Predicting Life Expectancy and Rethinking the Classification of A Country's Development

Name: Nathan Kai Ching Chan
GT ID: 698
Email: nchan64@gatech.edu
Date: 4/20/2025
Course: ISYE 7406

Abstract

This analysis explores the most important determinants of life expectancy and evaluates the classification of countries by development status. Using a dataset of 179 countries across 16 years from 2000 to 2015, multiple models were compared to estimate life expectancy based on socioeconomic and health variables. Two of the best-performing models were the two ensemble methods (Gradient Boosting and Random Forest), as they showcased that the most influential predictors of life expectancy are mortality rates, followed by alcohol consumption and years of schooling.

A country's development is classified using K-means clustering, which is directly compared to the World Health Organization's binary classification of "developed" and "developing." K-means clustering revealed that the majority of countries fell into the upper and middle development groups, indicating that existing labels may be oversimplifying development. Although the WHO's method focuses on economic indicators, the clustering approach introduces healthcare and educational factors to capture the sophisticated nature of development.

Both objectives illustrate the drawbacks to using simple approaches to drive policymaking and highlight the benefits of ensemble methods and unsupervised learning. Policymakers can allocate resources to prioritize access to healthcare, education, and substance abuse training.

Introduction

Life expectancy is an essential indicator of a country's health and economic well-being. Institutions such as the World Bank, United Nations, and national governments leverage this measurement as guidance for resource allocation decision-making in areas that affect healthcare planning and economic investments. Despite its significance, many countries continue to suffer from the misallocation of resources in public health, which can be attributed to a lack of understanding of what factors drive life expectancy. The inefficient allocation of resources can lead to preventable deaths, limited economic development, and restricted access to healthcare. Therefore, the goal is to identify the most important healthcare and socioeconomic factors of life expectancy so that policymakers can effectively invest into resources to improve a country's well-being.

An additional objective of this analysis is to examine the discourse surrounding the classification of a country's development. A widely-used classification approach is the binary labeling of countries as either "developed" or "developing" – a method utilized by the World Health Organization/United Nations. While it is straightforward, some may argue that a country's development is oversimplified by focusing on economic indicators while paying less attention to other factors such as education level and mortality rates. For instance, the United States may perform exceptionally well on an economic development scale, but while it is often regarded as a "Developed country", it also struggles with healthcare access and has a lower life expectancy

than certain developing countries such as Cuba. Hence, this analysis will include a secondary objective to classify countries by a number of factors beyond economic indicators.

Thus, this project addresses two points. First, it investigates the most crucial healthcare and socioeconomic variables influencing life expectancy. Secondly, it utilizes an unsupervised clustering technique to develop alternative labeling methods for a country's development, and compares these with existing labels to evaluate their differences.

Literature Review

The literature review explores recent studies on the topic of life expectancy determinants. In the first literature by Deshpande et al. (2023), the researchers used a similar dataset to this analysis, and imputed the mean value of columns and the mode for string features to address missing values. The compared models include multiple linear regression and machine learning algorithms, and concluded that the random forest regressor had the best predictive power. The most important predictors were GDP per capita, alcohol consumption, prevalence of diseases, and immunization coverage. However, their analysis lacked robustness and generalizability due to the lack of cross-validation and baseline model comparisons (e.g. simple linear regression).

Roffia et al. (2022) conducted a longitudinal study on the life expectancy of OECD countries. Their methodology involved a fixed-effects regression framework given the nature of panel data, and incorporated lagged variables (between 1 to 5 years) to account for delayed effects. To address endogeneity concerns, the study incorporated multiple regression analysis with year and country dummies, backward stepwise selection, and Arellano-Bond estimators. The key predictors found in this study were education, public health services, physician density, and hospital density. The study also recognizes limitations to data availability, and explanatory power is limited to OECD countries. Thus, it is also worth considering that the variables in the analysis also face similar concerns regarding how potentially essential variables were not included in the dataset.

Data Sources

The dataset is obtained from Kaggle and is named "Life Expectancy WHO (Updated). Each of the variables in the dataset is sourced from a variety of accredited institutions:

- World Health Organization (WHO): Life expectancy, mortality rates, alcohol consumption, vaccine immunizations, BMI, HIV, thinness, and an economic development dummy variable.
- World Bank: GDP per capita, population.
- Our World In Data: Years of schooling.

