Life Expectancy in changing times

Milestone: Project Report

Group 12

Meet Vasani

Mihir Kakadiya

857-757-0787

857-340-9465

vasani.m@northeastern.edu

kakadiya.m@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Meet Vasani

Signature of Student 2: Mihir Kakadiya

Submission Date: 04/21/2023

Table of Content

Problem Setting

The problem focuses on how health of a human depreciates as there is an increase in number of disease. As new disease arise and transform day by day, the life expectancy rates are decreasing. There are various study articles and research which states that the life span is decreasing and not up to the mark as it used to be before. This is an alarming sign and governing body for every nation should take these analysis seriously, as they also indirectly impact on the overall environment. As the time moves on, generation after generation there are various factors which acts as a barrier for humans to survive. These barriers are sometimes not a great deal to humans and they do cause health issues, which indirectly affects the longevity of life. The life span reduces as the intoxication increases. The life expectancy also depend on bad habits of those humans. The ecosystem around plays a major role.

Problem Definition

We have a dataset which contains all sorts of variables which will help us determine the life expectancy of a normal human being based on the various environmental and health conditions at a particular instance of time. From this we can figure out in future what will be the scenario of a particular country based on the facts and given results. Which vaccines are needed most to revive those numbers up. What other hospital utilities are to be requires for the upcoming years can be determined to stop the widespread fear among the population.

Data Sources

The Life Expectancy dataset is taken from Kaggle. The link to the dataset is: Life Expectancy Dataset (WHO).

Data Description

The dataset contains a total of 22 columns and 3000 rows, where in the output or dependent variable gives life expectancy in age. The independent variables includes Immunization related factors, Mortality factors, Economical factors and Social factors, which are in fact strong factors to identify the outcome we are trying to achieve.

Data Exploration

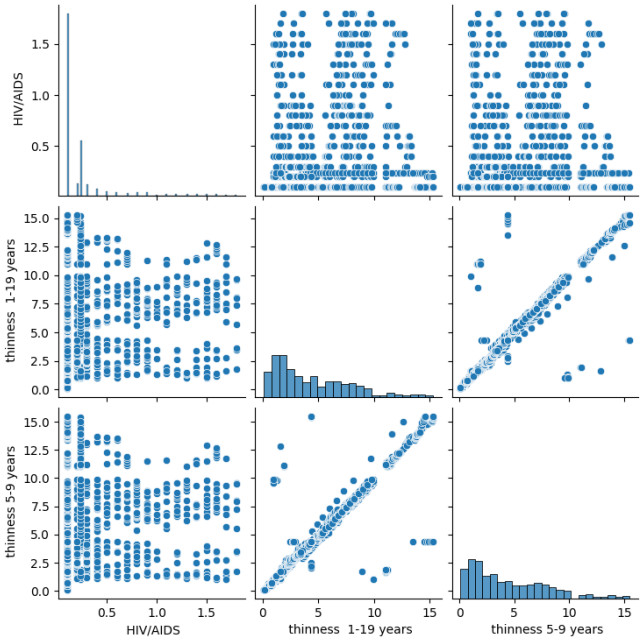

Fig 1. Pair plot for HIV/ AIDS vs. different age groups

- From the above pair plot, we can see that people do get thin with an HIV/ AIDS disease. Moreover Thinness for age group 5-9 have a positive correlation with thinness for age group 1-19. We used Seaborn library to make this plot and have passed columns as the arguments.

