Intersect, Vol 18, No. 1 (2024)

Change in Life Expectancy Across Countries

Sutej Reddy Mandadi
Redmond High School

Abstract

This paper used machine learning (ML) techniques to examine which factors contribute the greatest to life expectancy levels. Firstly, through background research, life expectancy was found to be an effective representation of a country's overall health. Next, initial data analysis was done to analyze which features of the data were relevant to this study by looking at the factors affecting life expectancy. After the features were selected, three ML models were fitted to the data: multiple linear regression, random forest regression, and decision tree regression. The ML models were instrumental in identifying how these features interact with each other and life expectancy. The random forest regression model returned the highest R-squared value so that is the model used for this study. The R-squared value communicates how accurately the model makes predictions compared to the actual test data. To decide which of the features affected life expectancy greatest, feature importance was used. Feature importance is a metric that shows how greatly features are affecting the output value in an ML model. After running feature importance on the random forest regression model, the graph showed that the gross domestic product (GDP) of the country most greatly affected life expectancy. GDP encompasses the value of total final output of goods and services produced by the economy of a nation in a year. This conveys the importance of economic involvement to a country's overall health. When a resource-constrained country does better economically and improves its GDP, it increases output of goods and services resulting in job creation and more money in the nation. The additional financial resources will provide an opportunity for resource constrained nations to spend more money on institutions like health care and education which in turn impact life expectancy positively.

1. Introduction

The Office for National Statistics [2] defines life expectancy as, "a statistical measure of the average time someone is expected to live, based on the year of their birth, current age and other demographic factors including their sex" (ONS, 2023, 1. Life expectancy). Fundamentally, the life expectancy of a country is a great indicator of the overall health of the people in the country. [3] The measure of life expectancy is especially important for underdeveloped nations because they tend to have lower life expectancies than that of the wealthier countries. It is in the best interest of underdeveloped nations to determine what can be improved to better the life expectancy of their country. This study aims to discover what factors generally are more impactful on the life expectancy measure of a country using data from the World Health Organization. By using this study to narrow down what factor(s) underdeveloped nations should use to better the overall health in their countries, resources and funds of these nations are saved because unlike the wealthier nations, these countries cannot spend billions of dollars trying to fix a multitude of problems at once. While there are many issues in the world, this study will focus on socio-economic factors that are viable for improvement. One factor that may be relevant to this study is schooling. [3] "Education showed a dose–response relationship with all-cause adult mortality, with an average reduction in mortality risk of 1.9% (95% uncertainty interval 1.8–2.0) per additional year of education" (Balaj et al., 2024, Findings). [4] Another factor that may be useful is the gross domestic product (GDP) of a country. GDP encompasses the value of total final output of goods and services produced by the economy of a nation in a year. "GDP per capita increases the life expectancy at birth through increasing economic growth and development in a country and thus leads to the prolongation of longevity" (Miladinov, 2020, Data and methods). The occurrence of other diseases and these two metrics are available to be used in the model for our research study. After a sound model is created, the different features will be altered slightly to test which of the features affects the output, life expectancy, the greatest.

2. Literature Review

Life expectancy is used as a key indicator to measure the health and relative well-being of a nation's population. It is not surprising that there has been significant research in this area to study different factors that impact life expectancy. The previous research studies have covered the

impact of different factors like health, education, GDP etc. on life expectancy.

2.1. Interconnectedness of Overall Health and Life Expectancy

The usefulness of this research paper fundamentally relies on the assumption that life expectancy correlates directly to the overall health of a nation. It is easy to see that the health of a country's population has a clear correlation with life expectancy. [5] Yanping Li et al. (2018) aimed to discover why Americans had a shorter life expectancy compared the populations of other wealthy countries. They concluded that "Adopting a healthy lifestyle could substantially reduce premature mortality and prolong life expectancy in US adults" (Li et al., 2018, Conclusions). Because life expectancy can be traced back to healthy lifestyles, we can hypothesize that unhealthy patterns like alcohol abuse in countries may lead to lower life expectancies. This study also helps validate our research procedure because it shows that life expectancy does not solely depend on whether a country is developed or undeveloped. As demonstrated by the average life expectancy of the US population, there are likely other factors at play. [6] In another related but more specific study, Nagai et al. (2011) studied the impact of walking on life expectancy. Their results showed that, "Participants who walked >=1 h per day have a longer life expectancy from 40 years of age than participants who walked <1 h per day" (Nagai et al., 2011, Results). Walking is an aspect of a healthy lifestyle, so this study continues to demonstrate the evident link between a healthy lifestyle and life expectancy. As displayed by this research study, there are specific factors that directly influence life expectancy. Our research study will utilize this knowledge and take it a step further by determining which factor is most significant.

