improving health status. With improved health status the country's life expectancy is projected to improve in future. Therefore, the county governments are required to:

- Allocate more resources in order to ensure a hundred percent immunization coverage.
- Improve the GDP per capita growth in their counties. The government has placed emphasis on job creation by issuing low interest loan Hustler Fund and Women Enterprise Fund to youths and women. However, these initiatives have not had a full impact on employment creation due to challenges such as poor book keeping, financial over dependence and disintegration of investment groups that has led to the collapse of small scale and medium enterprises. To counter these, further capacity building of entrepreneurial skills should be done to the funds beneficiaries as well as monitoring of their business and appropriate business support should be availed to them.
- The study shows a negative linear relationship between life expectancy and HIV prevalence thus making it one of the important factors of consideration. The county governments especially in the Nyanza region have to allocate more resources in curbing the effect of HIV. There has to be an increase in the number of public awareness campaigns on the causes, effects of HIV and its impact on life expectancy.
- As per the graphical evidence Central region has averagely the most allocation to the health sector and thus has the highest life expectancy while North Eastern allocated a small percentage of its expenditure to the health sector hence the low life expectancy. The county governments should therefore increase the budget allocation to the health sector in order to improve the life expectancy of its citizens.
- In view of the effect of infant and adult mortality rates on life expectancy, the county governments should allocate more resources to health programs that concentrate on infectious disease like HIV/AIDS, Malaria and tuberculosis. Non-communicable disease like diabetes, cancer, heart and lung disease should also be a priority in Kenyan counties. Mostly, these morbidities are linked to lifestyle choices, but also can worsen due to poor diet, smoking, alcohol use, or degraded environment. With the upcoming middle class in

Kenya, lifestyle is also changing: increasing abuse of alcohol and drugs etc. Obesity, a major risk factor for diabetes, heart disease and some types of cancers are also an important health issue. The general public should be made aware of these through public education. The county governments should also increase resources to be used in curbing the above stated diseases.

- The national government should also introduce Universal Health Coverage to enable every citizen get access to healthcare without any hindrances.
- The government should increase the minimum wage given to employees and also create more employment opportunities in order to reduce the poverty rate in the country. This is in order for the citizens to be able to afford basic needs such as healthcare in order to improve the life expectancy.

5.4 Recommendation for further research.

There exists a gap in the field of study relating to the study of factors affecting life expectancy, further studies can look into factors such as schooling, crime rate state of development of each county, main occupations per counties, unemployment, literacy, foreign exchange rate etc. These variables were not used in the study due to data constraints. Our study recommends that future studies to include the above factors when investigating the factors affecting life expectancy in Kenyan counties.

REFERENCES

Abhinaya.V[1], Dharani.B.C[2], Vandana.A[3], Dr.Velvadivu.P[4], Dr.Sathya.C.[5]. Statistical analysis on factors influencing life expectancy. International Research Journal of Engineering and Technology (IRJET)

Alboukadel Kassambara(2018). Machine Learning Essentials.

Beck, James Vere; Arnold, Kenneth J. (1977). Parameter Estimation in Engineering and Science.

James Beck. p. 287. ISBN 978-0-471-06118-2.

C.Z. Janickow, "Fuzzy Decision Trees: Issues and Methods", *IEEE Trans. Systems Man and Cybernetics B: Cybernetics*, vol. 28, no. 1, pp. 1-14, 1998.

Calot G, Sardon J-P. Methodology for the calculation of Eurostat's demographic indicators. Detailed report by the European Demographic Observatory

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

Gruber, Marvin (1998). *Improving Efficiency by Shrinkage: The James--Stein and Ridge Regression Estimators* . CRC Press. p. 2. ISBN 978-0-8247-0156-7.

Gruber, Marvin (1998). Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators. Boca Raton: CRC Press. pp. 7–15. ISBN 0-8247-0156-9.

Hilt, Donald E.; Seegrist, Donald W. (1977). Ridge, a computer program for calculating ridge regression estimates. doi:10.5962/bhl.title.68934.

Hoerl, Arthur E.; Kennard, Robert W. (1970). "Ridge Regression: Applications to Nonorthogonal Problems". Technometrics. **12** (1): 69–82. doi:10.2307/1267352. JSTOR 1267352.

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression 2nd edn John Wiley & Sons. *Inc.: New York, NY, USA*, 160–164.