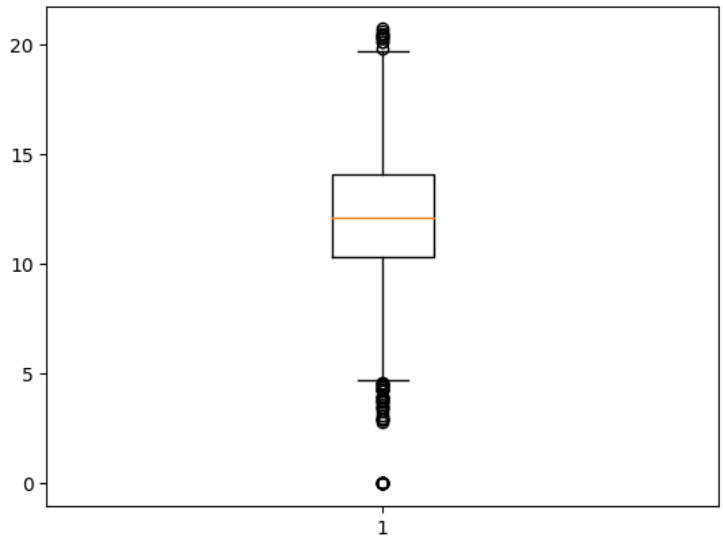

Fig 2. Before Removing the Outliers of a single variable

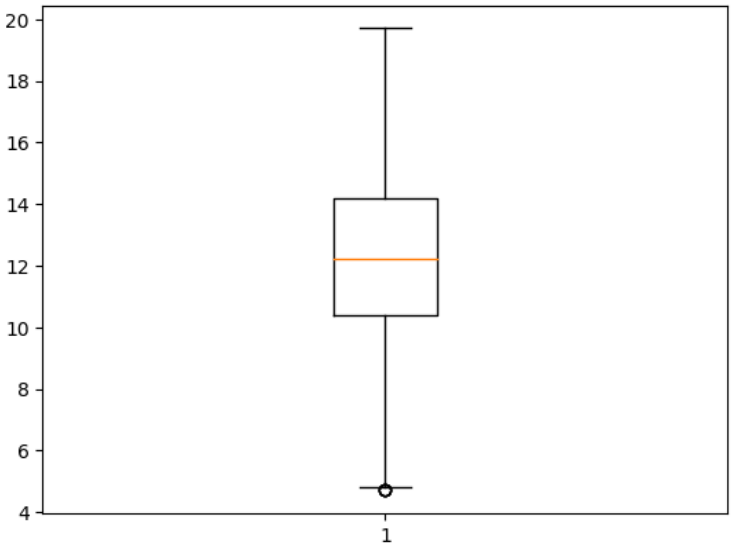

Fig 3. After Removing Outliers of a single variable

- Used Box plots in order to remove out the Outliers from the dataset for all the input variables.

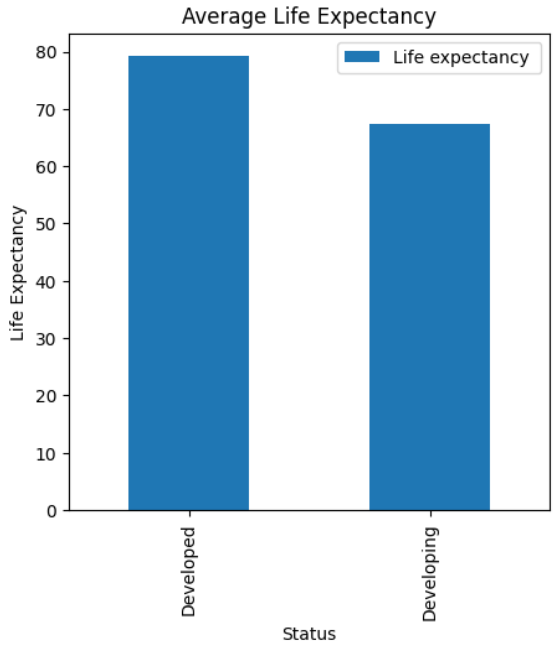

Fig 4. Bar chart for Average Life Expectancy with Developed and Developing Country

- Average Life Expectancy for Countries which are already developed in higher than the countries which are still developing. Used plotly with type as bar, added X label, Y label, fig size for better understanding.