2.2. Using Machine Learning for Life Expectancy Prediction

[7] Lakshmanarao et al. (2022) researched the feasibility of life expectancy prediction based on immunization and Human Development Index (HDI) factors. For context, this study utilizes the same data set that our life expectancy project used. They used different machine learning (ML) models like "logistic regression, SVM, Decision Tree, and random forest regression and achieved a good R^2 value with the random forest algorithm" (Lakshmanarao et al., 2022, Abstract). Support vector machine (SVM) is a supervised learning algorithm that tries to find the best line, or decision boundary. This project by Lakshmanarao et al.

provides insight into some of the potential models that we could use to carry out this project.

2.3. Impact of Countries' Economic Factors on Life Expectancy

The socio-economic status of a country also has a fairly significant impact on life expectancy, and this has been studied in previous research papers. [4] Goran Miladinov (2020) studied the relationship between socioeconomic development and life expectancy. They concluded that income per capita and infant mortality play a large role in determining life expectancy. This paper used the Full Information Maximum Likelihood (FIML) method and statistical model to estimate life expectancy based on other input variables. This paper did not use machine learning, which is one difference in our project. The use of machine learning allows me to evaluate a larger range of factors than that seen in the paper by Miladinov. [8] Hummer and Hernandez studied the impact of educational attainment on life expectancy. Their study showed that the "Highly educated adults in the United States have lower yearly mortality rates than less-educated people in every age, gender, and racial/ethnic subgroup of the population" (Hummer and Hernandez, 2013, A HEAD: Educational Differences in Adult Mortality). This quote does not give insight as to the link between education and mortality rates in all countries. We can build off this study with our own plan to analyze schooling rates as one of our features in our machine learning model.

2.4. Learning from Past Studies

Previous researchers have tried to measure the impact of different social, health and economic factors on life expectancy. Our research will look at all of these factors together to see which factors have the greatest impact and also make a recommendation to developing nations as to where they should invest in order to get the most benefits. Our project differentiates itself from other socio-economic studies because our study utilizes machine learning to effectively make predictions. While some studies above include machine learning models, this study will take it a step further by determining the greatest factor that changes life expectancy.

3. Data Evaluation

3.1. Data Description

The data set was downloaded off Kaggle. It is based on data published by The World Health Organization (WHO) and United Nations. The data set

includes life expectancy and different health factors that describe life expectancy for 193 countries that is published by WHO. The corresponding economic data was collected from United Nations web site. The data was collected for each country from years 2000 to 2015. In the end, we had 2938 rows representing the data for 193 countries across these years for 22 different fields. The data was recorded for the following fields: Country, Year, Status (Developed or Developing), Life expectancy, Adult mortality, Infant deaths, Alcohol, Percentage expenditure, Hepatitis B, Measles, BMI, Under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, Thinness 1-19 years, Thinness 5-9 years, Income composition of resources, and schooling.

To start, we completed exploratory data analysis to determine which factors captured in this data set our research should focus on. We excluded some factors like Year, Status because the research was not trying to focus on the year nor the status of the nation. As part of the exploratory data analysis, we plotted life expectancy against the factors in the data set to analyze the impact of each factor on life expectancy.

3.2. Data Distribution

To get a high-level understanding of the data distribution for each of the features, we created a histogram. Here is how the distribution looked:

