Prediction of Life Expectancy using Support Vector Machine (SVM) Algorithm

Prepared by

Mmaduanusi Arinze Uche

(C2009193)

Applied Data Science

Data Science Foundation ICA

January 11, 2023

Table of Contents **Page**

List of Tables **Page**

List of Figures

Abstract

Life expectancy refers to a statistical and analytical measure of the average period an organism is expected to live based on their birth year, current age, and many other demographic factors. With more coverage than the narrow metric of using infant and child mortality which only focus on mortality at a young age, life expectancy is the key metric for assessing population health, capturing mortality throughout the entire life course. Factors considered in this study include schooling, income composition of resources, BMI, GDP, population, development status of a country, alcohol consumption rate, government expenditure on health, measles cases, HIV/AIDS deaths, immunization coverage, thinness diseases, per capita expenditure on health, and mortality rates in infants and adults. The aim of this study is to predict life expectancy automatically using support vector machine (SVM) through optimizing performance metrics and exploring factors that statistically contribute mostly to life expectancy. Indicated by the result, schooling has the highest positive correlation of **+0.71** while adult mortality has the strongest negative correlation of **-0.70** with life expectancy respectively. Also, the SVM linear machine learning algorithm performed well by predicting the life expectancy with an **r-square** value of **0.7399 and mean absolute error of 3.1850** after optimizing the model.

Keywords: Life expectancy, Support Vector Machine (SVM), Machine Learning, Linear Regression, Exploratory Data Analysis (EDA), Performance Parameter.

Problem Identification

Efforts to achieve a sustainable society by optimizing longevity through healthier living and other socio-economic factors are materializing in many countries while still dragging in others. Life expectancy remains the most accurate statistical measure of the average time an individual is expected to live. Thus, accurate prognostication of life expectancy using relevant factors like adult mortality, infant death, BMI, alcohol consumption, diseases, socio-economic conditions, and demographic factors would be very instrumental to accurate health decision making and improving health services. This problem statement provides a way to predict life expectancy of people living in a country considering their demographic, socio-economic, and immunization factors including year of each occurrence.

Introduction

Life expectancy refers to a statistical and analytical measure of the average time an organism is expected to live based on its birth year, current age, and many other demographic factors like gender, lifestyle, location, education, etc.**[1]** The common and most used measure for life expectancy is **life expectancy at birth** *(LEB)*, which could be described using two dimensions: **cohort LEB**, meaning length of life of a birth cohort (all individuals born in a given year) and can be computed only for cohorts born so long ago that all their members have died and **period LEB** is the mean length of life of a hypothetical cohort **[1][2]** assumed to be exposed, from birth through death, to the mortality rate observed at a given year **[3]**. Early civilizations experienced the all-time highest mortality rates and lowest life expectancies at every age for both male and female genders. For example, in the bronze and iron ages, human LEB was 26 years, and this is due to diseases, wars, famine, lack of access to good health system and basic amenities **[4]**. However, in recent years, LEB in Eswatini (formerly known as Swaziland) is 49, while LEB in Japan is 83. It was observed that factors like high infant mortality, death of adolescents from accidents, diseases, poverty, lack of education, wars, and childbirth, before the universal access to modern medicine significantly affect LEB **[5]**. Over the past two centuries, human life expectancy has taken a remarkable positive turn, and this was achieved during the demographic transition of societies from regimes of high mortality to regimes of low mortality rates **[6][7]**. Japanese women currently have the highest life expectancy at birth which is above 87 years. In 1840, the record was held by Swedish women, with an average life span of 46 years **[8]**. This development ushered in an increase in longevity because of optimized quality of life and lifestyle changes.

Background of Study

Many previous studies considered factors affecting life expectancy from a selective perspective of either demographics, income composition, mortality factors, or a few combinations of these. Factors like alcohol consumption, education, gross domestic product, and population size were somewhat overlooked. Moreso, these studies were conducted using multiple linear regression models based on dataset of one year for one or more of the countries. However, this study seeks to take a more holistic approach by considering many factors that significantly contribute to life expectancy, taking account of multiple data points obtained from World Health Organization (WHO) data repository from a period of 15 years for 193 countries and investigate their inter-dependencies. Additionally, this study will use support vector machine learning algorithm to predict life expectancy. Since the observations of this dataset are based on actual datapoints from different countries, the result of this study will provide useful insights to the government of these countries to make informed decisions that will efficiently improve life expectancy since the significant factors contributing to lower value of life expectancy would be investigated.

Scope and Limitation of Study

This study relies completely on the accuracy of the data. Dataset was collected from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO) data repository website and its corresponding economic data was collected from United Nation website**[9].** The datasets are made available to the public for the purpose of health data analysis. Only the most significant health-related factors were chosen for this analysis to provide a more specific result. Observations suggest that in the recent past, the human mortality rate has declined drastically in the developing countries due to improvements in healthcare as compared to three decades ago. On this account, this study will use 15 years data (2000 to 2015) from 193 countries for further analysis.

Justification of Study

i. Life expectancy prediction has a greater impact in our modern society because of our lifestyle, different disease types, social-economic factors, and ever-changing environmental conditions.
ii. Life expectancy provides one of the most important factors in end-of-life decision making, therefore, a good prediction would help in many ways to optimize the quality of human life.
iii. The prediction of life expectancy will provide useful insights that will support good decision/policy making by the government to enhance human longevity using key metrics like education and income composition of resources.
iv. As an essential metric for measuring human longevity, accurate prediction of life expectancy would be very pivotal in healthcare management and advanced care planning.

Study Aim and Objectives

This research has two main aims:

i. To critically investigate the factors with the highest impact on life expectancy among the range of presented attributes.
ii. To predict human life expectancy using support vector machine (SVM) learning algorithm.

Literature Reviews

The subject of life expectancy has gained remarkable contributions from different authors and researchers over the years as it is the most reliable measure to determine the overall mortality level of a population. It also serves as a pointer to the government for effective resource allocation and economic optimization. Inferences from some studies pointed towards birth year and gender as the chief dependent factors of life expectancy in developed countries, while some studies established that illiteracy, infant mortality, and diseases are the leading contributors of low life expectancy in the developing countries, especially in Africa. This study will, however, consider the above-mentioned factors and several others using SVM learning algorithm to accurately arrive at a decision.

Chetty *et al*., (2016) investigated the relationship between life expectancy and income in the United States between 2001 and 2014 using a simple linear regression method **[10].** They established that income and longevity have a linear relationship, and differences in life expectancy across income groups increased over time.