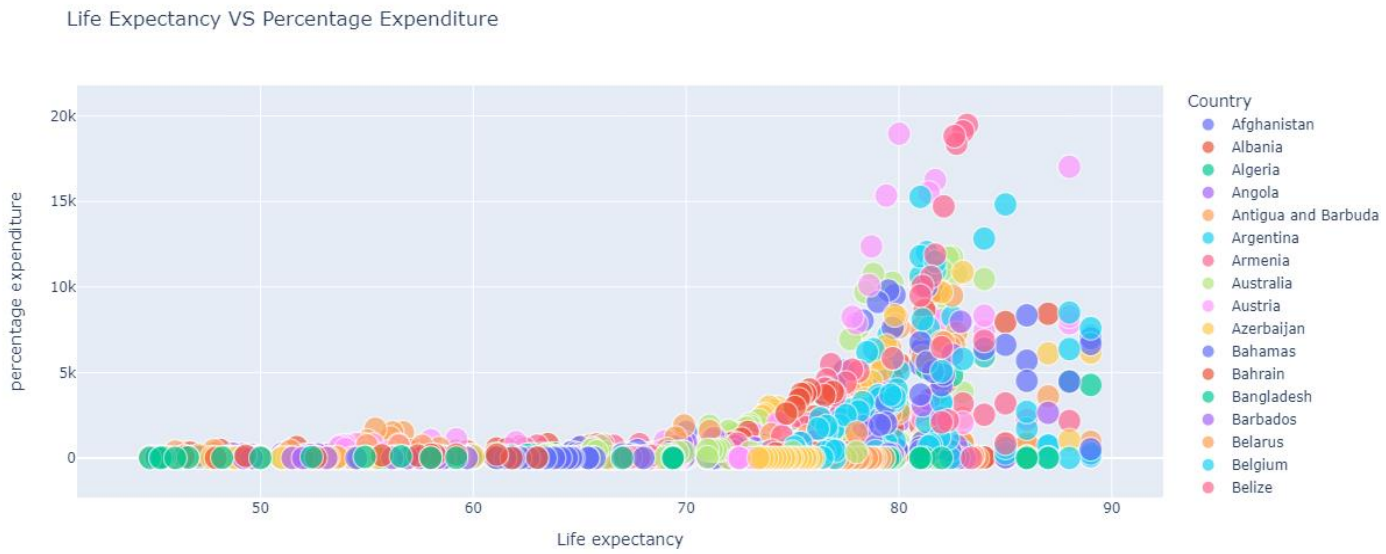

Fig 5. Life Expectancy vs. Percentage Expenditure

- From the above graph, as the age increases, percentage expenditure for Health Sector increases.
- We can also depict that higher Life Expectancy leads to higher expenditure on Healthcare.
- Used Scatter plot with three variables.

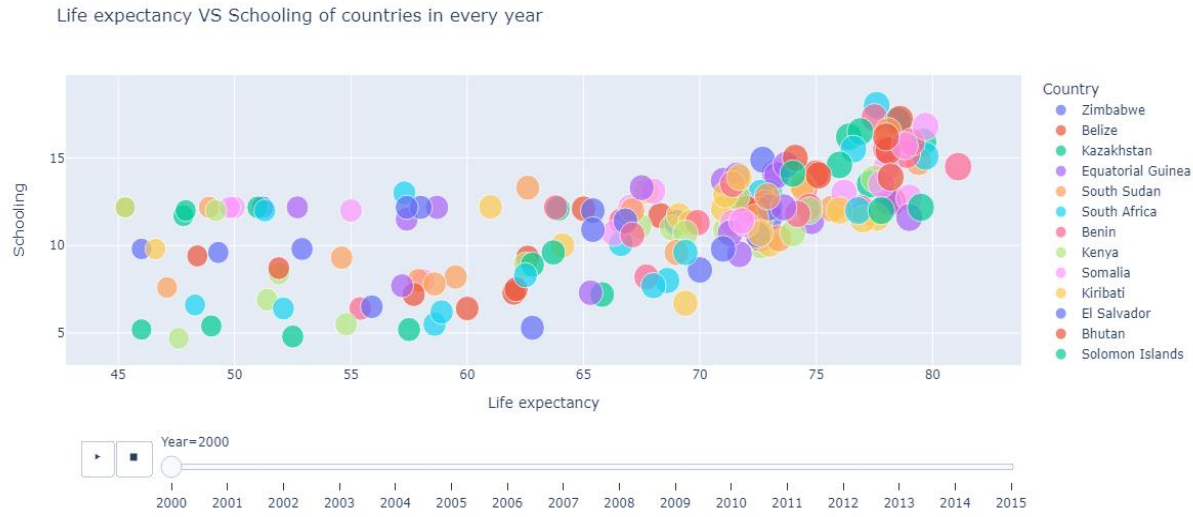

Fig 6. Life Expectancy vs. Schooling Year

- As the years pass on we can see that Life Expectancy for each and every country has increased as the number of years for schooling has increased.
- Used animated plot to figure out the scatter plot based on the various years.