The initial dataset consists of 21 variables and 2864 observations (179 countries across 16 years from 2000 to 2015). This dataset also builds upon the original Kaggle dataset made 5 years prior, and the key difference is that the author of the updated dataset included their

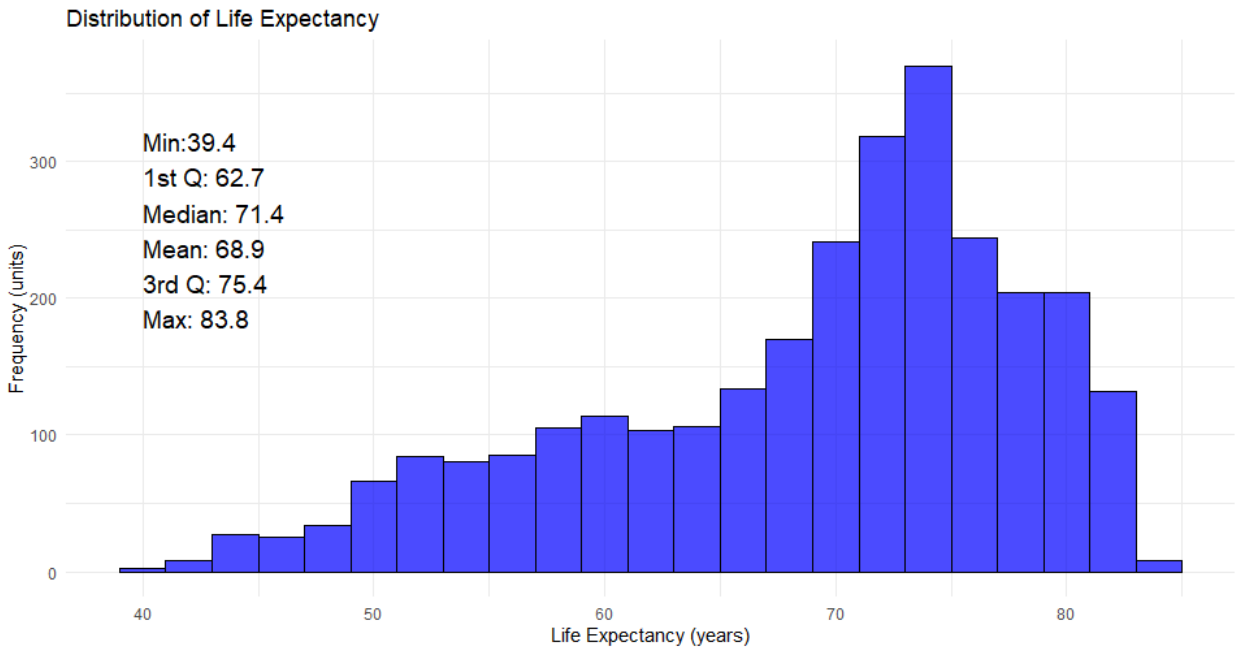
imputation methods to address missing data and outliers. The author's justification of imputation includes:

- For a country's missing values in any year: the closest three-year average.
- For a country's missing values in all years: average of the Region.
- Certain countries with more than four missing columns were omitted.
    - Down from 193 countries in the original dataset to 179 countries in the current dataset.
- The original Kaggle dataset also utilized Missmap in R for imputation to address missing data for smaller countries such as Togo and Cabo Verde.

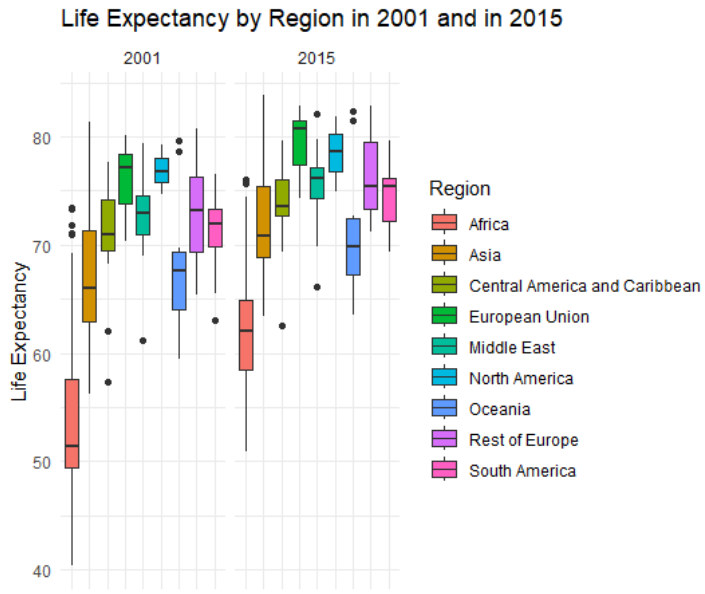
Exploratory Data Analysis

A number of visualizations are created below to provide some insights into the dataset.