Kasichainula *et al.,* (2020) applied several machine learning methods of both regression and classification to predict life expectancy. They eventually adopted a random forest classifier because it delivered the better accuracy of 89% **[11].**

Ayshwaryaa, N. *et al.,* (2020) predicted human life span using multiple machine learning algorithms. They monitored the correlation between attributes like diseases, gender, ages, and environmental factors and reached an accuracy of 90% when predicted with random forest algorithm **[12].**

Khulood, K.F. et al., (2021) and his team critically investigated life expectancy based on machine learning and structured predictors using four different algorithms: linear regression, decision tree regression, random forest regressor and voting regressor. Upon comparative analysis, voting regressor gave the best accuracy **[13].**

Bhosale, A.A. and Sundaram, K.K., (2019), proposed that expectancy of the life mainly target on predicting models using trends. They proposed life expectancy considering weight, adult mortality, heart rate, and respiration rates. The inspection provides the standard life expectancy forecasted by variables that can be easily calculated **[14].**

Mahumud, R.A. *et al.,* (2013) implicitly investigated the relationship and correlation between life expectancy and educational background of people, health plans, economic stability, diseases, BMI and how the lifestyle of the people in Bangladesh also affect their life expectancy. They used a linear regression algorithm and the model showed that education has a very positive correlation to life expectancy. This means that the more people are educated, the more aware they become about healthy living and the risks of unhealthy practices **[15].**

Byvatov and Schneider (2003) applied SVM in bioinformatic classification of small organic molecules that potentially modulate the function of G-protein coupled receptors. Their report stated that SVM showed a classification accuracy of 90% with a correlation of 0.78 **[16].**

Longlong Zhang et. al., (2013) applied SVM in the review of remaining useful life prediction for engineering application. They discovered that SVM could accurately predict this because it has been proven to solve problems considering structure risk and has better generalization ability compared with conventional machine learning methods such as artificial neural networks. **[17]**

From the above study reports, many researchers used different ML algorithms based on either accuracy or choice. In this study, I seek to use a different measure that has not gained much use in this subject to arrive at a similar or even a better result.

I will use SVM algorithm because it has been proven to produce lower prediction error as compared to regressors like Linear Regression and Logistic Regression, and classifiers like K-Nearest Neighbor based on other methods like artificial neural networks, especially when large numbers of features are considered for sample description.

Additionally, as the target label has continuous data that spans across years, it has become a regression case that will use a suitable regressor with high predictability and low errors which SVM regressor could conveniently provide.

Materials and Methods

Description of Data Set

This dataset contains 22 attributes and 2938 instances, and the description of the features with their respective data types is as shown in table 1. The dataset was downloaded from Kaggle which was originally obtained from the World Health Organization (WHO) data repository website for 15 years (2000 – 2015) dataset concerning life expectancy for 193 countries. The corresponding socio-economic datapoints were collected from the United Nations database[9]. The data set will be used to predict the life expectancy of each represented country for the period of 15 years.

Data Preparation and Loading into RStudio

This data set was first prepared in Microsoft Excel by moving the target label (life Expectancy) from column 4 to column 22 for easy analysis. The data was then loaded into the RStudio as a csv file for pre-processing and analysis. The structure and summary were checked to confirm the data type of each feature, their statistical characteristics, and the presence of null values.

Table 1: Life expectancy dataset with data types (Dataset from Kaggle – WHO database)

S/N	Features	Descriptions	Data Type	Null Values
1	Country	Names of countries under study and their number 193	Character	0
2	Year	The year from which the information was collected, and the years from 2000 to 2015	Integer	0
3	Status	Clarifies the country's status defined by developed or developing	Character	0
4	Adult Mortality	The mortality rate for adults of both sexes for those aged from 16 to 60 years per 1000 population	Integer	10
5	Infant Deaths	The number of infant deaths in the country during the specified year per 1000 population	Integer	0
6	Alcohol	The rate of consumption per capita in the country in litters of pure alcohol	Numeric	194
7	Percentage Expenditure	Expenditure on health as a percentage of gross domestic product per capita	Numeric	0
8	Hepatitis B	The percentage of one-year-old children vaccinated for Hepatitis B	Integer	553
9	Measles	Number of reported cases of people with measles per 1000 population	Integer	0
10	BMI	Abbreviation for Body Mass Index, and it calculates the average for the whole population	Numeric	34
11	Under-Five Deaths	Number of deaths of children aged less than 5years per 1000 population	Integer	0
12	Polio	The percentage of one-year-old children that immunized for Polio (Pol3)	Integer	19
13	Total Expenditure	The percentage of government expenditure on health relative to total expenditure	Numeric	226
14	Diphtheria	The percentage of one-year-old children that immunized for diphtheria-tetanus toxoid and pertussis (DTP3)	Integer	19
15	HIV/AIDS	Number of deaths in children in the age from 0 to 4 years, because of HIV / AIDS for 1 per 1000 population	Numeric	0
16	GDP	Abbreviation for Gross Domestic Product, and it calculates the per capita (in USD)	Numeric	448
17	Population	Census of the population of the country during the specified year	Character	0
18	Thinness 10 -19 Years	The percentage of thinness among children and adolescents between the ages of 10 to 19 years	Numeric	34
19	Thinness 5-9 Years	The percentage of thinness among children between the ages of 5 to 9 years	Numeric	34
20	Income Composition of Resources	A human development index ranging from 0 to 1 in terms of income composition for resources	Numeric	167
21	Schooling	The number of educational years	Numeric	163
22	Life Expectancy	Data of life expectancy for each countries represented	Numeric	10

Data Pre-processing

Since quality dataset leads to better models and more accurate predictions, data pre-processing has become a very vital step and fundamental integral part of the data science processes. Data pre-processing involves data cleaning in the form of casting, dealing with null values etc. to increase the quality of the model and increase the prediction accuracy.

Data Cleaning

The objective for data cleaning is to find the easiest way to rectify quality issues. Here, the data type of population column was converted into numeric from character. The missing or null values in all numeric data were fixed with median values. If mean or mode were used, it will give less accurate results because most of the features have skewed distribution.

Exploratory Data Analysis (EDA)

EDA is one of the critical steps to take in data analysis to investigate dataset and summarize its significant features using statistical graphics or any other data visualization methods. It essentially aims at spotting patterns and trends, identifying anomalies, and testing theories. The density plot and histogram distribution show various distribution of other attributes with the target label (life expectancy). From figure 1, HIV/AIDS, GDP, Infant Deaths, Measles, Percentage Expenditure, Population, and Under Five Deaths followed an exponential distribution, hence would be dropped. One of the Thinness would be dropped since they both have the same shape and behavior. Figure 2 shows that attributes like Adult Mortality, Alcohol, Thinness have negatively skewed distribution while Diphtheria, Hepatitis B, and Polio are positively skewed. Schooling, Income Composition of Resources, Life Expectancy, and Total Expenditure have near normal or Gaussian distribution. BMI clearly has a bimodal distribution.