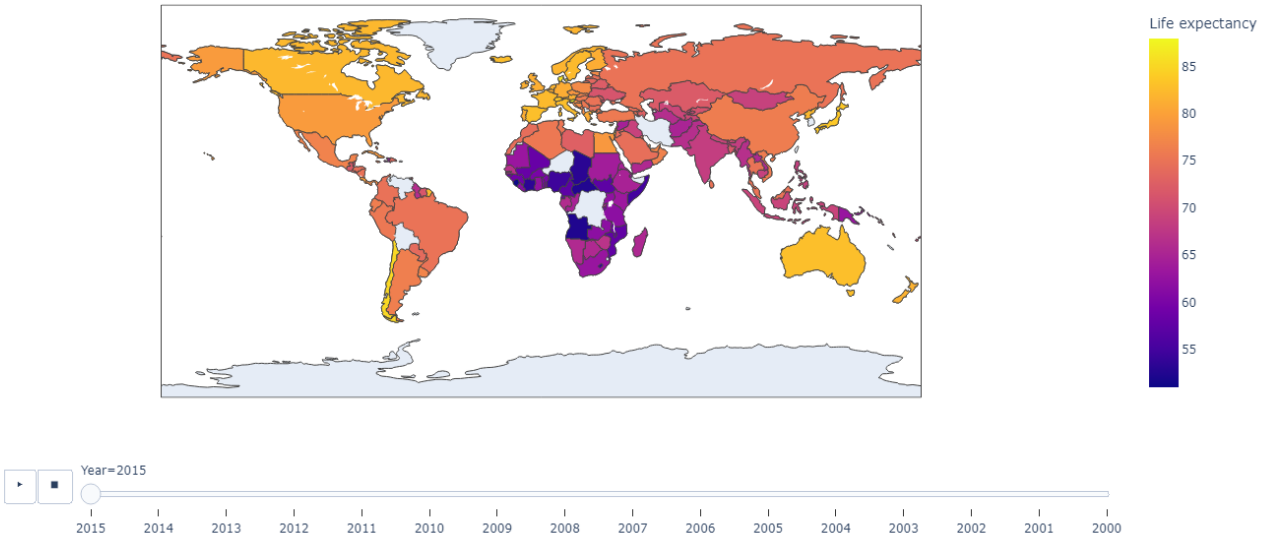

Fig 7. Life Expectancy through various years for Different countries

- Africa, as we can has the lowest Life Expectancy amongst all the other Countries from year 2000 to 2015
- We converted the countries into its area code via pycountry library, as we do not have country code column, which we need in order to plot this map. After that we used plotly library to plot the graph

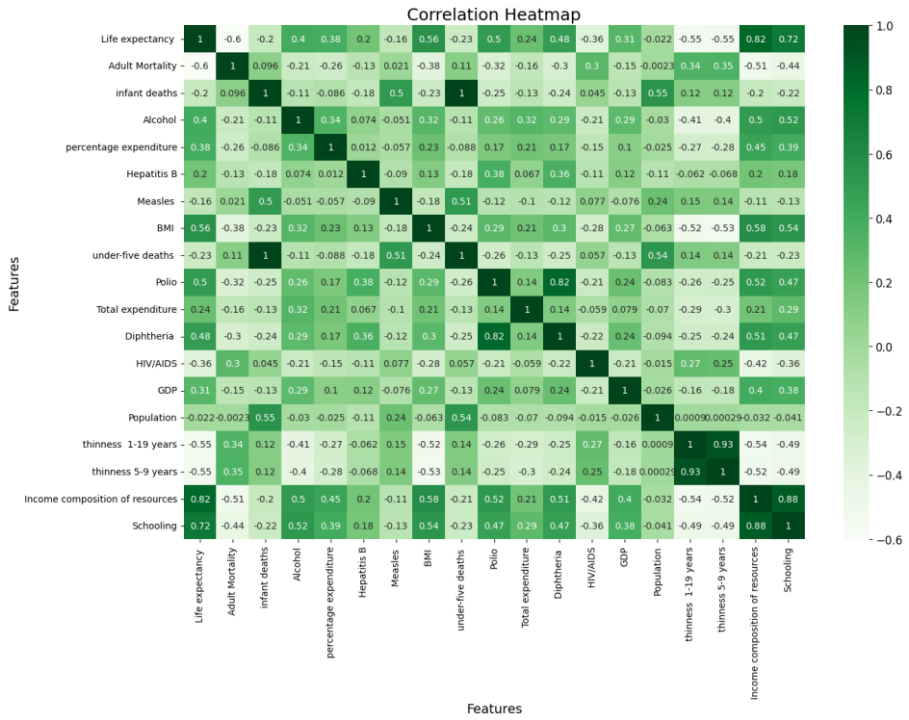

Fig 8. Heat map to find the correlation between variables

Data Mining Tasks

Detecting Outliers: Made Box plots for some variables which were out of range in order to detect the outliers and removed it.

Missing Values: After removing of the outliers, used mean in order to fill the missing values, based on different countries.