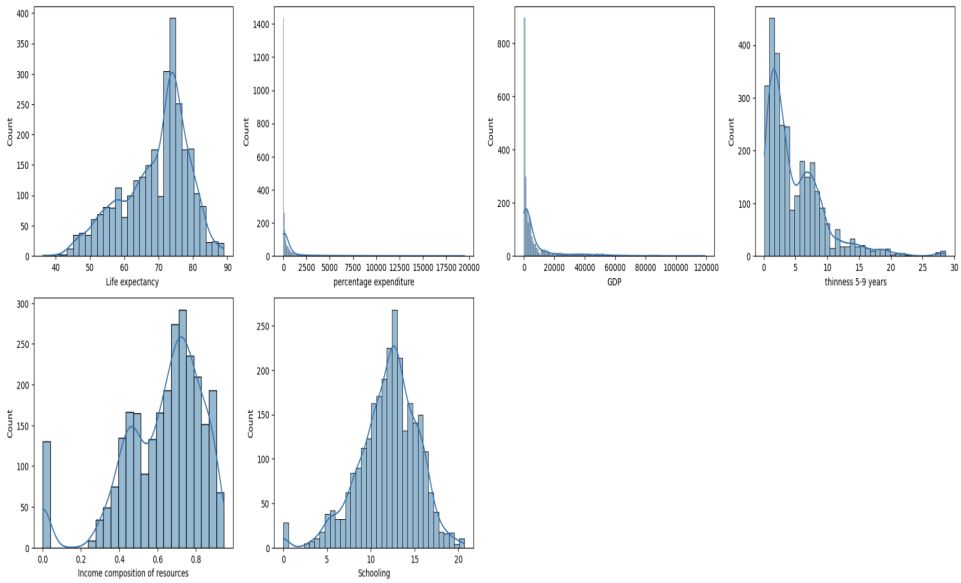

FIGURE 1. Distribution of Features

3.3. Features that showed clear correlation with life expectancy

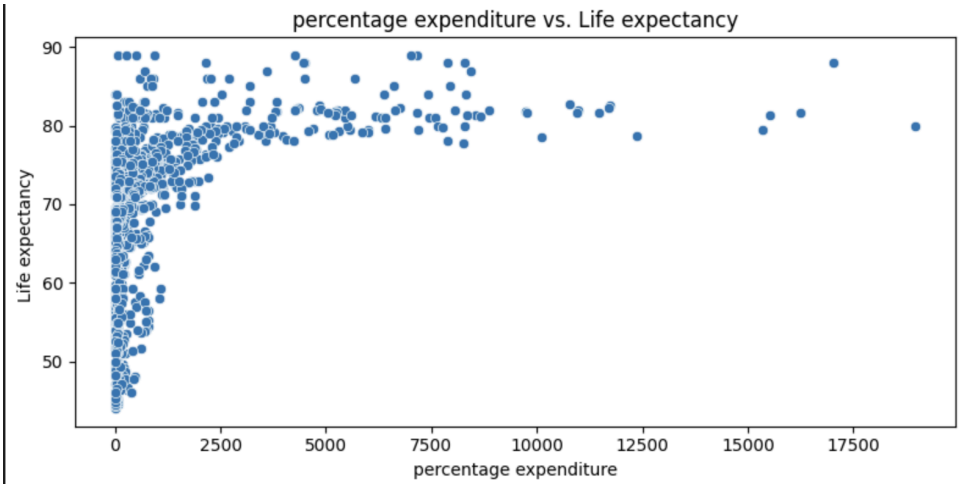

FIGURE 2. Distribution of Percentage Expenditure and Life Expectancy

Figure 2 shows life expectancy plotted against health expenditure (as a percentage of GDP). There seems to be a clear impact of health expenditure on life expectancy with positive impact on life expectancy with an increase in health expenditure by a nation.

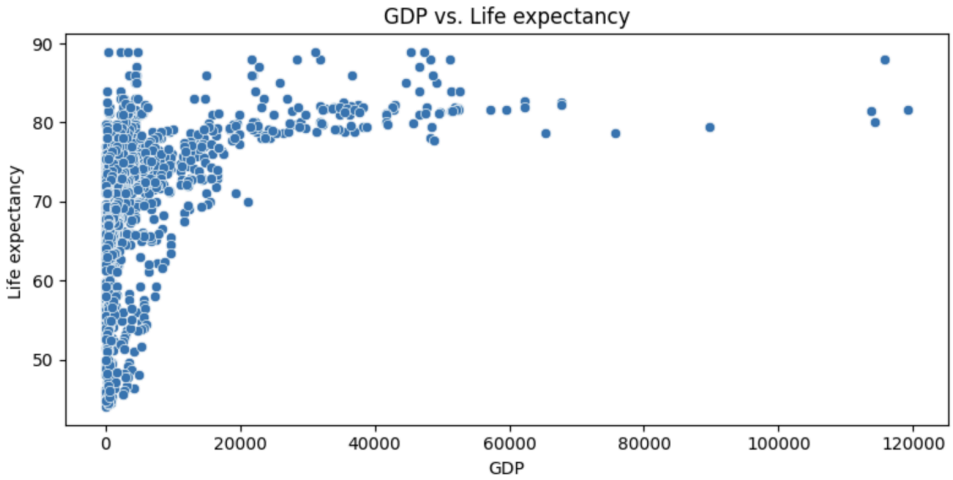

FIGURE 3. Distribution of GDP and Life Expectancy

Figure 3 shows life expectancy plotted against per-capita GDP. There seems to be a clear impact of GDP on life expectancy with positive impact on life expectancy with an increase in GDP of a nation.