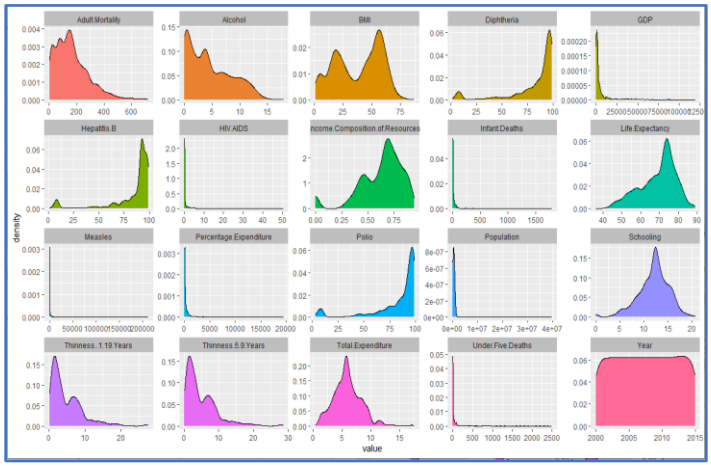

Figure 1: Density plots of features

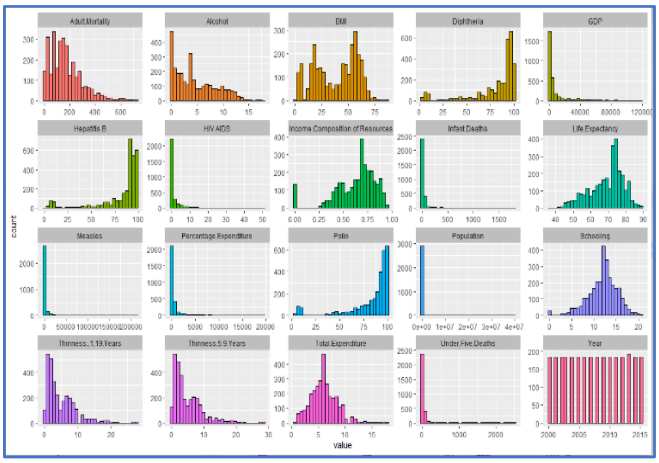

Figure 2: Histogram distribution of features

The box plot in figure 3 shows how the dataset performed in terms of spread and outliers. All numerical independent variables have outliers except for BMI (and Year). Income Composition of Resources has a little outlier and positively skewed. The correlation matrix in figure 4 shows that only four features have a strong correlation with the target. These include Schooling, Income Composition of Resources, BMI, and Adult Mortality, with schooling having the highest positive correlation, and Adult Mortality having the highest negative correlation. Correlation is of two types: positive and negative correlation. Positive correlation implies that the feature increases the target increase and vice versa. Negative correlation means that an increase in the features yields a corresponding decrease in the target, essentially, they have opposite or inverse relationship.