Standardization: Used Standard scalar in order to make all the dimensions equal for all the variables and not to make a single variable dominate.

Dimension Reduction: After using Standard Scalar, we applied PCA (Principal Component Analysis) in order to reduce the dimension with conserving the same variance. Found Covariance Matrix with Eigenvalues and Eigenvectors, selected the eigenvectors, which were of use and applied dot product with the original data values in order to get PCA values which will be used in model training.

Data Mining Models

The models used were used purely in its raw form, without hyper tuning or changing of its parameters. We used scikit-learn library in order to fit the dataset into all the models.

Linear Regression: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The method assumes a linear relationship between the dependent variable and the independent variable(s), which means that the relationship can be described by a straight line.

Decision Tree Regression: Decision tree regression is a non-parametric supervised learning method used for predicting the value of a dependent variable based on the values of one or more independent variables. Unlike linear regression, decision tree regression models nonlinear relationships between the dependent and independent variables. The method involves constructing a tree-like model, where each node represents a decision based on a feature or attribute of the independent variables. The tree is constructed recursively by splitting the data into smaller subsets based on the chosen feature and value.

Random Forest Regression: Random forest regression is an ensemble learning method that combines multiple decision trees to make predictions on a dependent variable based on the values of independent variables. It is a supervised learning method used for both classification and regression problems. The method involves constructing multiple decision trees, each of which is trained on a random subset of the data and a random subset of the independent variables. This randomness helps to reduce over fitting and improves the generalization of the model.

Support Vector Regression: Support vector regression (SVR) is a type of supervised learning algorithm used for regression analysis. SVR is a variation of the popular support vector machine (SVM) algorithm, which is primarily used for classification. SVR works by finding the hyper plane in a high-dimensional space that has the maximum margin with respect to the training data points. The hyper plane is determined by identifying a subset of the training data points, known as support vectors, which lie closest to the hyper plane.

Lasso Regression: Lasso regression, also known as L1 regularization, is a linear regression technique used for feature selection and regularization. It is a type of linear regression that adds a penalty term to the ordinary least squares (OLS) regression to prevent over fitting and improve the model's generalization performance. Lasso regression works by adding a penalty term to the OLS regression that is proportional to the absolute value of the coefficients of the independent variables. The penalty term forces the coefficients of some independent variables to become zero, resulting in a sparse model that only includes the most important variables.

Ridge Regression: Ridge regression, also known as L2 regularization, is a linear regression technique used for regularization. It is a type of linear regression that adds a penalty term to the ordinary least squares (OLS) regression to prevent over fitting and improve the model's generalization performance. Ridge regression works by adding a penalty term to the OLS regression that is proportional to the square of the coefficients of the independent variables. The penalty term forces the coefficients of some independent variables to become smaller, resulting in a model that is less sensitive to the noise in the data.

Elastic Net Regression: Elastic net regression is a linear regression technique used for feature selection and regularization. It is a combination of both L1 regularization (lasso regression) and L2 regularization (ridge regression) techniques.

K Nearest Neighbor Regression: K-nearest neighbor regression, or KNN regression, is a non-parametric supervised learning algorithm used for predicting the value of a continuous target variable based on the values of the k-nearest neighboring data points in the feature space.

Gradient Boosting Regression: Gradient boosting regression is an ensemble learning method that combines multiple weak prediction models to create a strong prediction model. It is a type of supervised learning algorithm used for regression analysis.

In gradient boosting regression, a series of decision trees are sequentially added to the model, each one trying to correct the errors of the previous tree. The model is trained by minimizing a loss function, such as mean squared error (MSE), through gradient descent optimization.

ADA Boosting Regression: ADA Boost regression, short for Adaptive Boosting regression, is a type of ensemble learning method used for regression analysis. It is a boosting algorithm that combines multiple weak prediction models to create a strong prediction model.

Extra Tree Regression: Extra Trees Regression, short for Extremely Randomized Trees Regression, is a type of ensemble learning method used for regression analysis. It is a variation of the Random Forest algorithm that builds a large number of decision trees and uses averaging to improve the predictive accuracy and reduce over fitting.