1) Distribution of Life Expectancy

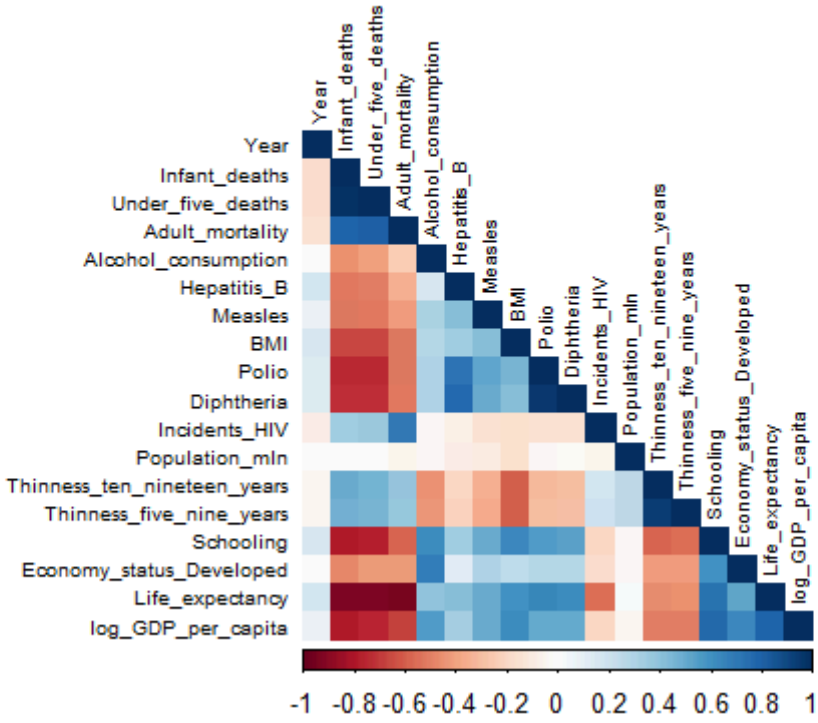


The distribution of life expectancy is right-skewed, with the majority of instances falling between 70 to 80 years. This also implies that techniques that assume a normal distribution will struggle.



2)

The disparities of life expectancy in 2001 were much larger than in 2015, and there is an overall increase in life expectancy throughout the years for all regions concerned. It may be useful to include a regional classification as a feature in model building.



3)

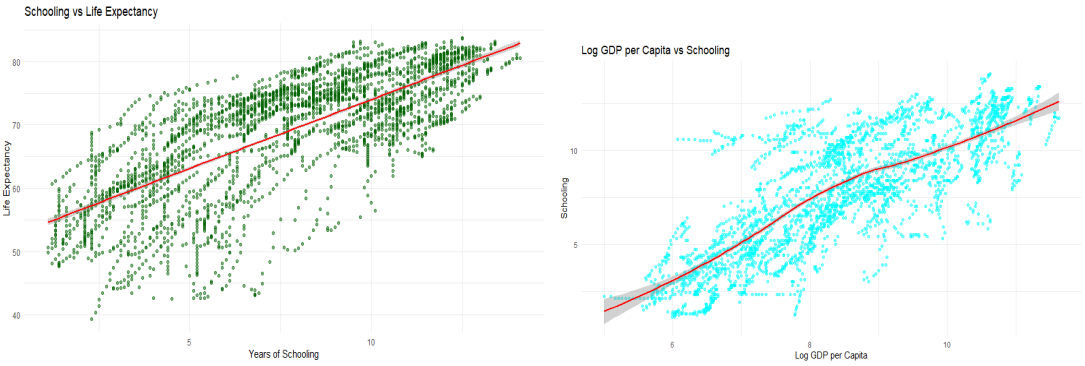
In the correlation matrix, there are some notable correlations against life expectancy:

- Strong positive correlations include schooling and GDP per capita.
- Strong negative correlations include all the mortality factors (infant, under 5, and adult)
- Weaker correlations include vaccination rates and economy status

To address multicollinearity concerns, AIC will be applied for feature selection, particularly to address the correlations between GDP per capita vs. schooling, and the mortality rates.