J.R. Quinlan, "Decision Trees as Probabilistic Classifiers", *Proc. Fourth Int'l Workshop Machine Learning*, pp. 31-37, 1987.

J.R. Quinlan, "Induction of Decision Trees", *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

Jolliffe, I. T. (2006). Principal Component Analysis. Springer Science & Business Media. p. 178. ISBN 978-0-387-22440-4.

Kennedy, Peter (2003). A Guide to Econometrics (Fifth ed.). Cambridge: The MIT Press. pp. 205–
206. ISBN 0-262-61183-X.

Khomsan, A., Dharmawan, A. H., Sukandar, D., & Syarief, H. (2015). *Indikator Kemiskinan dan Misklasifikasi Orang Miskin*. Yayasan Pustaka Obor Indonesia.

Leung, Kenneth Towards Data Science, 2022, April,6: Web*https://towardsdatascience.com/principal-component-regression-clearly-explained-and-implemented-608471530a2f*

Makowski Dominique, S Ben-Shachar Martan, Patil Indrajeet, Ludecke Daniel(2020). Methods and Algorithms for Correlation Analysis in R. *Journal of Open-Source Software,5(51),2306,2020.*

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

Morduch, J. (1994). Poverty and vulnerability. *The American Economic Review*, *84*(2), 221-225.

R.L.P. Chang and T. Pavlidis, "Fuzzy Decision Tree Algorithms", *IEEE Trans. Systems Man and Cybernetics*, no. 1, pp. 28-35, 1977.

Rino Rappuoli1, Mariagrazia Pizza, Giuseppe Del Giudice, and Ennio De Gregorio (2014) Vaccines, new opportunities for a new society. Novartis Vaccines, 53100 Siena, Italy

Souza, César R. “Kernel Functions for Machine Learning Applications.” 17 Mar. 2010. Web<http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html>.

Suthaharan, S. (2016). Support Vector Machine. Integrated Series in Information Systems, 207–235.

Teguh Widodo1*, Indang Dewata2, Iswandi Umar3 and Aprizon Putra4(2021). Does the poverty take effect to the life expectancy?Journal of Xi'an Shiyou University, Natural Sciences Edition.

Theory Biosci. (2003) 1222:313-320 CCC, Urban & Fisher Verlag:

http://www.urbanfischer.de/journals/theorybios

Unwin Antony (2015). Graphical Data Analysis with RWeb<.*https://www.routledge.com*>

World Health Organization (2013) Global Vaccine Action Plan2011 - 2020. Available at www.who.int/immunization/global_vaccine_action_plan/GVAP_doc_2011_2020/en/index.html.Accessed June 1, 2014.

Yiu,Tony Towards Data Science,2019,June,12: Web*https://towardsdatascience.com/understanding-random-forest-58381e0602d2*

F. Esposito, D. Malerba and G. Semeraro, "A Comparative Analysis of Methods for Pruning Decision Trees", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 476-491, 1997.

Zhang, J., Zhang, J., and Lee, R. D. (2001). Mortality declines and long-run economic growth. *Journal of Public Economics*, *80* (3), 485-507.

APPENDICES.

Codes.

DATA ANALYSIS STA 450 PROJECT

GROUP 10

2023-02-04

Loading packages

```
library(dplyr)

library(ggcorrplot)

library(tidyverse)

library(dplyr)
library(ggplot2)
library(vtable)

library(scales)

library(ggcorrplot)
```

Plotting graphs

```
expec<-read.csv(file.choose())

#getting summary statistics for the data.
st(expec)
```

Subsetting the data.

```
expec1=expec %>% mutate(county_expenditure=county.expenditure,adult_mortality=adult.mort
,under5_mortality=under_five.mort)

expec1=expec1[,c(-6,-8,-10)]

expec2<-expec1[,c(3,4,5,7,8,9,10,11,12,14,15,16)]
```

Plotting a correlation plot.

```
d<-round(cor(expec2),2)
ggcorrplot(d,lab=TRUE,lab_size = 5,type="lower",pch=4,pch.col="black",legend.title = "Correl
ation",title="Correlation plot at 0.05 signif level",sig.level = 0.05,insig="blank")
```