Performance Evaluation

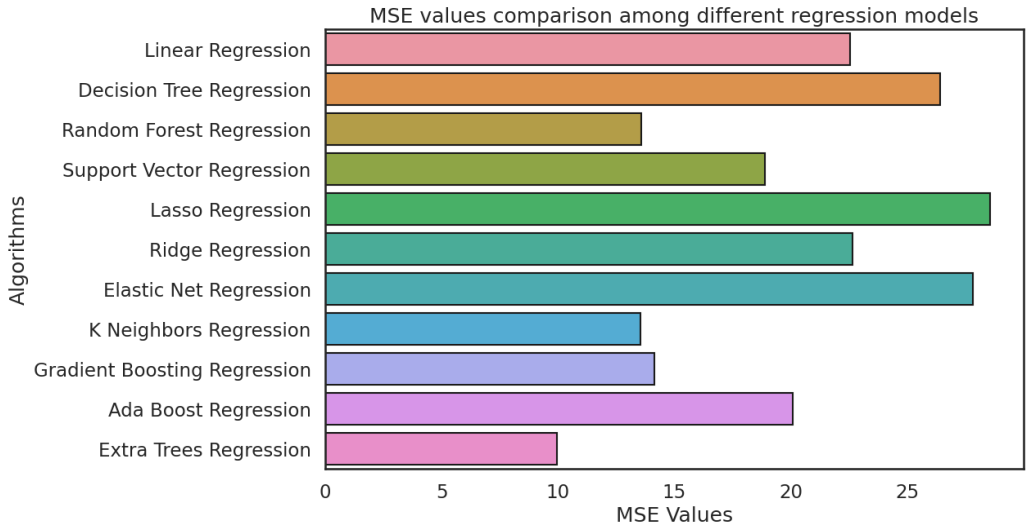

Fig 9. MSE Values for various Regression Models Used

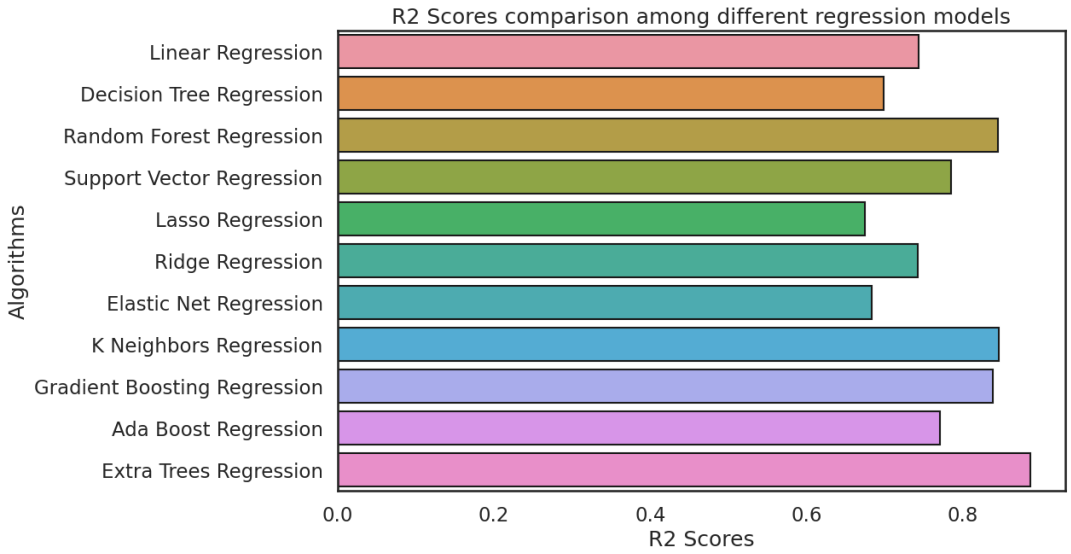

Fig 10. R2 Score for various Regression Models Used

Based on the MSE values and R2 scores, we can say that lower the MSE and R2 Score gets the chance for best classifier, which can be sufficed by Extra Tree Regression, Random Forest Regression and K Neighbors Regression

The following Performance Metrics were used in order to evaluate our Machine Learning models:

Mean Squared Error: Mean squared error (MSE) is a commonly used measure of the average squared difference between the predicted and actual values in a regression analysis. It is a statistical measure that assesses how well a predictive model performs on a given set of data. MSE is calculated as the average of the squared differences between the predicted and actual values for each data point in a dataset. The squared differences are calculated to ensure that the negative and positive differences do not cancel each other out.

The formula for MSE is: $MSE = (1/n) * \Sigma(y - \hat{y})^2$,

Where n is the total number of data points, y is the actual value, $\hat{y}$ is the predicted value, and Σ represents the summation operator.