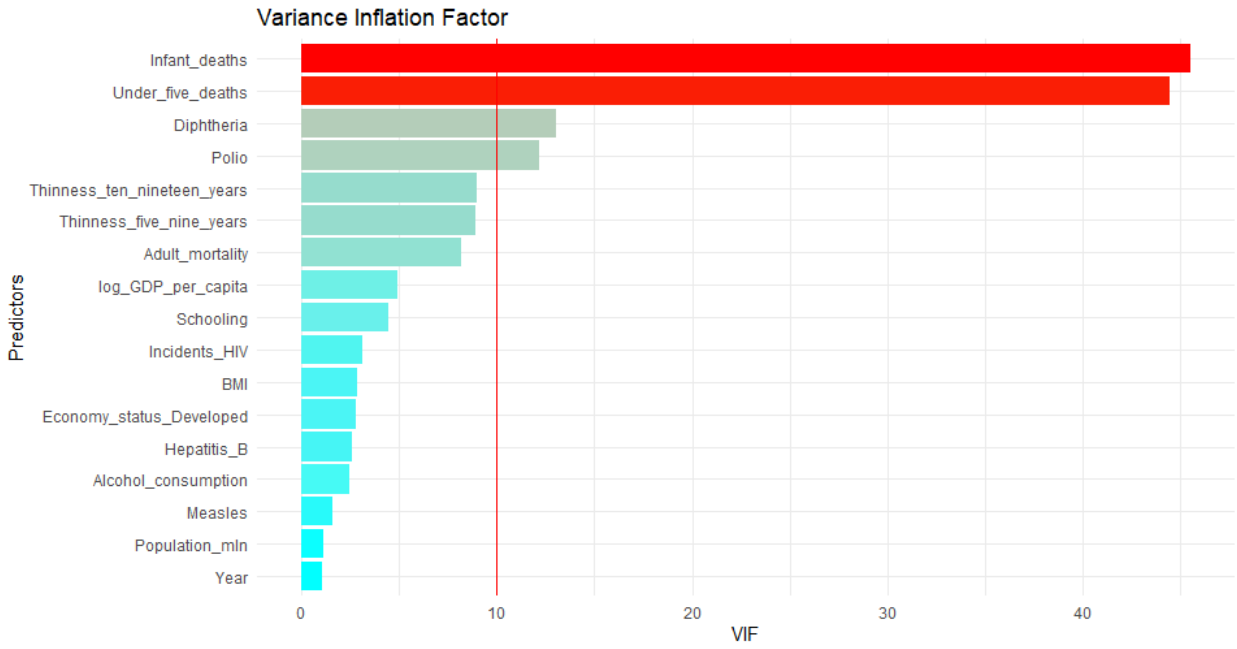
4)

There is a strong nonlinear relationship between GDP per capita and life expectancy. In particular, it shows diminishing returns of life expectancy at higher GDP levels. It makes sense to implement a log transformation on GDP per capita, as the trendline differences (using LOESS) indicates that logGDP per capita appears to be much more linear. I also considered transformations for other variables, but the justification for transformations is not the same in every case (e.g. hinders interpretability). In addition, Random Forest and Boosting do not assume linearity, unlike certain baseline models.



5)

Based on the green plot, years of schooling and life expectancy are strongly positively correlated, and by plotting log GDP per capita against years of schooling in the cyan plot, I immediately thought to test for multicollinearity.



6)

After testing for multicollinearity using VIF, the VIF values are below 5, suggesting that there is some correlation between log GDP per capita and schooling, but it is not serious. In addition, this led me to investigate other variables for multicollinearity. 4 variables were above 10–a loosely defined threshold to indicate severe multicollinearity. Of those four variables, two of them have a value of around 45 (infant deaths per 1,000) and 44 (under five deaths per 1,000). This is extremely undesirable and unsurprising, and I opted to remove infant deaths because under five deaths include infant deaths. Diphtheria (DTP3) and Polio (both represented by the percent coverage of immunization among 1 year olds) have high factors and can be attributed to the likelihood of having coverage of both vaccines. In other words, one may assume that if a country has a high DTP3 vaccination rate, then it likely has high Polio vaccination rates as well. Therefore, Polio is also removed from model considerations as DTP3 is often used as an immunization benchmark (UNICEF, 2024).

Ultimately, the variables selected for modeling are:

- log_GDP_per_capita
- Schooling
- Alcohol_consumption
- BMI
- Diphtheria vaccination
- Hepatitis B vaccination
- Under-five deaths
- Adult mortality
- Year
- Economy_status_Developed (dummy variable)

Proposed Methodology

The methodology will be split into two parts:

1) Model comparisons to find the most significant determinants of life expectancy
2) Clustering to categorize a country's development

Part 1

For cross-validation, it will not utilize a random train-test split but instead a time-aware train-test split of 80/20 due to the longitudinal nature of the dataset. Therefore, the train-test split will be created based on data from the years 2001 to 2012 (training data), and tested on 2013 to 2015, which resembles real-life efforts of predicting future years. The main and baseline methods will be compared by the testing error in a single run evaluation, followed by cross-validation of 100 iterations.

1. Linear Regression

Linear Regression is the first baseline and serves as a benchmark for the succeeding models because of its simplicity. It is expected to struggle with the existence of nonlinear relationships in the dataset. It will estimate the relationship between life expectancy and all predictors, with the exclusion of variables that contributed to multicollinearity.

2. Regression using AIC Stepwise

Feature selection is implemented through minimizing AIC. Population (in millions), Measles, and Thinness from 5-9 years were removed to minimize RSS using the stepwise function, which helps to strike a balance between simplifying models and maintaining predictive power.

3. KNN

KNN is utilized because it is a more straightforward method of tackling nonlinear patterns and classification. Although the optimal k was found to be $k = 1$ based on minimizing test errors, it is not particularly meaningful and is prone to overfitting; $k = 3$ is used instead, as the jump in testing error is relatively minor.