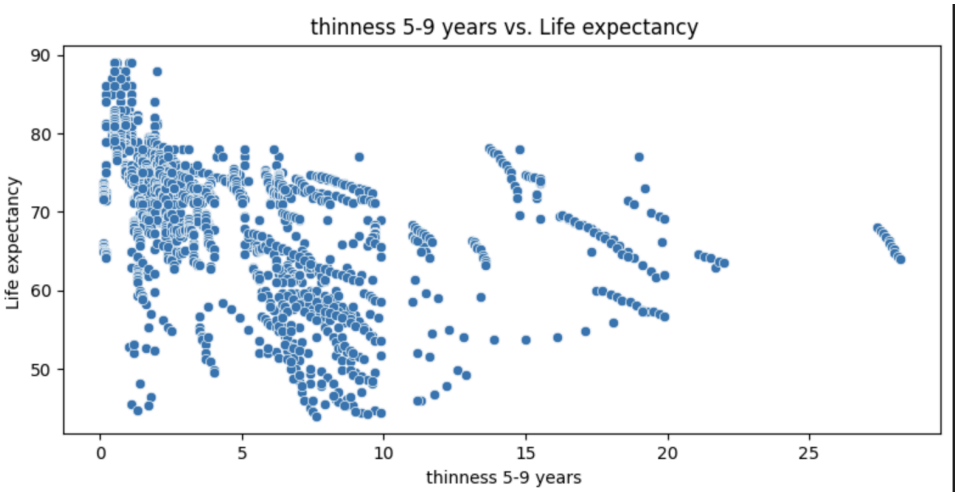

FIGURE 4. Distribution of thinness 5-9 and life expectancy

Figure 4 shows life expectancy plotted against prevalence of thinness in 5-9 year old children. There seems to be a slight correlation between prevalence of thinness of children and life expectancy with negative impact on life expectancy with an increase in prevalence of thinness of children.

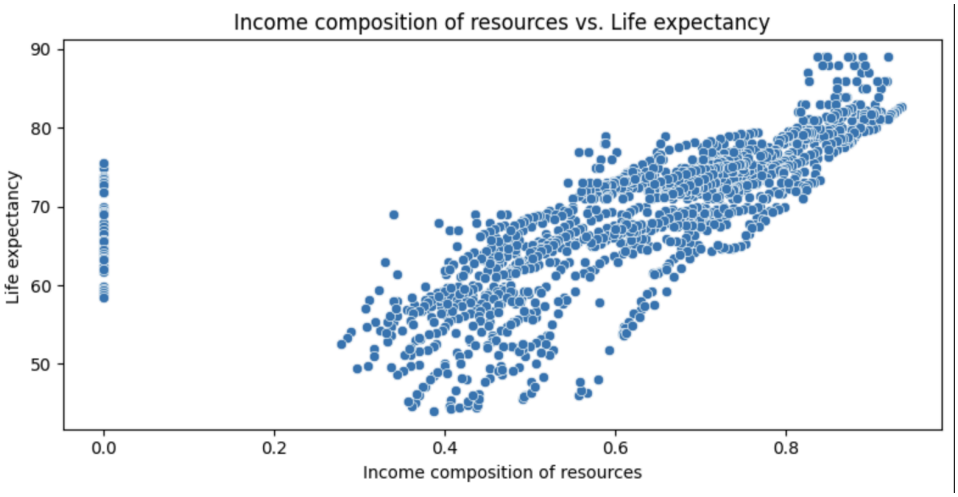

FIGURE 5. Distribution of Income composition of resources and Life expectancy

Figure 5 shows life expectancy plotted against income composition of resources. There seems to be a clear impact of income composition of resources on life expectancy with positive impact on life expectancy with an increase in income composition of resources.

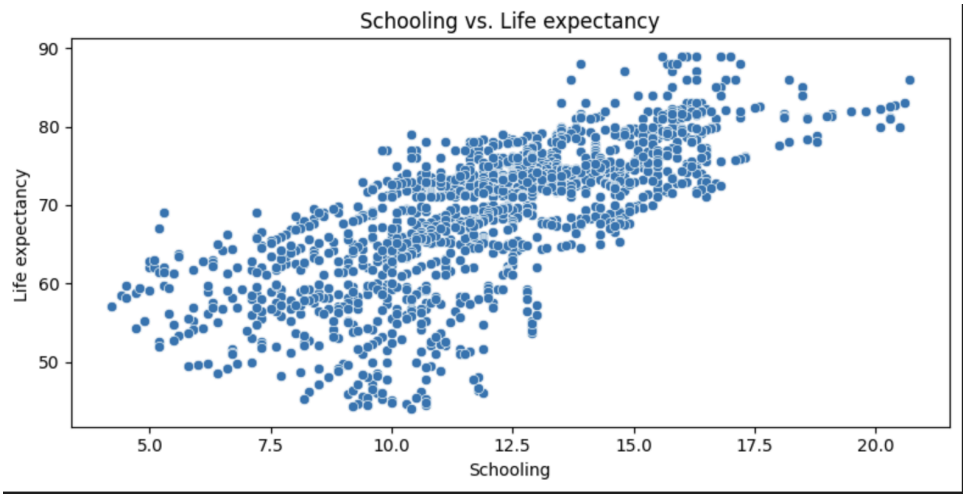

FIGURE 6. Distribution of Schooling and Life Expectancy

Figure 6 shows life expectancy plotted against number of years of schooling. There seems to be a clear impact of number of schooling years on life expectancy with positive impact on life expectancy with an increase in the number of schooling years.

3.4. Features that don't have clear correlation with life expectancy

In our data exploration, we found the following features did not demonstrate a strong correlation with life expectancy: per capita alcohol consumption, Hepatitis immunization, Polio immunization.