The MSE is a non-negative value, with a value of zero indicating a perfect fit between the predicted and actual values. A higher value of MSE indicates a greater difference between the predicted and actual values, indicating that the model may not be performing well.

Root Mean Squared Error: Root Mean Squared Error (RMSE) is a commonly used measure of the average deviation between the predicted and actual values in a regression analysis. It is similar to the Mean Squared Error (MSE), but the RMSE takes the square root of the MSE to ensure that the units of measurement of the error metric are the same as the units of the predicted and actual values.

The formula for RMSE is: $RMSE = \text{sqrt}((1/n) * \Sigma(y - \hat{y})^2)$,

Where n is the total number of data points, y is the actual value, $\hat{y}$ is the predicted value, and Σ represents the summation operator. RMSE is a non-negative value, with a value of zero indicating a perfect fit between the predicted and actual values. A lower value of RMSE indicates a better fit between the predicted and actual values, indicating that the model may be performing well.

Absolute Error: Absolute Error is a measure of the magnitude of the difference between the predicted and actual values in a regression analysis. It is calculated as the absolute value of the difference between the predicted and actual values.

The formula for Absolute Error is: $\text{Absolute Error} = |y - \hat{y}|$,

Where y is the actual value and $\hat{y}$ is the predicted value. The Absolute Error is a non-negative value, representing the size of the error without considering the direction of the error (positive or negative). A smaller Absolute Error indicates that the predicted value is closer to the actual value.

Mean Absolute Error: Mean Absolute Error (MAE) is a commonly used measure of the average magnitude of the difference between the predicted and actual values in a regression analysis. It is calculated as the average of the absolute values of the differences between the predicted and actual values for each data point in a dataset.

The formula for MAE is: $MAE = (1/n) * \Sigma|y - \hat{y}|$

Where n is the total number of data points, y is the actual value, $\hat{y}$ is the predicted value, and Σ represents the summation operator. MAE is a non-negative value, representing the average size of the errors without considering the direction of the errors (positive or negative). A smaller MAE indicates that the predicted values are, on average, closer to the actual values.

R2 Score: R2 Score, also known as the coefficient of determination, is a statistical measure used to evaluate the goodness of fit of a regression model. It indicates the proportion of the variance in the dependent variable that is explained by the independent variables in the model. R2 Score is calculated as the ratio of the explained variance to the total variance.

The formula for R2 Score is: $R2 = 1 - (RSS/TSS)$

Where RSS (Residual Sum of Squares) is the sum of the squared residuals (the differences between the predicted and actual values) and TSS (Total Sum of Squares) is the sum of the squared deviations from the mean of the dependent variable. R2 Score can range from 0 to 1, with a value of 1 indicating a perfect fit of the model to the data. A higher value of R2 indicates that a larger proportion of the variance in the dependent variable is explained by the independent variables in the model. R2 Score is a commonly used evaluation metric in regression analysis, particularly in cases where the goal is to predict or explain the variation in a dependent variable based on one or more independent variables.

Based on the above performance metrics, we can say that Extra Tree Regression is performing better than other models. Extra Tree Regression is producing Low MSE, RMSE, Absolute Error, Mean Absolute Error and High R2 Score when compared with the other models.

Project Results

Below given is a table and graph which indicates Extra Tree Regression is the best model to fit this dataset.

index	Algorithms	MSE Values	RMSE Values	Absolute Error	Mean Absolute Error	R2 Values
0	Linear Regression	22.532641820086546	4.746855993190287	140.97443060823596	3.3319312051140204	0.7438917782587633
1	Decision Tree Regression	26.42820246921893	5.1408367479641806	152.67506206925574	3.3921774021482474	0.6996144530121047
2	Random Forest Regression	13.59390265815277	3.6869910032644193	109.49804630444665	2.385709101713319	0.8454903661940159
3	Support Vector Regression	18.898868598242327	4.347282898344933	129.10771512055248	2.7929676367443665	0.7851936018748399
4	Lasso Regression	28.55927189413672	5.344087564228034	158.71130334865435	3.9669215546088434	0.6753925084580364
5	Ridge Regression	22.654066490121725	4.759628818523743	141.3537641673803	3.33890510571018	0.7425116535238774
6	Elastic Net Regression	27.807850756180716	5.273314968421734	156.6094644871484	3.8574400593718043	0.6839332349719216
7	K Neighbors Regression	13.523581474406715	3.677442246236739 5	109.21446268890729	2.320556073608862	0.8462896436805947
8	Gradient Boosting Regression	14.152405057666051	3.761968242511631	111.72475670531334	2.555349910312751 5	0.839142373023945
9	Ada Boost Regression	20.15984856152261	4.489972000082251	133.3453652410272	3.3718807174063263	0.7708611796659557
10	Extra Trees Regression	9.920609903962172	3.149699970467373	93.54131672846302	1.9662207028168603	0.887241372699258