A graph of distribution of life expectancies by counties

```
library(tidyverse)
library(dplyr)

#Creating a grouping variable for life expectancy called expectancy1.
expec=expec %>% mutate(Life_expectancy1=round(Life_expectancy,0))
expec$expectancy1<-c(ifelse(expec$Life_expectancy1<=55,"0-55",
                          ifelse(expec$Life_expectancy1>=56.00 & dat11$Life_expectancy1<=6
0.99,"56-60",
                               ifelse(expec$Life_expectancy1>=61.00 & dat11$Life_expectancy
1<=65.99,"61-65",
                                   ifelse(expec$Life_expectancy1>=66.00 & dat11$Life_expecta
ncy1<=70.99,"66-70","71+")))))



plota <- expec %>%
  count(expectancy1) %>%
  mutate(pct = n/sum(n),
      pctlabel = paste0(round(pct*100,1),"%" ))
ggplot(plota,
     aes(x=(expectancy1),
       y=pct)) +
  geom_bar(stat = "identity",
       fill = "indianred",
       color = "black") +
  geom_text(aes(label = pctlabel),
        vjust=-0.25)+
  scale_y_continuous(labels = percent ) +
  labs(x ="Life Expectancy",
     y = "Percent")+ theme_bw() + theme(axis.text.x=element_text(size=15,"bold"),axis.text.y=e
lement_text(size=12,"bold"),axis.title=element_text(size=14,"bold"))
```

Boxplot of distribution of life expectancies by regions.

```
ggplot(expec,aes(regions,life_expectancy,fill=regions))+geom_boxplot() + theme_bw() + labs(y
="Life Expectancy", x="Regions",fill="Regions") + theme(axis.text.x=element_text(size=15,ang
le=45,hjust =1,"bold"),axis.text.y=element_text(size=9,"bold"),axis.title=element_text(size=14,"
bold"))
```

Scatterplot of distribution of life expectancies by regions.

```
ggplot(expec,aes(y=life_expectancy,x=county.expenditure,col=regions,shape=expectancy1)) + g
eom_point() +labs(x="County Expenditure on Healthcare",y="Life Expectancy")+ theme_bw() +
theme(axis.text.x=element_text(size=15,"bold"),axis.text.y=element_text(size=15,"bold"),axis.tit
le=element_text(size=14,"bold"))+theme(legend.text=element_text(size=rel(1.2)))
```

A Scatterplot of life expectancy against county expenditure on healthcare.

```
ggplot(expec,aes(y=life_expectancy,x=county.expenditure,col=2)) + geom_point(col='red') + geom_smooth(method="lm")+labs(x="County Expenditure on Healthcare",y="Life Expectancy") + theme_bw() + theme(legend.position = "none")+ theme(axis.text.x=element_text(size=13,"bold"),axis.text.y=element_text(size=15,"bold"),axis.title=element_text(size=14,"bold")) + theme(axis.text.x=element_text(size=15,"bold"),axis.text.y=element_text(size=15,"bold"),axis.title=element_text(size=14,"bold"))+theme(legend.text=element_text(size=rel(1.5)))
```

A Scatterplot of life expectancy against Adult Mortality.

```
names(expec)

ggplot(expec,aes(x=adult.mort,y=life_expectancy,col=regions)) + geom_point() +  labs(x="ADULT MORTALITY", y="Life Expectancy")+ theme_bw() + theme(axis.text.x=element_text(size=15,"bold"),axis.text.y=element_text(size=15,"bold"),axis.title=element_text(size=14,"bold"))+theme(legend.text=element_text(size=rel(1.2)))
```

A Scatterplot of life expectancy against Adult Mortality.

```
ggplot(expec,aes(x=adult.mort,y=life_expectancy)) +theme_bw() + geom_point(col="red") +geom_smooth(method="lm",color="blue")+ labs(x="Adult Mortality", y="Life Expectancy") + theme(axis.text.x=element_text(size=15,"bold"),axis.text.y=element_text(size=15,"bold"),axis.title=element_text(size=14,"bold"))+theme(legend.text=element_text(size=rel(1.2)))
```

A bar graph showing the distribution of life expectancies by regions.

```
ggplot(expec,aes(regions,fill=expectancy1))+geom_bar() + theme_bw() + labs(y="Life Expectancy", x="Regions",fill="Regions") + theme(axis.text.x=element_text(size=15,angle = 45,hjust = 1,"bold"),axis.text.y=element_text(size=12,"bold"),axis.title=element_text(size=14,"bold"))
```