3.5. Factors that the Research will Focus on

Based on the exploratory data analysis, we could see that some factors had significant impact on life expectancy while other factors had smaller or negligible impact. Our focus in this study was narrowed down to the five factors that seem to have the most impact on life expectancy. The five factors we identified for the research to focus on based on our exploratory data analysis are:

I. Per-capita GDP
II. Health expenditure (as a percentage of GDP)
III. Number of schooling years
IV. Prevalence of thinness in 5-9 year old children
V. Income composition of resources

3.6. Data Preprocessing

3.6.1 Removing Insignificant Columns from the Dataset

We ran the pre-processing step to remove all the insignificant columns to arrive at the dataset with the 6 significant columns, as stated above. This included the life expectancy column. Firstly, columns like "country" and "year" were dropped because they are insignificant for this study. We do not want the model to operate based on these values. A few other columns were dropped because they are close proxies for life expectancy. For example, the columns of "infant deaths" and "deaths under 5" directly correlate to life expectancy and they are obvious causes of decreased life expectancy. There is not much value in including these values in the study because if they are selected as the most important feature, we are not getting much value out of this study.

3.6.2. Removing non-numerical data from dataset

Once we had the required columns in the dataset, we worked on ensuring that the data for each of the columns was valid numerical data. During pre-processing, we found that there were some rows where some of the values were non-numerical data.

To start with, there were 2938 entries in our dataset. After removing rows with invalid data entries, we are left with a clean dataset of 6 columns and 2458 rows. We did a manual inspection of about 50 rows (~20%) of deleted data to ensure that removing this data would not introduce any bias into the subsequent training and prediction.

4. Methodology

4.1. Splitting Training and Test Data

The training and testing data will be split randomly, using each row entry of the data as one data point. The reason that random division works in this specific scenario is because we do not care about the tag on the data. For example, we do not care if the data is of the country Afghanistan or New Zealand. For training these models, we want the model to treat the data all the same. We will be using a standard percentage, 20% toward testing/validation and 80% toward training data.

4.2. ML Regression Analysis Models

We applied different ML algorithms on the dataset to first determine the one that is most successful at determining life expectancy based on these input parameters. The algorithms we considered are Multiple Linear Regression, decision tree and random forest.

4.2.1. Multiple Linear Regression (MLR)

[10] In simple linear regression, a model is built to predict the outcome of one dependent variable based on one independent variable. Multiple linear regression expands this idea, and the model is extended to predict the value of one dependent variable based on two or more independent variables. There are a few assumptions made by the multiple linear regression. For example, it assumes that the independent variables are not highly correlated and the observed values for independent variable are selected independently and randomly. In our research, the response variable is life expectancy and explanatory variables are per-capita GPD, health expenditure etc.

4.2.2. Decision Tree Regression

[11] Decision Trees (DTs) are a non-parametric supervised learning method. They are used for classification and regression. Decision tree regression creates a model that predicts the value of a target variable by learning simple decision rules. It operates by recursively partitioning the data into subsets based on input values. This partitioning creates a tree-like structure. The internal nodes of the tree represent a decision based on specific feature whereas lead nodes contain the final predicted outcomes. Decision trees are simple to understand, and simple to interpret. The cost of using a decision tree is logarithmic and performs well in most conditions.

4.2.3. Random Forest Regression

[12] Random forest regression analysis uses multiple decision trees to make a prediction. It combines the predictions from these decision trees to arrive at the final prediction. It is based on ensemble learning method. Ensemble learning is the process of using multiple models on the same data and averaging the results of each model to find a more predictive result. This approach reduces overfitting and improves accuracy. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of random forest regression classifiers depends on the accuracy of the trees that make up the forest and the correlation between them.

4.2.3.1 n-estimators in Random Forest Regression

[13] According to sci-kit learn, “A random forest is a meta estimator that fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control

over-fitting" (sci-kit learn, n.d., RandomForestRegressor). For this model we will be creating, we will need to find the optimal value for the n-estimators parameter in implementation. We will discuss the strategy to do this later in the results portion of this paper.

4.3. R^2 Value

In this paper, the models will be evaluated using the R^2 value. This R^2 value shows how well the model fits the data provided. Provided below is the formula to calculate the R^2 value. [14]

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SRR)}{total\ sum\ of\ squares\ (SST)} = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}$$

The method required to calculate this is found in the sci-kit learn library and we will pull it from there. The r^2 value will be outputted as a value between 0.0. and 1.0. A value closer to 1.0 shows that the value is effective at modelling the data and a value closer to 0.0 shows little fitting.

4.4. Feature Importance

After the model is chosen based on which R^2 value is lowest, we will use the model to test which feature is affecting the life expectancy greatest. It is provided in sci-kit learn how to implement this as there is a measure called mean decrease in impurity. The bounds for this are the same as for the R^2 value for the models. From 0.0 to 1.0, the higher the value from the mean decrease in impurity, the higher effect the feature is having on the y-factor, life expectancy.