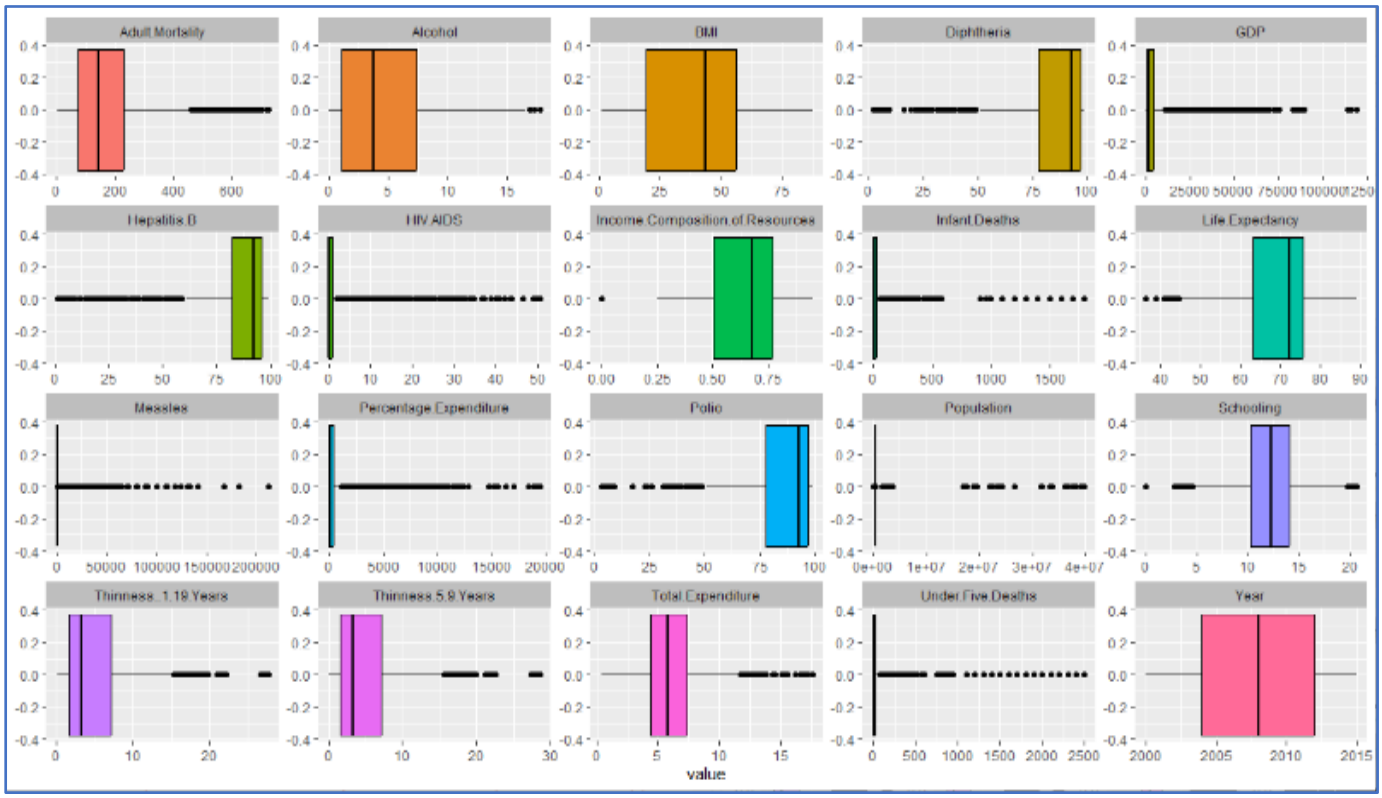

Figure 3: Box plot for outliers in features

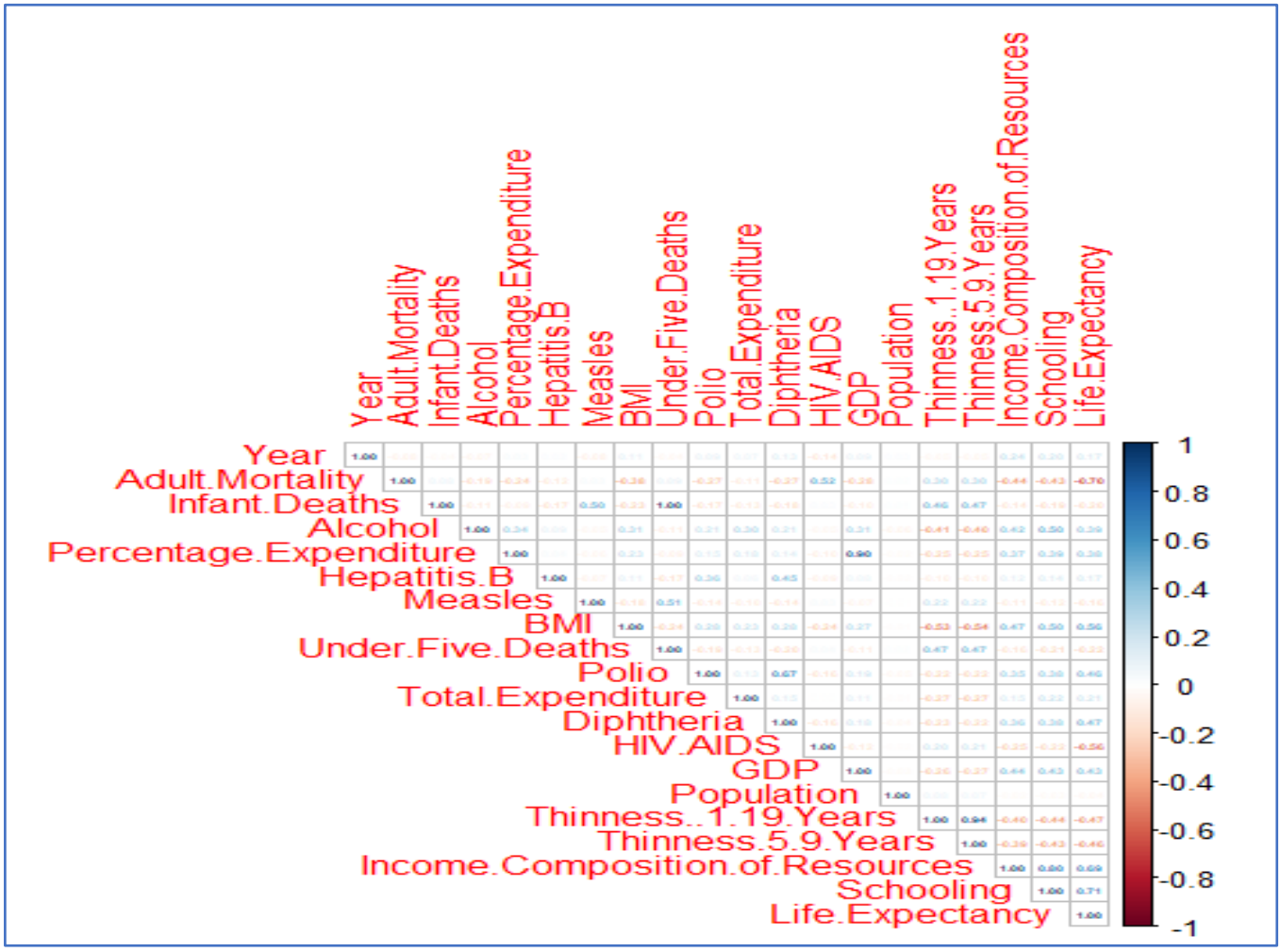

Figure 4: Correlation matrix of features and target label

Optimization of Dataset

The correlation matrix confirmed the redundant columns that followed exponential distribution as depicted on the density and histogram plots. They are removed to improve model accuracy. Categorical variables like country and status were encoded with Label Encoder before fitting them into the model. Machine learning generally performs better when the features have a similar scale and near normal distribution. Thus, a standard scaler was chosen which works by extracting the mean from each value and dividing by the standard deviation. This method is called standardization of data and it is chosen over normalization because normalization requires data to have a normal distribution while standardization works with either normal or skewed distribution. The mathematical expression of standardization is shown below.

$$Standardization = \frac{x - mean(x)}{Standard\ deviation}$$

Table 2: Selected features for analysis (Dataset from Kaggle – WHO database)

S/N	Selected Features	Descriptions	Data Type
1	Country	Names of countries under study and their number 193	Numeric
2	Year	The year from which the information was collected, and the years from 2000 to 2015	Integer
3	Status	Clarifies the country's status defined by developed or developing	Numeric
4	Adult Mortality	The mortality rate for adults of both sexes for those aged from 16 to 60 years per 1000 population	Integer
5	Alcohol	The rate of consumption per capita in the country in litters of pure alcohol	Numeric
6	Hepatitis B	The percentage of one-year-old children vaccinated for Hepatitis B	Integer
7	BMI	Abbreviation for Body Mass Index, and it calculates the average for the whole population	Numeric
8	Polio	The percentage of one-year-old children that immunized for Polio (Pol3)	Integer
9	Total Expenditure	The percentage of government expenditure on health relative to total expenditure	Numeric
10	Diphtheria	The percentage of one-year-old children that immunized for diphtheria-tetanus toxoid and pertussis (DTP3)	Integer
11	Thinness 10 -19 Years	The percentage of thinness among children and adolescents between the ages of 10 to 19 years	Numeric
12	Income Composition of Resources	A human development index ranging from 0 to 1 in terms of income composition for resources	Numeric
13	Schooling	The number of educational years	Numeric
14	Life Expectancy	Data of life expectancy for each countries represented	Numeric

Technical Implementation

Data was prepared for modeling by splitting it into training and testing sets, a process otherwise known as data splicing. An important step before the split was to initialize the split procedure by setting up the training and testing in 80:20 ratio using the **createDataPartition()** function. This means that the machine learning model would be trained with 80% of the dataset and be tested with the remaining 20%. SVM (using the regression kernel) is the chosen machine learning algorithm for this analysis because the dataset is a supervised one with the target label having continuous values. The goal of the SVM is to find a distinctive divisor called the hyperplane in the N-dimensional space that uniquely classifies multiple data values. The designed model was used to predict life expectancy and the accuracy was measured with the key performance metrics.

Performance Evaluation

The metrics applied in measuring the performance and accuracy of this model include mean absolute error, root mean square error, and r square.