Table 1: KNN Testing Error by K Value

K Value (1-8)	Testing Error (1-8)	K Value (9-15)	Testing Error (9-15)
1	1.058901	9	1.55711352277168
2	1.063841	10	1.59594040968342
3	1.174558	11	1.61175755459624
4	1.267886	12	1.68449029386234
5	1.339035	13	1.71385544852961
6	1.407628	14	1.77625469075277
7	1.444596	15	1.79196994846627
8	1.505638		

4. LOESS

LOESS is limited to 4 predictors in R and is low-dimensional in nature. It is a smoother that allows for fitting many regressions around different points in a dataset to capture complexities beyond the linear scale. Since I have already transformed GDP per capita, I decided on other nonlinear predictors: Alcohol consumption, under-five deaths, adult mortality, and BMI. A sequence of 0.1 to 1 by every 0.05 intervals was tested to find the optimal span parameter. This is found to be 0.9, where the testing error is minimal.

5. Random Forest

Random Forest and GBM are selected because of their ability to handle nonlinear relationships–something that is highly prevalent in this dataset. Due to computational cost, the parameters were tested within a sparse range:

- ntree: 100, 200, 400, 600, 800. Optimal: 600
- nodesize: 1, 3, 5, 7, 9. Optimal: 1
- Mtry: 1, 3, 5, 7, 9, 11, 13, 15. Optimal: 5

6. GBM (Gradient Boosting Machine)

GBM is also helpful for complex data, as it repeatedly introduces decision trees trained on residuals at a rate called shrinkage. Sharing similar computational issues that Random Forest struggles with, the parameters to consider are as follows:

- n.trees: 100, 200, 400, 800, 1600. Optimal: 1600 (more could result in overfitting)
- shrinkage: 0.01 - 0.1 by increments of 0.01. Optimal: 0.09
- interaction.depth: 1, 3, 5. Optimal: 5
- n.minobsinnode: 5, 10, 15. Optimal: 5

For GBM Boosting, 1600 trees is a significant number of iterations to compute reasonably, and a larger number would result in more trees being added to the model to improve fit. The larger the

shrinkage, the more each sequential tree has an effect on the model. Thus, 0.09 is reasonably small given the significant number of trees. Interaction depth at 5 is also a reasonable pairing with the other parameters because, given the number of features in the dataset, a higher interaction depth allows for many variables to interact with each other in a decision tree. Lastly, n.minobsinnode at 5 means that a leaf in each decision tree has to contain no more than 5 data points, which reduces overfitting risks.

Part 2

Using the preprocessed dataset (addressed multicollinearity, transformations, and scaling), K-means clustering is utilized in an attempt to explore how unsupervised learning can introduce alternative methods to the classification methods that are widely used today. K-means clustering is implemented using K = 2 based on the optimal tuning via silhouette method, and K = 3 to provide further context into deeper groupings. As the classifications consider health, education, and socioeconomic variables, the K = 3 clusters were labeled as “High Development”, “Moderate Development”, and “Low Development” to represent a holistic approach.

It is important to note that a dummy variable already exists based on WHO’s 2014 definition and labeling of a country’s economic development. Naturally, there will be comparisons made between the following approaches:

1. WHO’s dummy variable
2. K = 2
3. K = 3

The clustering was only used on 2015 data, as it is the most recent and meaningful year while also maintaining a fair comparison to the WHO definition. Lastly, a confusion matrix compares the first two approaches, along with classification metrics to illustrate immediate differences.

Analysis and Results

Part 1

In the single run scenarios, the testing errors are as listed:

Table 2: Testing Errors (Single Run)

Model	Testing Error
Linear Regression	1.9513384
AIC Regression	1.9500819
KNN (k = 3)	1.1745582
LOESS	1.5799213
Random Forest	0.7029022
GBM	0.2190355

Based on the table of results above, the best-performing models were the ensemble methods of GBM and Random Forest. KNN follows this at k=3, then LOESS, and the traditional regression methods perform the worst. This suggests that the dataset requires complex tools to tackle the abnormal relationships and distributions. Parameters were tuned mainly to optimal settings, with the caveat of computational power and time limitations among the ensemble methods.

Using Monte-Carlo Cross-Validation of 100 iterations, the testing error results are as follows:

Table 3: Mean and Variance of Testing Error (MCCV 100 iterations)

Method	Mean Testing Error	Variance Testing Error
Linear Regression	1.8423995	0.0114867
AIC Regression	1.8425658	0.0114378
KNN	0.4850340	0.0066211
LOESS	1.3395970	0.0057742
Random Forest	0.2337865	0.0011862
GBM	0.2212887	0.0031794

These results seem to indicate that GBM is still the strongest performer. Random Forest significantly improves its mean test error to almost match GBM, while having the lowest variance testing error. Another significant improvement is KNN, which suggests that this model is highly sensitive to random train/test splits. The order largely remains the same, beyond the changes in the order of the main methods, with stepwise regression using AIC as the weakest performer. Thus, the MCCV testing reinforces the notion that this dataset has strong nonlinear relationships and that ensemble methods overshadow linear models due to their ability to capture nonlinear patterns.