5. Results

5.1 R^2 Values

	Multiple Linear Regression	Decision Tree Regression	Random Forest Regression
Trial 1	0.3463009644606968	0.7542407273441734	0.8706403006370265
Trial 2	0.3463009644606968	0.7461722186830037	0.8706726840605479

Trial 3	0.3463009644606968	0.7592901770785977	0.869775030874452
Average	0.3463009644606968	0.75323437436	0.8700742486

FIGURE 7. Table showing R^2 Values for Different Models

From, the sci-kit learn library, linear regression, decision tree regression and random forest regression were implemented. The R^2 value that was outputted for multiple linear regression had an average of 0.3463. This is a very low accuracy level, but linear regression is a simple model so hopefully, the future models do a better job at fitting this data set. The R^2 value for decision tree regression was an average of 0.7532 for the three trials. This is more than the first model, and this is supported by the knowledge that decision tree regression is more complex than multiple linear regression. The last model used was random forest regression which outputted an average R^2 value of 0.8700742486. Because random forest regression was the best fit, we will use this model going forward.

5.2. Concern with Overfitting

Overfitting, meaning the model performs very well on training data, but not so well on unseen data is a common issue when testing machine learning models. A simple way of describing overfitting is that the model memorizes the training data and cannot generalize itself for unseen data. A way to test a model for overfitting is comparing the R^2 values for the model results on training data and testing data. If there is a large difference, it means that the model is overfitted because it is performing overwhelmingly better on seen data than unseen data. As seen in Figure 7, the R^2 value for random forest regression was 0.870 for the testing date. After applying the same model to the training data, the R^2 value was 0.983. Since both these values are close and the test data R^2 value is very high, we do not need to be concerned with overfitting.

5.3. Feature Importance

The first method used for feature importance is mean decrease in impurity (MDI).

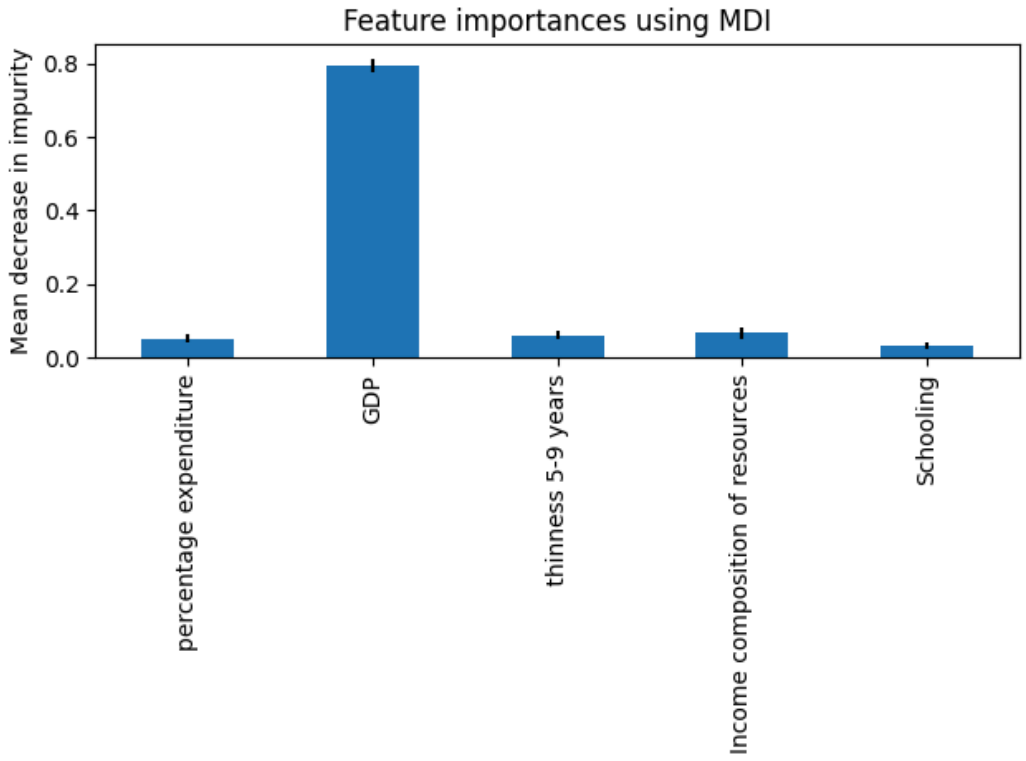

FIGURE 8. Graph with Feature Importances using Mean Decrease in Impurity

As seen in *Figure 8*, GDP is the largest factor affecting life expectancy. The second method used permutation importance.