Mean Absolute Error (MAE): Refers to the average of all absolute errors without considering their cardinality. Absolute error is the difference between the measured value and the absolute value. MAE measures accuracy for continuous variables and could be mathematically expressed as below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

where,

n = number of errors, $\sum$ = summation symbol, and $|x_i - x|$ = absolute error

Root Mean Square Error (RMSE): This is a statistical metric that measures the difference between the predicted values and the observed values. Essentially, it tells how far apart the predicted values are from the observed values in a regression analysis. The lower the RMSE value, the better the model can 'fit' the dataset. It is mathematically expressed below.

$$RMSE = \sqrt{\frac{\sum(P_i - O_i)^2}{n}}$$

where,

P_i = Predicted value for the ith observation in the dataset, O_i = Observed value for the ith observation in the dataset, and

n = number of observations or instances

R-Squared (R^2): This measures the difference or variance in statistical terms for a dependent variable that an independent variable(s) can explain. It simply defines the extent the data fits the regression model. The R-Squared value of 1 is a perfect fit while that of 0 is a worst fit. R-Squared could be mathematically expressed using the formula below.

$$R^2 \quad \left[\frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2} - (\sum x)^2 * \sqrt{n\sum y^2} - (\sum y)^2}\right]^2$$

where,

n = number of instances in the given dataset, x = first variable in the context, and y = second variable in the context.

Table 3: Performance metrics and accuracy at different tune lengths

Table of Accuracy using Performance Metrics								
Tune length	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MAE	3.1848	3.1850	3.1853	3.1850	3.1866	3.1846	3.1866	3.1858
RMSE	4.8535	4.8509	4.8537	4.8540	4.8551	4.8462	4.8586	4.8556
R^2	0.7396	0.7399	0.7396	0.7396	0.7394	0.7396	0.7396	0.7394

Results

The result of the trained SVM model was optimized by iterating the model at varying tune lengths from **3 to 10** until better accuracy of the performance metrics was achieved. **Table 3** shows the different values of the performance metrics at different tune lengths, and it was observed that the maximum value of $\mathbf{r^2}$ value was obtained at the tune length of **4**. This tune length was used to retrain the model and a graph of predicted vs. observed was plotted showing a linear relationship between both values as shown in **figure 9**. The graph of the performance metrics at different tune lengths are displayed in figures **11 to 13** where the maximum $\mathbf{r^2}$ was recorded at tune length of **4** while RMSE recorded the highest value at **9**. MAE recorded two identical highest values at **7 & 9.**