Summary statistics for Regions with Lowest and Highest Life expectancies.

```
expec3<-subset(expec,regions==c("Central","Rift Valley"),select= c(-year,-county,-regions,-expectancy1))
expec4<-subset(expec,regions==c("Nyanza","North Eastern"),select= c(-year,-county,-regions,-expectancy1))
st(expec3,factor.counts = FALSE)

st(expec4,factor.coounts= FALSE)
```

RANDOM FOREST MODEL

```
library(randomForest)

library(caret)#Access varImp function

r1<-read.csv(file.choose())
r2<-r1[,c(-1,-2,-15,-16)]
```

Splitting the datasets to test and training sets.

```
set.seed(123)
trainsamp<-r2$life_expectancy %>% createDataPartition(p=0.8,list = FALSE)
```

```
train<-r2[trainsamp,]
test<-r2[-trainsamp,]
```

Selecting the best Mtry

```
bestmtry<-tuneRF(train,train$life_expectancy,stepFactor = 1.2,improve=0.01,trace=T,plot=T)
```

Creating a random forest model.

```
model<-randomForest(life_expectancy~.,data=train)
model
```

Getting the important variables details while predicting life expectancy.

```
importance(model)
```

Plotting Importance plots.

```
varImpPlot(model)
```

The importance plot explains how the model found out the variables above to be important in predicting life expectancy. Making predictions on test data

Making the predictions and getting R2 and RMSE values

```
pred_test<-predict(model,newdata =
                    test)

data.frame(RMSE=RMSE(pred_test,test$life_expectancy),
R2=R2(pred_test,test$life_expectancy))
```

PRINCIPAL COMPONENT REGRESSION

```
library(pls)

x1<-read.csv(file.choose())
x2<-x1[,c(-1,-2,-7,-16)]
set.seed(1000)
model<-pcr(life_expectancy~.,data=x2,scale=TRUE,validation="CV")
summary(model)
```

#VALIDATION RMSEP. This table tells us the test RMSE calculated by the k-fold cross validation. If we only use the intercept term in the model, the test RMSE is 3.648. If we add in the first principal component, the test RMSE drops to3.460.We can conclude that adding additional components actually leads to an increase in test RMSE. Thus it appears that it would be optimal to use only two principal components in the final model. ##2. TRAINING : % VARIANCE EXPLAINED. This table tells us the percentage of the variance in the response variable explained by the principle components. We can see the following: -By using the first PC, we can explain 34.62 of the variation in the response variable. -By adding in the second PC, we can explain 55.23% of the variation in the response variable. This implies that we will be able to explain more variation by adding more PC's. Visualizing Cross Validation plots.

Plotting validation plots.

```
validationplot(model)

validationplot(model,val.type="MSEP")
```

```
validationplot(model,val.type="R2")
```

From each plot we can be able to see that model fit improves by adding more PC’s.Thus the optimal model would include the 11 PC’s. Defining training and test sets. Splitting the dataset to training and test sets.

```
set.seed(1000)
names(x2)

##  [1] "life_expectancy"    "population"         "GDP_mill"
##  [4] "county.expenditure" "adult.mort"         "poverty_rate"
##  [7] "under_five.mort"    "HIV_prev"           "BCG"
## [10] "diptheria"          "hepatitis"          "polio"

train<-x2[1:120,]
y_test<-x2[120:188,1]
test<-x2[120:188,1:12]
```

Using the model to make predictions on a test set.

```
library(caret)
model1<-pcr(life_expectancy~population+GDP_mill+county.expenditure+adult.mort+poverty_rate+under_five.mort+HIV_prev+BCG+diptheria+hepatitis+polio,data=train,scale=TRUE)

pcr_pred<-predict(model1,test,ncomp=5.1)
sqrt(mean((pcr_pred-y_test)^2))
data.frame(
  RMSE=caret::RMSE(pcr_pred,test$life_expectancy),
  R2=caret::R2(pcr_pred,test$life_expectancy)
)
```

Taken together cross-validation identifies ncomp=11 as the optimal number of pcs that minimize the prediction error(RMSE) and explains enough variation in the predictors and in the outcome. 100% of the information contained in the predictors are captured by 11 principal components(ncomp=11). Additionally, setting ncomp=11, captures 74.56% of the information in the outcome variable(life_expectancy), which is good. Calculation RMSE.