Given the significance of the ensemble methods, this analysis will consider what the two methods consider to be the most important factors that affect life expectancy.

1) Random Forest

Random Forest does not look at what variables are significant in the way regression analysis does. However, it is still possible to extract the importance of each factor by looking at percentage increase in test MSE if a specific variable is permuted. In order of importance:

- Adult mortality, population (millions), under-five deaths, and alcohol consumption

2) GBM

GBM can also highlight the most important variables by looking at their relative importance. The top factors in order of importance are:

- Under five deaths (54%), adult mortality (43%), log GDP per capita, alcohol consumption, and schooling
- Mortality rates represent 97% of the relative influence. These two are extremely informative about life expectancy.

It is to be expected that the mortality rates represent the most important determinants of life expectancy. High mortality directly relates to shorter expected lifespans. Besides mortality rates, alcohol consumption is also a strong contender.

Part 2

Using K-means clustering at K = 2, a comparison is drawn with the Economy_status_Developed variable.

Table 4: Confusion Matrix For K=2

	Developing (0)	Developed (1)
Cluster 1	63	0
Cluster 2	79	37

Cluster 1 consists of 63 countries that the WHO classifies as developing, while containing zero countries that the WHO classifies as developed. Cluster 2 can be considered the “developed”

cluster as it contains all the WHO's economically developed countries, while also consisting of 79 countries that would be considered economically developing by WHO's standard. The classification metrics state that recall is perfect because the clusters successfully captured all WHO-developed countries. Still, precision is extremely low at 32%, given that 79 countries would not be classified in cluster 2 by the WHO. The reason for this is that the cluster not only considers similar economic indicators to the WHO, but it also accounts for countries with strong development in health and education (schooling, immunization rates, and low mortality).

Table 5: Confusion Matrix For K=3

	Developing (0)	Developed (1)
High Development	75	37
Low Development	16	0
Moderate Development	51	0

K = 3 helps to illustrate a crucial point in showing that of the 179 countries, only 16 would be classified as Low Development, whereas the majority of what belonged to the developing cluster in K = 2 are seemingly represented as the in-between cluster, or the Moderate Development cluster. All WHO-developed countries are classified as High Development in this clustering method.

Conclusion

Among the models explored, the two ensemble models outperformed other methods in a single run and a Monte-Carlo Cross-Validation of 100 iterations test. Based on the two ensemble methods, life expectancy is predominantly influenced by mortality rates. In addition, alcohol consumption and schooling emerged as important secondary factors. These suggest that policymakers should prioritize investments in public health and education. Children need to be educated on substances, diet, and exercise, and there needs to be improved access to education and healthcare resources for the young. Thus, an approach targeting healthcare systems, education, and substance abuse for young people may be the key to achieving longevity and good health.

The clustering analysis revealed remarkable differences from the WHO's classification of economic development. The majority of countries were grouped into clusters that indicate higher developmental characteristics. This difference can be explained by the inclusion of socioeconomic and health factors, unlike the WHO's one-dimensional approach. The K = 3 approach also illustrated a significant middle-development tier, insinuating that only a handful of countries are truly falling behind in terms of a multidimensional approach. Both classifications serve different purposes, as a policymaker who is only interested in economic indicators, such as GNI per capita, will want to utilize the WHO's classification. However, another policymaker who is interested in a multitude of factors beyond economics may look to consider the K-means clustering, which can provide a better understanding of development as a whole instead of development in terms of economic indicators.

For future work, it would be beneficial to conduct proper panel data analysis techniques (such as exploring fixed effects). Furthermore, looking at different unsupervised learning methods beyond K-means clustering to try and find more reasonable classifications of development

Lessons I Have Learned

1. Literature Review

I received TA feedback for homework 5 to explore the literature review on this topic. The literature review was highly beneficial in helping me explore possible avenues for my analysis. In my case, my 2nd literature explored fixed effects–a concept I did not understand initially (as with panel data analysis in general). Although I did not incorporate panel data analysis techniques, it helped me recognize the inherent flaws in my approach to my dataset. Literature 1 was also helpful because of the similarities in the dataset and machine learning algorithms used, and guided the implementation of methods learned in this course. Reading literature also taught me the debate surrounding the classification of development, which is why I decided to include a secondary objective later. Thus, I would highly recommend the literature review step for applicable cases.