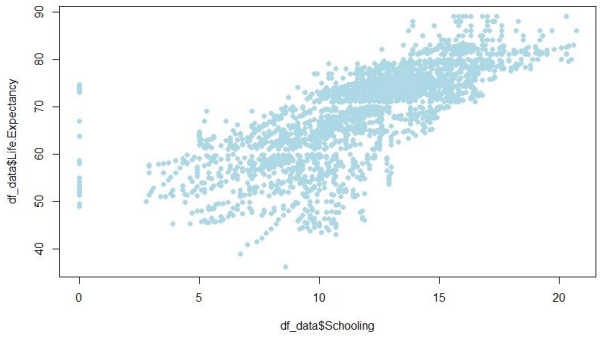

Figure 5: Life expectancy vs. Schooling

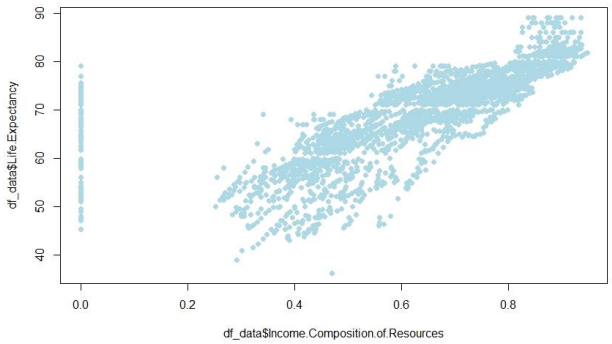

Figure 6: Life expectancy vs. Income composition of resources

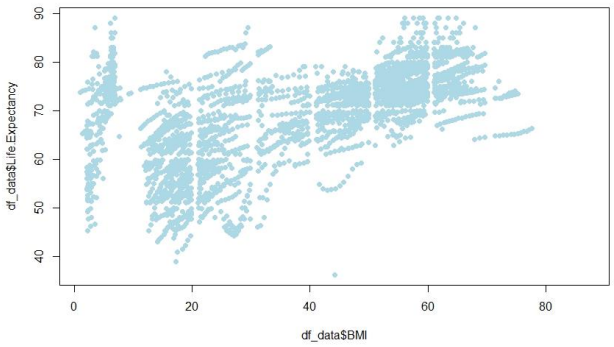

Figure 7: Life expectancy vs. BMI

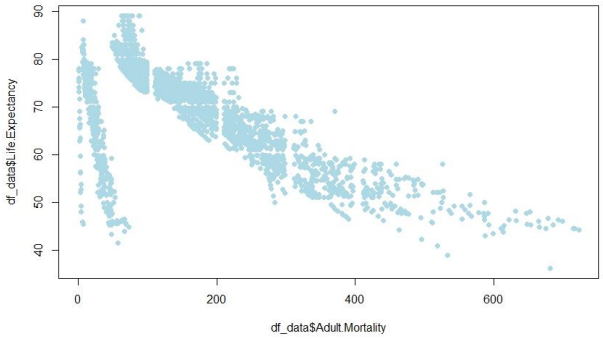

Figure 8: Life expectancy vs. Adult mortality

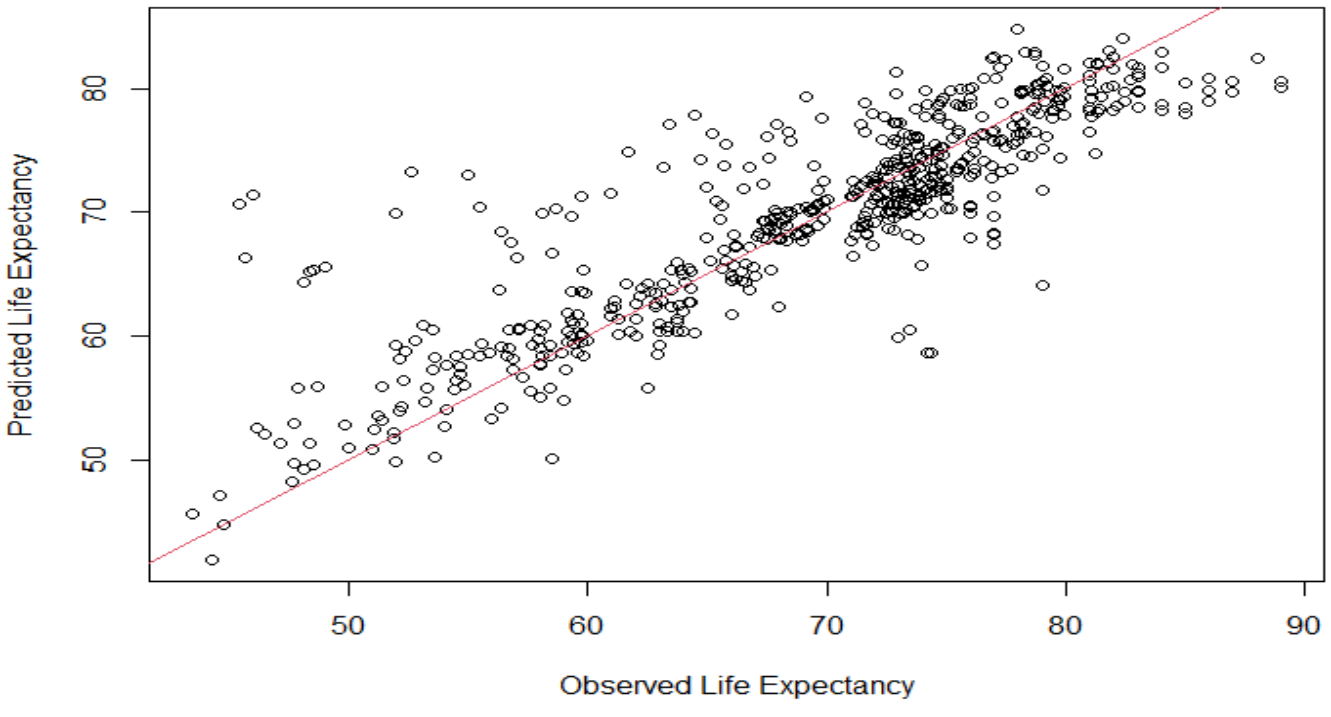

Figure 9: Plot of predicted vs. observed life expectancy

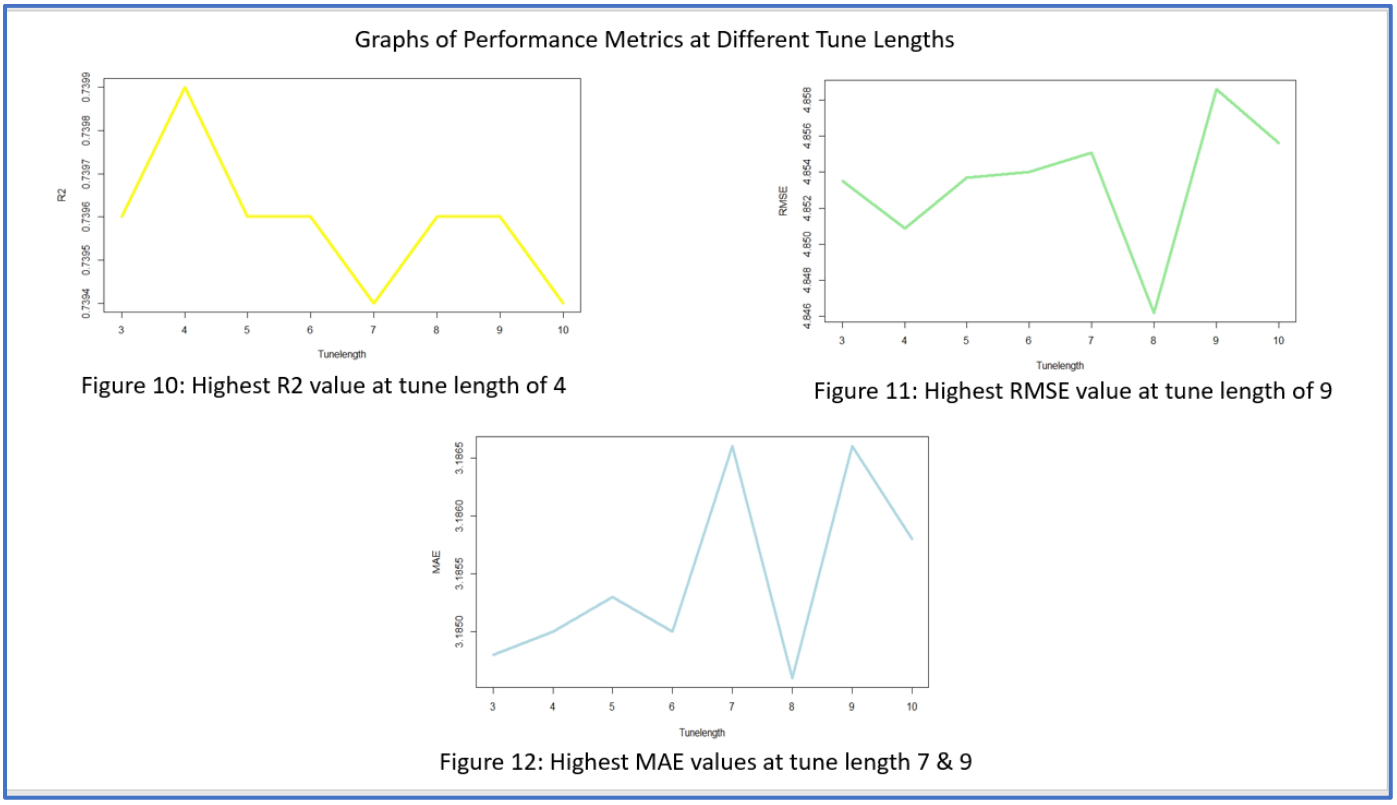

Figure 10: Highest R2 value at tune length of 4

Figure 11: Highest RMSE value at tune length of 9

Figure 12: Highest MAE values at tune length 7 & 9

Discussion

Limitation: Prior to properly understanding the dataset and how machine learning algorithms work in general, I tried using KNN algorithm for this analysis but got stuck at a point. I later understood by research that it did not work because my target label has continuous variables. Thus, knowing that it is a regression problem instead, I picked SVM regressor (after trying out multiple linear regression initially) because it has higher predicted performance according to literature reviews of previous works.

As seen in figure 4, among all the observed factors affecting life expectancy, schooling has the highest positive correlation of **+0.71** which means that the more literate people become, the more they improve their life expectancy. Also, adult mortality has the highest negative correlation of **-0.70** with life expectancy and this implies that as adult mortality increases, life expectancy is reduced and vice versa. The correlation graphs of all factors having a high correlation with life expectancy are plotted in figures **5 to 8.**

The SVM linear machine learning model predicted the target with a good r-squared accuracy of **0.7399 (~74%)** after the optimization process and this means that it could be adopted as the suitable algorithm for life expectancy prediction.