Fig 11. Various Performance metrics comparison between models

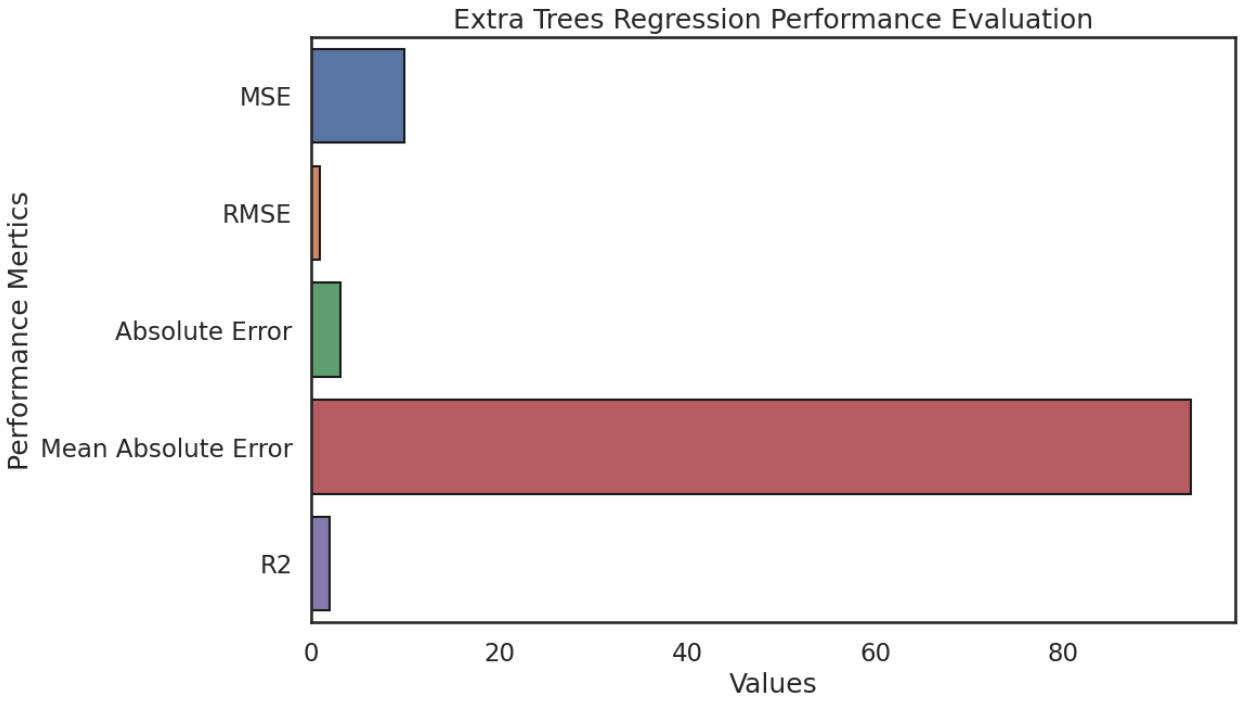

Fig 12. Performance Metrics for Extra Tree Regression

- We can see from the performance results, among multiple regression models used, Extra Tree Regression is giving the best result possible.
- Based on the graphs, charts we can see that Life Expectancy does depend on multiple factors and can impact an outcome drastically.
- While many factors dominate, there are some which does not harm if removed based on Dimensionality Reduction.

Impact of the Project Outcomes

- Analyzing this data could help identify regions or populations where improvements in these areas could lead to increases in life expectancy.
- Furthermore, if a project is implemented in a region or population where life expectancy is relatively low, it may be possible to track changes in life expectancy over time to assess the impact of the project. For example, a project that provides access to clean water and sanitation in an area with high rates of water-borne illness may lead to a decrease in mortality and an increase in life expectancy over time.
- Overall, while the impact of a project on life expectancy will depend on numerous factors, using data sets such as the WHO life expectancy data set can help identify areas where interventions may have the greatest impact and provide a basis for measuring the success of a project.
- With more complexity models, by implementing Deep Learning and many more, we might increase on the accuracy, but can also lead to over fitting.