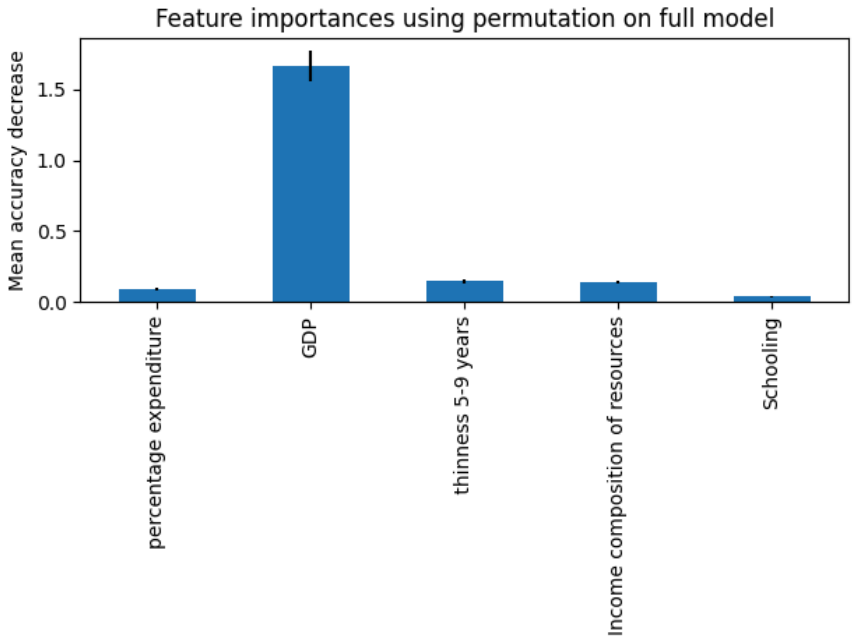

FIGURE 9. Graph with Feature Importances using Permutation Importance

As seen by *Figure 9* as well, the GDP has the greatest feature importance. In both these graphs, schooling had the least effect on the final life expectancy.

5.6. Relative Uncertainty of Feature Importance Values

Relative uncertainty using margin of error is important when analyzing feature importance. Low margins of error show confidence for the feature importance value. Here is a plot of the relative uncertainty values for random forest regression:

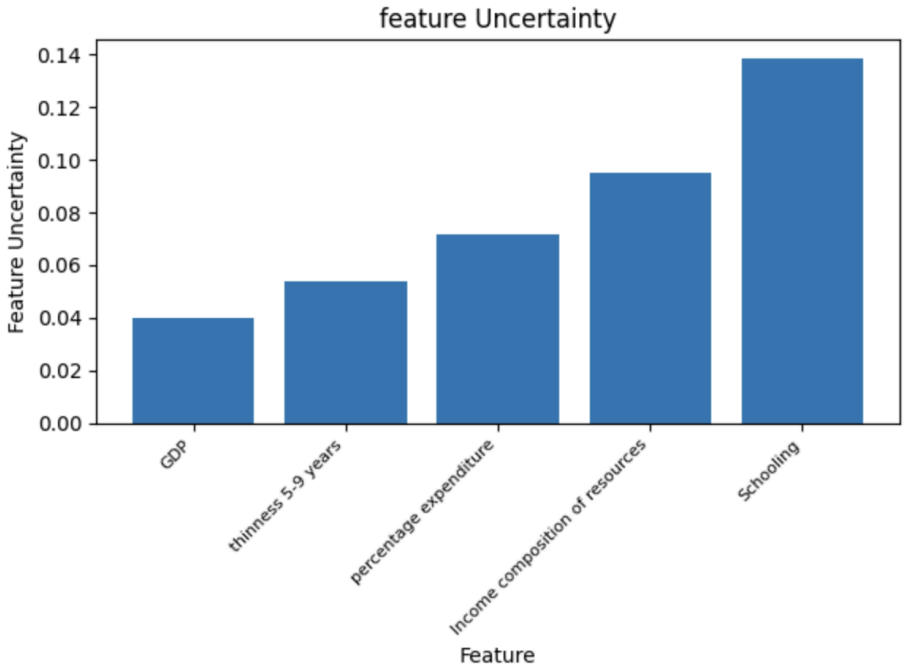

FIGURE 10. Feature Uncertainty Chart

A good value to be under for feature uncertainty is 0.2. Since all values are under this value, and GDP is well under 0.2, we can have confidence in these feature importance values.

5.5. Consistency of Feature Importances Across Models

We analyzed the feature importances of the other three models: linear regression, decision tree regression and random forest regression. Our analysis showed that the feature importance was fairly consistent across all 3 models. Across all 3 models, we found that GDP had the highest feature

importance and GDP's feature importance was almost 10 times higher than the next highest feature importance value. We saw consistency across all models that GDP is the most important feature in determining life expectancy.

6. Limitations and Future Work

6.1. Limitation of R^2 Value

There are a couple of limitations of the R^2 value. The first main issue is that this metric does not output the "goodness" of the model or how well it represents real world patterns. With a higher R^2 value, all that one can understand is that the model is fitted better. The model might be arriving at correct results for different reasons.

Another limitation of this paper was only using the one metric to evaluate our models, the. R^2 value. The reason this was done was because there was a sizable gap in the R^2 values and it was easy to grasp that the random forest regression model was the best by far. More metrics for evaluation would have made our testing more robust, but they would not have changed the ultimate result of this paper.