2. EDA

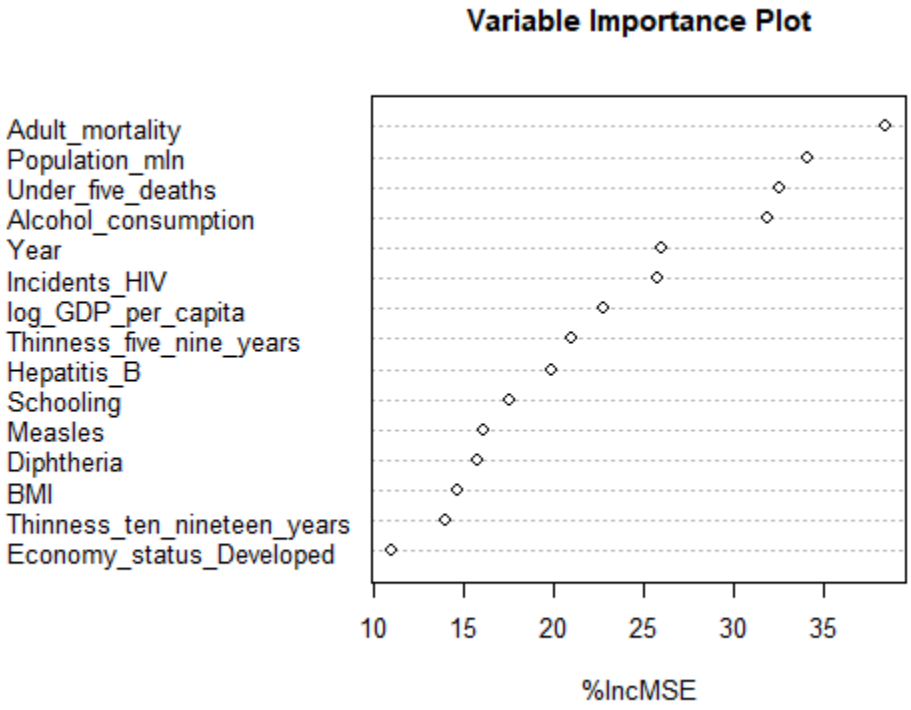
EDA is another crucial step I was able to learn and apply throughout the different homework and this project. It helped me to understand the structure of the dataset and to investigate the relationship among the variables. There were attempts to address violations of linear regression assumptions, such as non-linear patterns and heteroskedasticity in GDP per capita. The result is a much more reliable baseline model using linear regression.

3. Secondary Objective

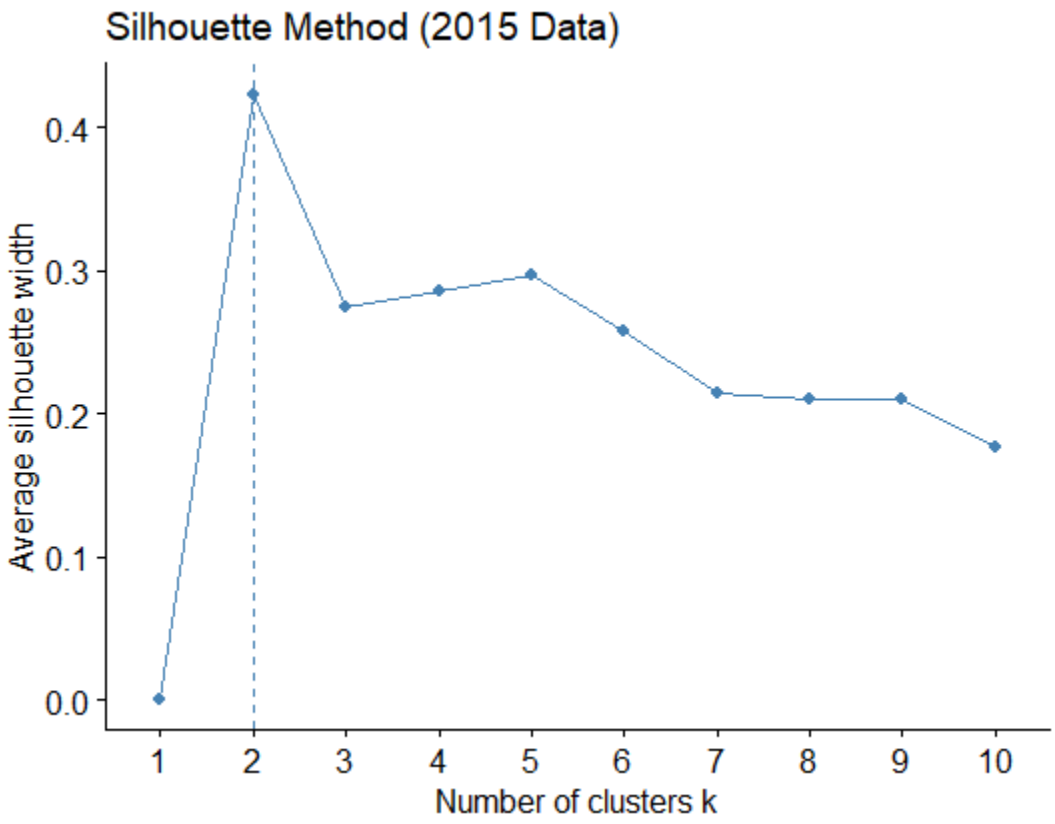
The outcome of the clustering analysis was quite different to what I had envisioned. In particular, I did not expect that many countries to be classified into the stronger development cluster into the K = 2 and K = 3 clustering approaches. This challenged my assumptions as many institutions around the world tend to classify fewer countries into the stronger category. This section raises concerns over the definition and measurement of development, and highlights the difficulty of capturing the nuances with binary labels and simple models.