The model could have performed better if more features from geographical, economical, and financial evaluations were included because those factors could make major impact in the longevity of a population.

Nevertheless, recalling from the aim of this study which is to investigate relevant factors affecting life expectancy and to predict life expectancy using SVM algorithm, it is safe to summarize that when seeking a way to improve life expectancy, schooling and adult mortality should be critically considered as they are the major contributors.

Conclusion and Future Work

This study described how life expectancy of countries was predicted using a support vector machine (regression) algorithm. It also pointed out the factors with the most correlation with life expectancy which schooling, BMI, Income composition of resources and adult mortality rate took the lead. Through the various stages of designing and deploying of the SVM algorithm, the accuracy of the various performance metrics like R2, RMSE & MAE were optimized to their best obtainable values. I also established that as factors correlate more with life expectancy, the prediction becomes more accurate and provides stronger insights to government of each country on the best sectors to allocate more resources to.

Lastly, these predicted factors that may affect the life expectancy in the future would assist government and healthcare organizations make adequate preparations for circumstances such as pandemics, scarcity etc. This study could further extend to more research by adding more features like years and geographical data to the existing dataset. More advanced algorithms like deep learning could be used to get better results as well as adding more performance metrics could help optimize the results.

References

[1] E. Ortiz-Ospina, “Life Expectancy” – What does this actually mean?" Our World in Data, 28 8 2017. [Accessed 19 2 2021].

[2] "Period and cohort life expectancy explained: December 2019 – Office for National Statistics". www.ons.gov.uk. Retrieved August 31, 2020.

[3] S. Shryock, J. S. Siegel et al. The Methods, and Materials of Demography. Washington, DC, US Bureau of the Census, 1973

[4] J. Oeppen, J. W. Vaupel, Broken limits to life expectancy. *Science* **296**, 1029–1031 (2002).

[5] J. C. Riley, *Rising Life Expectancy: A Global History* (Cambridge University Press, 2001).

[6] D. Kirk, Demographic transition theory. *Population Study (Camb.)* **50**, 361–387 (1996).

[7] F. W. Notestein, “Population—the long view” in *Food for the World*, T. Schulz, Ed. (University of Chicago Press, Chicago, 1945), pp. 36–57.

[8] University of California, Berkeley; Max Planck Institute for Demographic Research (Germany), Human mortality database. https://www.mortality.org/. Accessed 22 July 2019.

[9] K. Rajarshi, "Life Expectancy (WHO)," Kaggle, 2017. [Accessed 15 9 2020].

[10] Chetty R, Stepner M, Abraham S, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. *JAMA.* 2016;315(16):1750–1766. doi:10.1001/jama.2016.4226

[11] Kasichainula, V. et al., (2020) ‘Machine Learning Techniques for Life Expectancy Prediction’, *International Journal of Advanced Trends in Computer Science and Engineering*, 9(4), 4503 – 4507

[12] Ayshwaryaa, N. et al., (2020) ‘Human Life Span Prediction using Machine Learning; International Journal for Research in Applied Science & Engineering Technology, 8(9) ISBN: 2321 – 9653.

[13] Khulood, K.F. et al., (2021) ‘Life Expectancy Estimation based on Machine Learning and Structured Predictors’, 3rd International Conference on Advanced Information Science and System (AISS 2021), pp 8.

[14] Bhosale, A.A. and Sundaram, K.K., (2019) ‘Life prediction equation for human beings, ‘*International Conference on Bioinformatics and Biomedical Technology*, vol. IEEE, pp. 266-268.

[15] Mahumud, R.A. et al., (2013) ‘Impact of Life Expectancy on Economic Growth and Health Care Expenditures in Bangladesh.,’ *Universal Journal of Public Health*, 1(4) pp. 180-186.

[16] Byvatov E, Schneider G. Support vector machine applications in bioinformatics. Appl Bioinformatics. 2003;2(2):67-77. PMID: 15130823

[17] L. Zhang, Z. Liu, D. Luo, J. Li, and H. -Z. Huang, "Review of remaining useful life prediction using support vector machine for engineering assets," 2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), 2013, pp. 1793-1799, doi: 10.1109/QR2MSE.2013.6625925.

R Package Citation

caret: Classification and Regression Training. R package version 6.0-93. Max Kuhn (2022). https://CRAN.R project.org/package=caret

ggplot2: Elegant graphics for data analysis. Springer-Verlag New York, 2016. H. Wickham. https://ggplot2.tidyverse.org

tidyverse: Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” _Journal of Open-Source Software_, *4*(43), 1686. doi: 10.21105/joss.01686 (URL: https://doi.org/10.21105/joss.01686).

corrplot: Visualization of a correlation matrix version 0.92. Taiyun Wei and Viliam Simko (2021). Available from https://github.com/taiyun/corrplot.

dplyr: A grammar of data manipulation. R package version 1.0.10. Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller (2022). https://CRAN.R-project.org/package=dplyr.

tidyr: Tidy messy data. R package version 1.2.1. Hadley Wickham and Maximilian Girlich (2022). https://CRAN.R-project.org/package=tidyr

CatEncoders: Encoders for categorical variables. R package version 0.1.1. ML Zhang (2017). https://CRAN.R-project.org/package=CatEncoders.