6.2. Limitation of Removals

As noted in 3.4.2, data was removed to reduce the number of rows of data from 2938 to 2458. While if these removals were isolated without a pattern between the removed data, there would not be a problem. However, a trend is that developing countries tend to have more data missing than developed countries. Hence, much of the data removed was data of developing countries, so the data probably was not as representative of the real world as we would have liked.

6.3. Future Work

This research showed us the importance of economic standing to a country's success. An interesting idea for future work is doing a similar project for a dependent variable relating to economy of countries. We would then investigate what factor(s) affect this dependent variable greatest. Using both these projects and additional research, a plan could be proposed to governments of underdeveloped nations of how to improve their economies, and in turn, improve health in their countries.

7. Conclusion

The three ML models were applied based on the data from the World Health Organization and the model, random forest regression, was chosen to go forward with. Using some methods for feature importances, we have concluded that out of the factors we have been considering for this study, GDP has the greatest impact on life expectancy. GDP means gross domestic product, and literally, it refers to how much product a country exports each year. While GDP does not intuitively tie back to life expectancy, it is important to understand that GDP indicates development and economic status of a country. [4] Essentially, when a country is higher up in the economic standing, the country has more facilities and overall development, leading to prolongation of people's lives. Our study shows that it is in the best interests of countries around the globe to get more economically involved. Firstly, this leads to more global trade opportunities. Finally, economic involvement will lead back to improving the lives of the people in the nation.

References

[1] Freeman, T., Gesesew, H. A., Bambra, C., Giugliani, E. R. J., Popay, J., Sanders, D., Macinko, J., Musolino, C., & Baum, F. (2020, November 10). Why do some countries do better or worse in life expectancy relative to income? an analysis of Brazil, Ethiopia, and the United States of America - International Journal for equity in health. BioMed Central. https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-020-01315-z

[2] Period and cohort life expectancy explained. Period and cohort life expectancy explained - Office for National Statistics. (n.d.). https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/methodologies/periodandcohortlifeexpectancyexplained#:~:text=Life%20expectancy%20is%20a%20statistical,demographic%20factors%20including%20their%20sex.

[3] Effects of education on adult mortality: A global systematic ... (n.d.). https://www.thelancet.com/journals/lanpub/article/PIIS2468-2667(23)00306-7/fulltext

[4] Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: Evidence from the EU Accession Candidate countries. Genus, 76(1). https://doi.org/10.1186/s41118-019-0071-0

[5] Correction to: Impact of healthy lifestyle factors on life expectancies in the US population. (2018). Circulation, 138(4).

https://www.ahajournals.org/doi/10.1161/CIRCULATION%20%20HA.117.032047

[6] Nagai, M., Kuriyama, S., Kakizaki, M., Ohmori-Matsuda, K., Sone, T., Hozawa, A., Kawado, M., Hashimoto, S., & Tsuji, I. (2011). Impact of walking on life expectancy and Lifetime Medical Expenditure: The ohsaki cohort study. BMJ Open, 1(2). https://doi.org/10.1136/bmjopen-2011-000240

[7] Lakshmanarao, A., A, S., T, S. R., G, L., & K, V. K. (2022). Life expectancy prediction through analysis of immunization and HDI factors using machine learning regression algorithms. International Journal of Online and Biomedical Engineering (iJOE), 18(13), 73–83. https://doi.org/10.3991/ijoe.v18i13.33315

[8] Montez, J. K., Hummer, R. A., & Hayward, M. D. (2012). Educational attainment and adult mortality in the United States: A systematic analysis of functional form. Demography, 49(1), 315–336. https://doi.org/10.1007/s13524-011-0082-8

[9] Yogi, T. N., B.C., P., Bhusal, A., Limbu, S., & Kafle, R. (2023). Alcoholic pellagrous encephalopathy: A case report on atypical presentation and diagnostic dilemma in alcohol-related disorders. Annals of Medicine & Surgery, 86(1), 501–506. https://doi.org/10.1097/ms9.0000000000001497

[10] Tranmer, M., Murphy, J., Elliot, M., and Pampaka, M. (2020) Multiple Linear Regression (2nd Edition); Cathie Marsh Institute Working Paper 2020-01. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/2020-1-%20multiple-linear-regression.pdf

[11] 1.10. decision trees. scikit. (n.d.). https://scikit-learn.org/stable/modules/tree.html

[12] Random forests Leo Breiman and Adele Cutler. Random forests - classification description. (n.d.-a). https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[13] Randomforestregressor. scikit. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[14] Numeracy, Maths and statistics - academic skills kit. (n.d.). https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html#:~:text=%C2%AFy)2.-,R%202%20%3D%201%20%E2%88%92%20sum%20squared%20regression%20(SSR)%20total,from%20the%20mean%20all%20squared.