MULTIPLE LINEAR REGRESSION

```
library(magrittr)
library(dplyr)
library(tidyverse)
library(caret)
x3<-read.csv(file.choose())

#Subsetting the data.
x4<-x3[,c(-1,-2,-15,-16)]
```

FITTING MULTIPLE LINEAR REGRESSION MODEL.

```
x<-lm(life_expectancy~.,data=x4)
#Getting residual plots.
plot(x)
```

The model meets all the assumptions of linear regression as per the residual plots.

Splitting the datasets into test and training in order to build and test model prediction

```
set.seed(123)
samp1<-x4$life_expectancy %>% createDataPartition(p=0.8,list=FALSE)
trainsamp<-x4[samp1,]
testsamp<-x4[-samp1,]
lm2<-lm(life_expectancy~.,data=trainsamp)
#making predictions
predictions<-lm2 %>% predict(testsamp)
#checking for RMSE and R2
data.frame(
  RMSE=RMSE(predictions,testsamp$life_expectancy),
  R2=R2(predictions,testsamp$life_expectancy)
)
```

FROM PYTHON.

```
In [1]:

import pandas as pd

import matplotlib.pyplot as plt

import numpy as np

from sklearn.tree import DecisionTreeClassifier

from sklearn.tree import DecisionTreeRegressor

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score

from sklearn.ensemble import RandomForestRegressor

from sklearn import metrics

from sklearn import tree

from io import StringIO

from IPython.display import Image

from sklearn.tree import export_graphviz

import pydotplus

import pydot

import sys

from sklearn import svm

from sklearn import model_selection

from statsmodels.tools.eval_measures import mse

from sklearn.metrics import mean_squared_error

from sklearn.linear_model import Ridge

import statsmodels.api as sm

In [ ]:

data=pd.read_csv(r'C:\Users\SILA\OneDrive\Desktop\Life Expectancy Data\dataproject.csv')

In [25]:
```

```
data.columns
In [26]:
dummies=pd.get_dummies(data[['county', 'year', 'life_expectancy', 'population', 'GDP_mill',
       'county.expenditure', 'adult.mort', 'poverty_rate', 'under_five.mort',
       'HIV_prev', 'BCG', 'diptheria', 'hepatitis', 'polio', 'expectancy1']])
X2=(dummies-dummies.min())/(dummies.max()-dummies.min())
X=X2.drop(columns=["life_expectancy"])
Y=X2["life_expectancy"]
X_train,X_test,Y_train,Y_test=model_selection.train_test_split(X,Y,test_size=.3,random_state=1)
In [27]:
h =svm.SVR(kernel='rbf')
h.fit(X_train,Y_train)
Y_pred=h.predict(X_test)
print(mse(Y_pred,Y_test))
h.score(X_test,Y_test)
metrics.explained_variance_score(Y_test,Y_pred)
In [28]:
model=Ridge()
model.fit(X_train,Y_train)
model.score(X_test,Y_test)
predicts=model.predict(X_test)
print(mse(predicts,Y_test))
metrics.explained_variance_score(Y_test,predicts)
In [53]:
model1=DecisionTreeRegressor()
model1.fit(X_train,Y_train)
predictions=model1.predict(X_test)
print(mse(predictions,Y_test))
metrics.explained_variance_score(Y_test,predictions)
```

```
In [13]:

fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (4,4), dpi=300)

tree.plot_tree(model1,filled=True,rounded=True)

plt.savefig('C:/ProgramData/jupyter/svm.png')


In [113]:

rf= RandomForestRegressor()

rf.fit(X_train,Y_train)

rfpredict=rf.predict(X_test)

print(mse(rfpredict,Y_test))

metrics.explained_variance_score(Y_test,rfpredict)

In [91]:

dummies1=pd.get_dummies(data[[ 'life_expectancy', 'population', 'GDP_mill',

       'county.expenditure', 'adult.mort', 'poverty_rate', 'under_five.mort',

       'HIV_prev', 'BCG', 'diptheria', 'hepatitis', 'polio']])

X3=(dummies1-dummies1.min())/(dummies1.max()-dummies1.min())

X2=X3.drop(columns=["life_expectancy"])

Y2=X3["life_expectancy"]

X2_train,X2_test,Y2_train,Y2_test=model_selection.train_test_split(X2,Y2,test_size=.3,random_state=1)

In [92]:

X2_train_sm=sm.add_constant(X2_train)

lm=sm.OLS(Y2_train,X2_train_sm).fit()

lm.params
```