Source Codes

```
#Step 1
#Importing data into a data frame in RStudio and Installing relevant packages
df <- read.csv("life_expectancy.csv")
#Check the structure and summary of dataset
str(df)
summary(df)
#Installing necessary packages
install.packages("caret")
install.packages("ggplot2")
install.packages("tidyverse")
install.packages("corrplot")
install.packages("dplyr")
install.packages("tidyr")
install.packages("CatEncoders")
#Loading packages into libraries
library(caret)
library(ggplot2)
library(tidyverse)
library(corrplot)
library(dplyr)
library(tidyr)
library(CatEncoders)
#Step 2: Data Pre-processing
#Data Cleaning
#1. Converting Population column from character to numeric
df$Population <- as.numeric(df$Population)
df$Population[df$Population == ""] <- ' '
df$Population <- as.numeric(df$Population)
str(df)
#Select only the numeric data leaving out the categorical data for now
df_num <- select_if(df, is.numeric)
#2. Filling the null/missing values with median
for (i in colnames(df_num)) { df_num[[i]] [is.na(df_num[[i]])] <- median(df_num[[i]], na.rm = TRUE)}
#Check for null values after the filling up to confirm the step
anyNA(df_num)
str(df_num)
summary(df_num)
```

```
#Step 3: EDA
#1: Plotting the Histogram distribution to show variable distributions and visualizations
df_num %>%
  gather(key = key, value = value) %>%
 mutate(key = as.factor(key)) %>%
 ggplot(aes(x = value, fill = key)) + geom_histogram(colour = "black") + facet_wrap(~key, scales = "free", ncol = 5) +
 theme(legend.position = "none", strip.background = element_rect(fill = "grey", size = 8))
          strip.text = element_text(colour = "black", face = "bold"), + labs(x = "values", y = "count")
#2: Plotting the Box plot to show variable outliers
df_num %>%
   gather(key = key, value = value) %>%
   mutate(key = as.factor(key)) %>%
   ggplot(aes(x = value, fill = key)) + geom_boxplot(colour = "black") + facet_wrap(~key, scales = "free", ncol = 5) +
   theme(legend.position = "none", strip.background = element_rect(fill = "grey", size = 8))
            strip.text = element_text(colour = "black", face = "bold"), + labs(x = "values", y = "count")
#3: Plotting the density distribution to show variable distributions and visualizations
df_num %>%
   gather(key = key, value = value) %>%
   mutate(key = as.factor(key)) %>%
   ggplot(aes(x = value, fill = key)) + geom_density(colour = "black") + facet_wrap(~key, scales = "free", ncol = 5) +
   theme(legend.position = "none", strip.background = element_rect(fill = "grey", size = 8))
            strip.text = element_text(colour = "black", face = "bold"), + labs(x = "values", y = "count")
#4: Plotting correlation matrix of the variable with the target
df_corplot <- cor(df_num)
corrplot(df_corplot, method = "number", type = "upper", number.cex = 0.30)
#5: Encoding of Categorical data
#Country and Status
Country <- LabelEncoder.fit(df$Country)
df$Country <- transform(Country, df$Country)
Status <- LabelEncoder.fit(df$Status)
df$Status <- transform(Status, df$Status)
#Step 4: Data Set Optimization
#1: Feature selection: selecting the correlated features and dropping redundant ones
drops <- c('Infant.Deaths', 'HIV.AIDS', 'Measles', 'Percentage.Expenditure',
        'Population', 'Under.Five.Deaths', 'GDP', 'Thinness.5.9.Years')
#Save the required features in a separate data frame.
df_data <- df[, !(names(df) %in% drops)]
#2: Standardization of only the predictors
df_std <- df_data %>% mutate_at(c('Country', 'Year', 'Status', 'Adult.Mortality',  'Alcohol', 'Hepatitis.B', 'BMI', 'Polio',
                    'Total.Expenditure', 'Diphtheria',  'Thinness..1.19.Years',  'Income.Composition.of.Resources',
```

```
                    'Schooling'), ~(scale(.) %>% as.vector))
#Filling the missing values again with median after standardization
for (i in colnames(df_std)) { df_std[[i]] [is.na(df_std[[i]])] <- median(df_std[[i]], na.rm = TRUE)}
anyNA(df_std)
summary(df_std)
#3: Plotting correlation graph of features and target
Co_plot1 <- plot(df_data$Schooling,df_data$Life.Expectancy, pch = 19, col = "light blue")
Co_plot1 <- plot(df_data$Income.Composition.of.Resources,df_data$Life.Expectancy, pch = 19, col = "light blue")
Co_plot1 <- plot(df_data$Adult.Mortality,df_data$Life.Expectancy, pch = 19, col = "light blue")
Co_plot1 <- plot(df_data$BMI,df_data$Life.Expectancy, pch = 19, col = "light blue")
#Step 5: Data Splicing
#1: Splitting Data for training and testing models
df_split <- createDataPartition(df_std$Life.Expectancy, p = 0.8, list = FALSE)
df_train <- df_std[df_split,]
df_test <- df_std[-df_split,]
#2: Checking the dimensions of training and testing data frame
dim(df_train)
dim(df_test)
#3: Implement train control method to control all computational overheads
train_ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 10)
#Install additional package
install.packages("e1071")
library(e1071)
#Step 6: Design of SVM ML Algorithm
df_svm.linear <- train(Life.Expectancy~., data = df_train,
                       method = "svmLinear", trControl = train_ctrl,
                       preProcess = c("center","scale"),tunelength = 10)
#Check the result of the trained model
df_svm.linear
#Step 7: Predict with trained model
df_pred <- predict(df_svm.linear, newdata = df_test)
#Step 8: Check for Model Accuracy
# Accuracy check using statistical parameters
df_MAE = MAE(df_test$Life.Expectancy, df_pred)
df_RMSE = RMSE(df_test$Life.Expectancy, df_pred)
df_R2 = R2(df_test$Life.Expectancy, df_pred)
#Step 9: Plot a regression graph for predicted and observed
plot(df_pred~df_test$Life.Expectancy, ylab = 'Predicted Life Expectancy ', xlab = 'Observed Life Expectancy ')
#Adding the correlation line
abline(0,1, col=2)
```

```
#Step 10: Model Optimization
for (i in 3:7){
  df_svm.linear<-train(Life.Expectancy~., data=df_train, method= "svmLinear", trControl = train_ctrl,
                       preprocess = c("center","scale"), tunelength=i)
  print(i)
  print(df_svm.linear)
  }
#Using a tune length of 4
df_svm.linear.4 <- train(Life.Expectancy~., data = df_train,
                       method = "svmLinear", trControl = train_ctrl,
                       preProcess = c("center","scale"),tunelength = 4)
#Predicting with an optimized tune length
df_pred.4 <- predict(df_svm.linear, newdata = df_test)
df_pred.4
#Plotting with the new predicted value at tunelength of 4
plot(df_pred.4~df_test$Life.Expectancy, ylab = 'Predicted Life Expectancy',
     xlab = 'Observed Life Expectancy')
#Adding a correlation line
abline(0,1, col=2)
#Step 11: Graphs of Performance Metrics at Tune lengths
#1: Creating a data frame for the performance metrics
df_met <- data.frame(
  Tunelength = c(3,4,5,6,7,8,9,10),
  MAE = c(3.1848, 3.1850, 3.1853, 3.1850, 3.1866, 3.1846, 3.1866, 3.1858),
  RMSE = c(4.8535, 4.8509, 4.8537, 4.8540, 4.8551, 4.8462, 4.8586, 4.8556),
  R2 = c(0.7396, 0.7399, 0.7396, 0.7396, 0.7394, 0.7396, 0.7396, 0.7394)
)
#2: Plotting the graphs of the performance metrics at varying tune lengths
plot(df_met$Tunelength, df_met$MAE, type = "l", col = "light blue", lwd = 4, xlab = "Tunelength", ylab = "MAE")
plot(df_met$Tunelength, df_met$RMSE, type = "l", col = "light green", lwd = 4, xlab = "Tunelength", ylab = "RMSE")
plot(df_met$Tunelength, df_met$R2, type = "l", col = "yellow", lwd = 4, xlab = "Tunelength", ylab = "R2")
```