Appendix

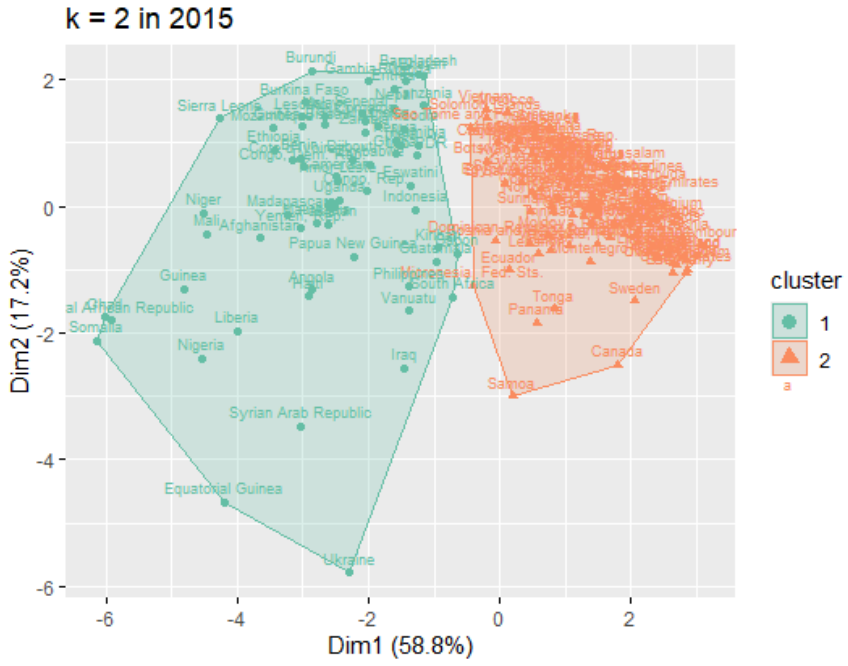
Appendix A: Random Forest Variable Importance Plot

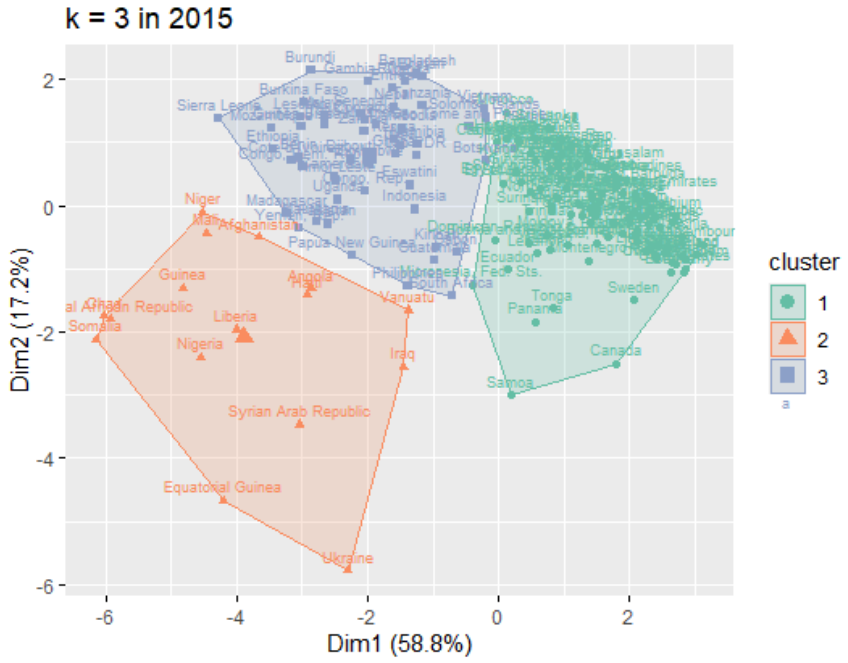


Appendix B: Silhouette Method For Optimal K Clusters



Appendix C: K-Means Clustering Visualizations





Bibliography and Credits

Deshpande, Dr. R., & Uttarkar, V. (2023). Life expectancy using data analytics. *International Journal for Research in Applied Science and Engineering Technology, 11(4)*, 972–978. https://doi.org/10.22214/ijraset.2023.50140

In Focus: Immunization. UNICEF Europe and Central Asia. (2024, November 1). https://www.unicef.org/eca/reports/focus-immunization-2024

Roffia, P., Bucciol, A., & Hashlamoun, S. (2022a). Determinants of life expectancy at birth: A longitudinal study on OECD countries. *International Journal of Health Economics and Management, 23(2)*, 189–212. https://doi.org/10.1007/s10754-022-09338-5

Dataset: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated