**Assessing the Impact of Economic, Social, Health, and Geographic Factors on Average Life Expectancy**

**(Group #1)**

**Prepared By:**

Bryce Hawkins, LeeAnn Markwalter,

TJ Augustine

Executive Summary:

Throughout this project, the overall data mining objective was to understand the impact that different health, political, and geographical factors have on average life expectancies across the world. More specifically, we aimed to create a predictive model that could predict the life expectancy of an individual given certain factors. To do this, life expectancy was set as the target variable throughout the entirety of the project.

Our data mining procedures were divided into two trials: trial 1 and trial 2. In trial 1, we utilized partition, filter, transformation, and principal component nodes to run and compare a variety of linear regression and neural network models. The linear regression on transformed significant variables had the highest adjusted r-square at .8719. Overall, the neural network was the most accurate model with calculated average square errors of 10.05 and 6.54. In trial 2, an impute node was used after the data was partitioned to make up for missing values that were present in the original raw dataset. The data was then transformed and regression, neural network, autoneural, and decision tree models were run. Overall, the neural and auto neural network models performed best and had the lowest calculated square error.

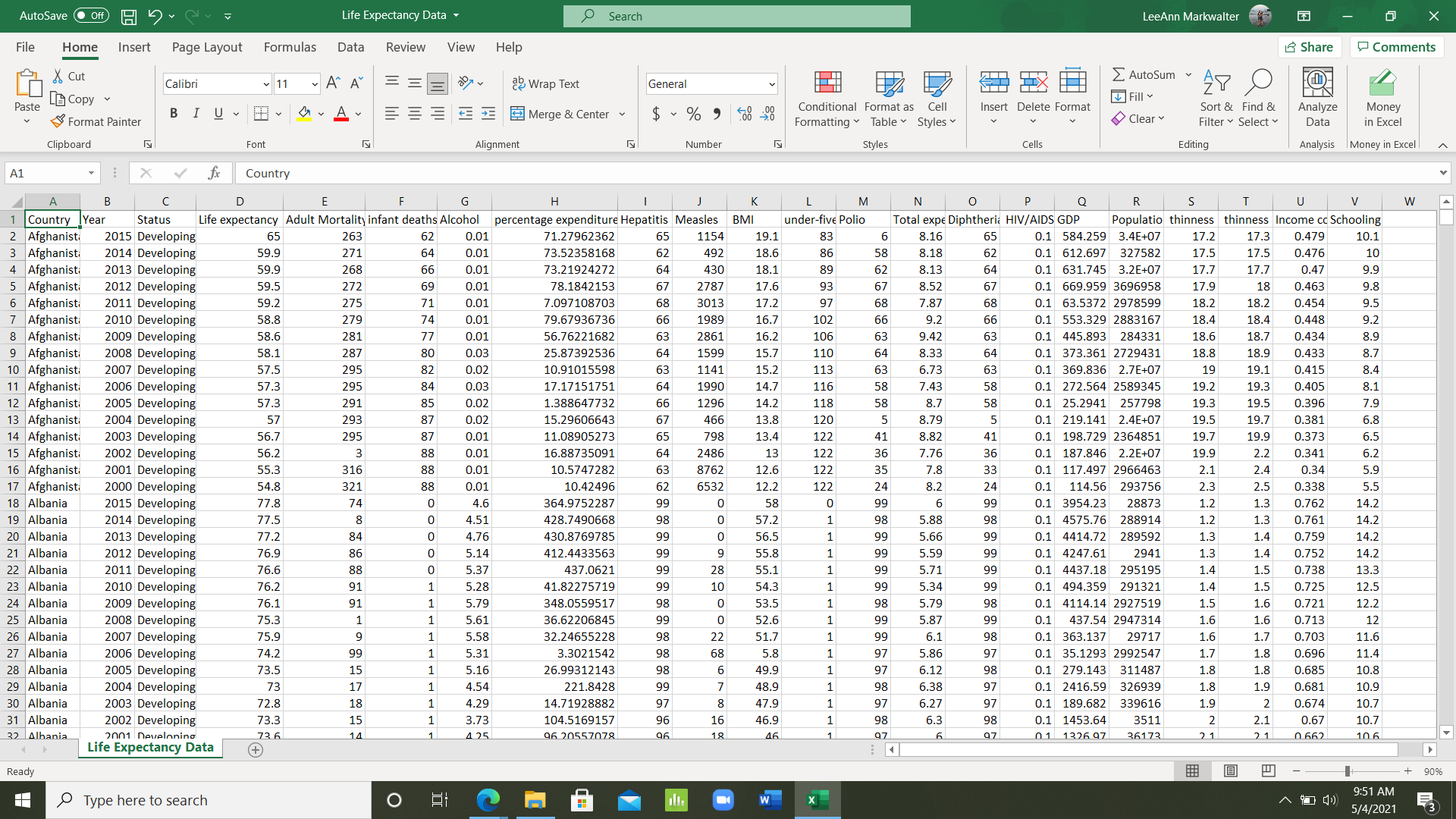
Overall, the predictive models that were run in Trial 1 without imputing any values produced the most accurate models. More specifically, the neural network of Trial 1 produced the lowest square error values. From this, we concluded that the best way to improve our models would be from gathering the actual data values from respective countries, not through imputations. Given these analyses, we found the most significant variables in predicting life expectancy to be adult mortality, diphtheria, income composition, polio, population, total expenditure, BMI, thinness 5-9 Years, year, HIV/AIDS, infant deaths, under 5 deaths, and status. Ultimately, we can include that a variety of health, economic, geographic, and social factors influence average life expectancy for a given country.

Project Motivation/ Background:

Life expectancy is the key metric for assessing overall population health. Understanding the impact that social, economic, health, and geographic factors have on average life expectancy has relevance and importance on a global scale. There is huge variation in average life expectancy from country to country. In some countries the average life expectancy is as low as 50, whereas others have an average of 85. Recognizing relationships and trends in our variables can help us to understand which factors are having the greatest impact on life expectancy. Utilizing these findings, we can help those countries with lower life expectancies to improve their systems, prolong lives of individuals, and address inequities.

Data Description:

The dataset used for the purposes of this project was retrieved from *kaggle.com* and is titled “[Life Expectancy Data](https://www.kaggle.com/kumarajarshi/life-expectancy-who).” The raw data contains 22 columns and 2,938 rows of both qualitative and quantitative variables that measure health and economic factors across 193 different countries around the world between the years of 2000 and 2015. Variables include country, year, economic development status, life expectancy, bmi, population count, gdp, and average years of schooling.

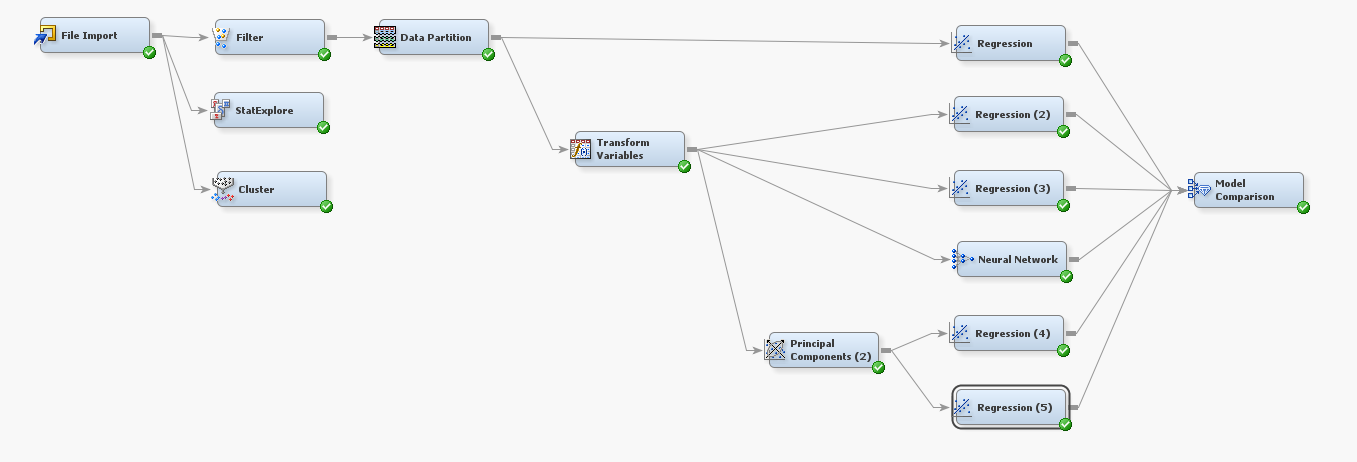
***Figure 1: Snapshot of raw data*** 

Data Preparation Activities:

For initial data preparations before models were built, we conducted simple preliminary data exploration analyses on the raw data set to achieve a basic understanding of our data. In our initial exploration, we noticed several missing values among the population, total expenditure, hepatitis b, alcohol, and gdp variables. Furthermore, in certain rows, values among variables for infant deaths, percentage expenditure, measles, under 5 deaths, and income composition of resources were 0. It is likely that many of these cells are actually missing values and not actually 0. No data values were specifically deleted or replaced in an initial move, but 2 trials were run to work around the missing values. One trial was run without an impute node, and another with a step imputing values to see if the model would improve or change when adjusting for the missing values. For both trials performed, life expectancy was set as the target variable. No variables were rejected initially.

Models/ Enterprise Miner Diagrams Used

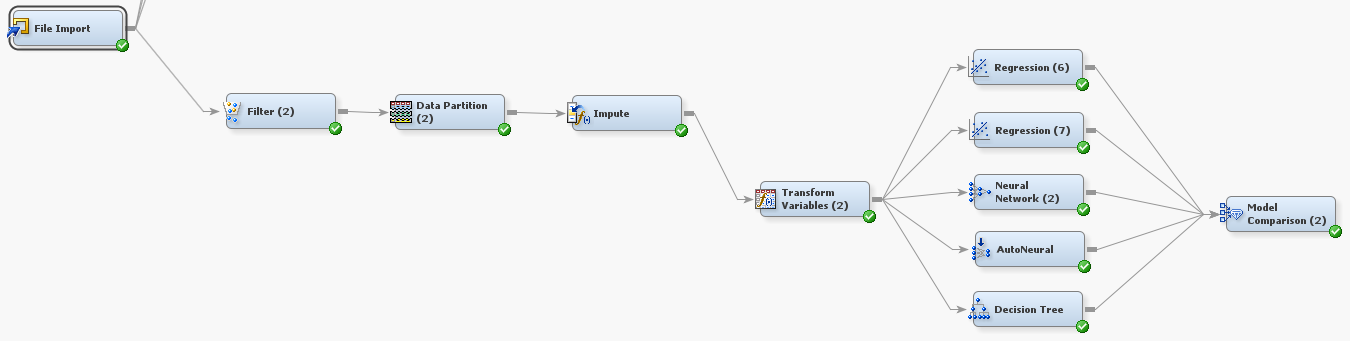
* ***Trial 1:***



***Figure 2: Trial 1 Diagram Overview***

The Trial 1 process began with exploring the dataset statistics and doing a cluster analysis. This revealed the relationships between the variables and also revealed if the variables had any outliers, the mean, and the distribution of the data. Filtering and data partition nodes were used to break up the data to make it more manageable for interpretation and modeling. After an initial regression was run without any changes, the variables were transformed using the best method, which was optimal binning. A second regression was run with the transformed variables to find the significant variables, and then the significant variables were used in another regression and a neural network. Once these were analyzed, the transformed variables were used in a principal component analysis. A regression was used to find the significant principal components, and then a final regression was run to see how accurate the final model was with only the significant components. A model comparison node was run with all of these to see how the average squared error of each compared with each other.

* ***Trial 2:***

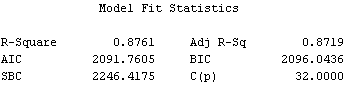


***Figure 3: Trial 2 Diagram Overview***

Trial 2 is very similar to Trial 1, but it includes no PCA and some of the data values were imputed to make up for missing values. The Trial began with filtering and data partition to make the data more manageable and have a training, validation, and test set. The impute node was used and a total of 733 values were imputed in the dataset. The majority of these imputed values were in the variables GDP, Hepatitis B, and Population. Once the impute process was completed, all the imputed variables were transformed to better fit the predictive model. An initial regression was run to find the significant variables, and then a second regression was run using the chosen significant variables. While also using the transformed variables, a neural network, and autoneural, and a decision tree were run. These all linked up to a model comparison to compare the average squared error of the predictive model.

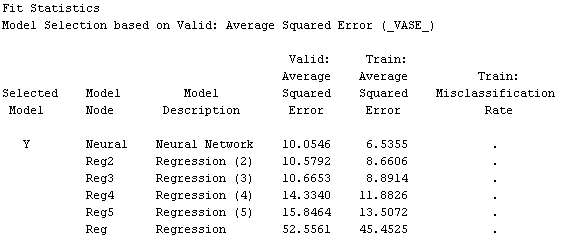
Findings:

* ***Trial 1:***

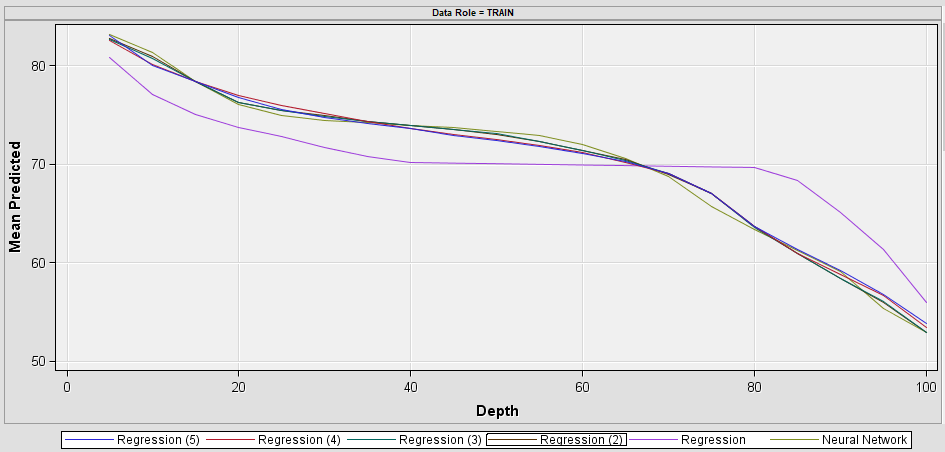


***Figure 4: Adjusted R Square of Regression #3***

The most accurate regression was #3 with an adjusted r-squared of .8719, which utilized the significant transformed variables. This revealed which variables had the greatest impact on life expectancy, and they predicted it generally with solid accuracy. These variables were all transformed using optimal binning, and they included adult mortality, diphtheria, income composition of resources, population, total expenditures, the year, BMI, HIV/AIDS, thinness between 5-9 years, infant deaths, and under 5 deaths. These all had an alpha value less than .05 and make sense when thinking of affecting the lifespan of a person.



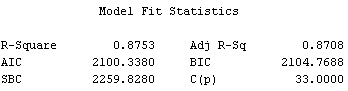
***Figure 5: Model Comparisons of Regressions by Average Square Error***



***Figure 6: Model Comparisons of Regressions utilizing Visualizations***

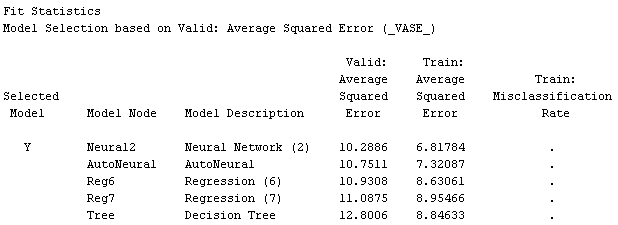
The neural network was the most accurate model overall because it had the smallest average squared error value for both the training data and validation data, with the results being 6.54 and 10.05, respectively. It also produced the lowest error values within the statistics portion of the output analysis. The PCA produced accurate results with an adjusted r-squared value of .8096. While this result is very good, it was not as accurate as the significant regression or the neural network . The least accurate regression was #1 with an adjusted r-squared of .7794, which had no changes to any of the variables.

* ***Trial 2:***

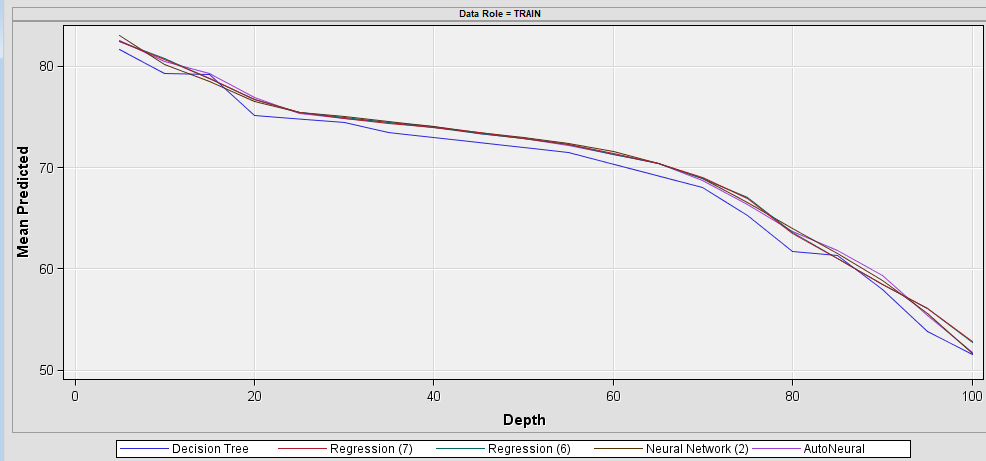


***Figure 7: Adjusted R Square of Regression #7***

The initial regression was run to test the significance of the variables, which found that the variables alcohol, GDP, hepatitis B, schooling, thinness 1-19 years, measles, and percentage expenditure all had p-values above a cutoff level of .05. These variables were not used in the remainder of the predictive models for this trial portion. While the initial regression had a higher adjusted r-squared value than regression #7, the latter did not include the 7 insignificant variables that the prior regression had included. It only dropped .18 to .8708 without the significant variables included, which shows that this model prioritizes the useful values.



***Figure 8: Model Comparisons of Regressions by Average Square Error***



***Figure 9: Model Comparisons of Regressions utilizing Visualizations***

Overall, the neural network and autoneural produced the most accurate models. Based on average squared error, the neural network produced the best values for both the valid and train set, with them being 10.29 and 6.82, respectively. The decision tree model, while not extremely inaccurate, produced the weaksets model of the 5 total models. As said before, regression #7 was more useful because it included only the significant variables. The PCA was not run here because it did not produce the most accurate results in the previous trial. The trial follows suit with the first trial where the neural network also produced the most accurate findings.

* ***Overall Findings***

The predictive models that were run in Trial 1 without imputing any values produced the most accurate models. The neural network in Trial 1 was the most accurate of them all, and the Trial 1 regression using the significant values also fared better than its counterpart. This shows that imputing the values would not help the model, and the only remedy would be to get actual data to improve the predictive power of the models.

The most significant variables were as follows:

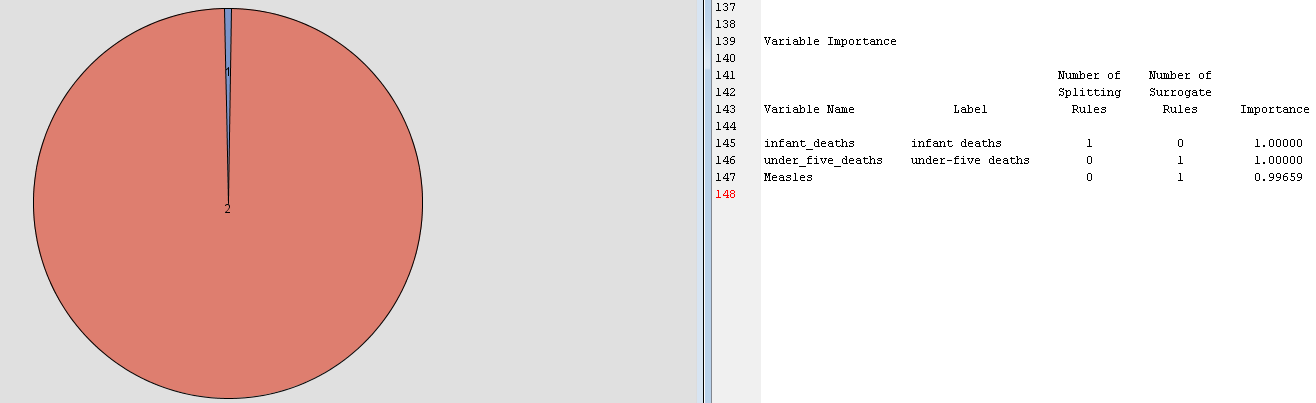
* Adult Mortality
* Diphtheria
* Income Composition
* Polio
* Population
* Total Expenditure
* BMI
* Thinness 5-9 Years
* Year
* HIV/AIDS
* Infant Deaths
* Under 5 Deaths
* Status

This list made sense from both a statistical and logical standpoint. Life expectancy is logically thought of as being longer when a person has little sickness like HIV or Polio and has better health metrics like BMI. Also, it covers all stages of life from Infant/Under 5, Thinness 5-9, and finally Adult Mortality. Together, it shows the overall health and having a healthy infant produced healthy lifestyles. Variables like Expenditure and Population represent the larger economic factors that can impact lifespan.

The insignificant variables were as follows:

* Alcohol
* GDP
* Hepatitis B
* Schooling
* Thinness 1-19 Years
* Measles
* Percentage Expenditure

This portion made sense because while Alcohol, Measles, and Hepatitis B can cause problems, they are not as prominent as they once were. Schooling, while valuable, does not have a high correlation to the health of a person. Similarly, the Thinness 1-19 years is a large period that is already covered in smaller increments by other variables.



***Figure 10: Cluster Analysis***

The original cluster analysis shows that infant deaths and under 5 deaths are very closely related. This is shown because the analysis only produced 2 cluster and eac variables had an importance of 1, while measles had an importance of .996. These variables both ended up being significant, while measles was not, so it shows that they are extremely similar and could have a relationship together when using them in the predictive models.

Managerial Implications & Conclusions:

1. **Managing deadly diseases**

Significant variables such as HIV, Polio, and Diphtheria show that the prevalence of deadly diseases are correlated with the lifespan of a person. Stressing the health of your people and ensuring that they do not get dangerous, life-threatening diseases should be of paramount importance. To mitigate the spread of these diseases and limit case numbers, a country’s healthcare system can take precautions such as expanding vaccination and public health awareness programs.

1. **Stressing healthcare at young age**

Variables such as Infant/Under 5 Deaths and Thinness 5-9 suggest that keeping kids healthy can lead to a long life. Promoting healthy habits at a young age can be the difference maker for how long a person lives. Furthermore, it is essential that parents are given the resources to feed their children and provide them with adequate care.

1. **Promote and test BMI & other health metrics**

Variables like BMI show that there are certain metrics about a person's body that can outline how healthy they are. Keeping a solid body fat index ensures that people have a lesser chance of developing heart disease or diabetes, which could shorten a person’s life.

1. **Spend more on healthcare each year**

Variables like Population and Total Expenditure show that there are outside factors beyond each individual person’s choices that impact how long they will live. Governments should be spending an adequate amount of money and resources on healthcare to improve life expectancy.

1. **Continue to manage dwindling diseases**

Although Hepatitis B and Measles are not as impactful as they once were, continuing to provide people with vaccinations and booster shots are important to prevent them from becoming dangerous. Even something like Alcohol is not as prevalent in society as it once was, with new laws and restrictions all the time. While they are insignificant to the overall life expectancy of a person, it is still important to manage these so that they do not become significant.

1. **Track necessary variables**

A variable like Thinness 1-19 years is overshadowed by the Thinness 5-9 and is a very broad variable. This only proves that a model can only be as good as the inputs it is utilizing. Breaking it down in 5 year subsections seems to have proven useful for the one variable, so finding the right methodology for what to collect and how to collect can provide greater insights.

1. **Understand the use of the variables**

Variables like Schooling and GDP are extremely important, just not for this particular predictive model. Just because they are not significant for life expectancy, does not mean that they do not have other societal benefits like utility. They both can improve society as a whole and can improve the creation of new technologies and markets that can possibly lead to world changing discoveries.

To conclude, there is no one factor that decides how long an individual lives. A variety of health, social, economic, political, and geographic factors all play a part in determining the average life expectancy of a given country. To ensure a high average life expectancy, both the systems that a country establishes and their citizens must work cohesively. Finally, it is important that we recognize and address inequities among countries that may be created due to lack of healthcare resources, affordability, or awareness.

References:

[How different factors have an influence on your life expectancy | by Mubarak Ganiyu | Towards Data Science](https://towardsdatascience.com/how-different-factors-have-an-influence-on-your-life-expectancy-7b807b04f33e#:~:text=The%20difference%20between%20developing%20countries%20and%20developed%20countries,have%20an%20average%20life%20expectancy%20of%2080.7%20years.)

[Life Expectancy - Our World in Data](https://ourworldindata.org/life-expectancy)

[WHO | Fact file on health inequities](https://www.who.int/sdhconference/background/news/facts/en/#:~:text=1.%20Today%2C%20there%20is%20a%2036-year%20gap%20in,under-five%20mortality%20rate%20is%2013%20out%20of%201000.)

<https://www.kaggle.com/kumarajarshi/life-expectancy-who>