# DIAML Assignment 4

Georgina Njorogoe - gnjoroge

November 2025

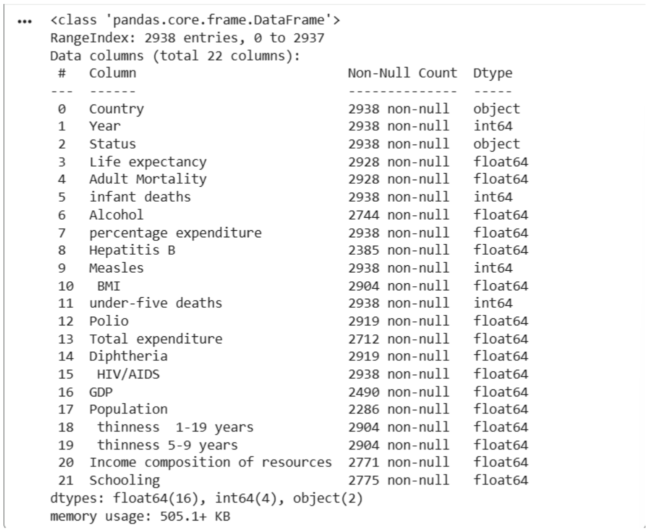
## Question 1: Exploratory Data Analysis

### Methods

The data were loaded from the csv file and viewed to gain understanding of the nature of the data. The data types and summary statistics were also analyzed for further exploration of the data. The column names were cleaned by removing the leading and trailing spaces. Finally, the average life expectancy for developed and developing countries was calculated and visualized as a line plot to analyze the trend.

### Results

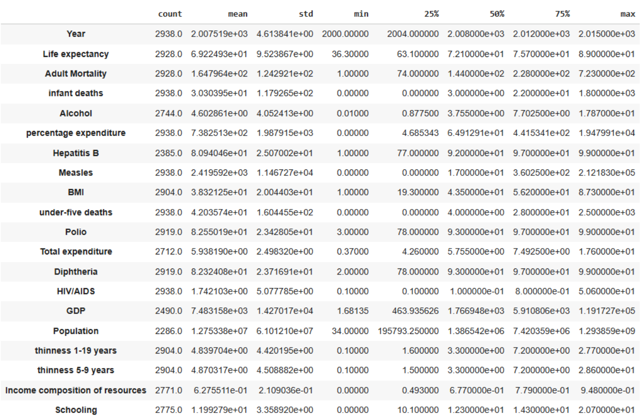
The following are the results of the processes carried out in this step.



```
... <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 2938 entries, 0 to 2937
    Data columns (total 22 columns):
     #   Column                           Non-Null Count  Dtype
    ---  ------                           --------------  -----
     0   Country                          2938 non-null   object
     1   Year                             2938 non-null   int64
     2   Status                           2938 non-null   object
     3   Life expectancy                  2928 non-null   float64
     4   Adult Mortality                  2928 non-null   float64
     5   infant deaths                    2938 non-null   int64
     6   Alcohol                          2744 non-null   float64
     7   percentage expenditure           2938 non-null   float64
     8   Hepatitis B                      2385 non-null   float64
     9   Measles                          2938 non-null   int64
     10   BMI                             2904 non-null   float64
     11  under-five deaths                2938 non-null   int64
     12  Polio                            2919 non-null   float64
     13  Total expenditure                2712 non-null   float64
     14  Diphtheria                       2919 non-null   float64
     15   HIV/AIDS                        2938 non-null   float64
     16  GDP                              2490 non-null   float64
     17  Population                       2286 non-null   float64
     18   thinness  1-19 years            2904 non-null   float64
     19   thinness 5-9 years              2904 non-null   float64
     20  Income composition of resources  2771 non-null   float64
     21  Schooling                        2775 non-null   float64
    dtypes: float64(16), int64(4), object(2)
    memory usage: 505.1+ KB
```

Figure 1: This figure shows the structure of the data, with the of columns, their data types and number of not null values.

From this figure, we can see that the data have 2938 rows and 22 columns. It has two categorical data types, country and status. The rest are numerical data types with infant deaths, measles, and under-five deaths being discrete data types, while the remaining data types such as GDP, Population, and BMI are continuous.



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Year** | 2938.0 | 2.007519e+03 | 4.613841e+00 | 2000.00000 | 2004.000000 | 2.008000e+03 | 2.012000e+03 | 2.015000e+03 |
| **Life expectancy** | 2928.0 | 6.922493e+01 | 9.523867e+00 | 36.30000 | 63.100000 | 7.210000e+01 | 7.570000e+01 | 8.900000e+01 |
| **Adult Mortality** | 2928.0 | 1.647964e+02 | 1.242921e+02 | 1.00000 | 74.000000 | 1.440000e+02 | 2.280000e+02 | 7.230000e+02 |
| **infant deaths** | 2938.0 | 3.030395e+01 | 1.179265e+02 | 0.00000 | 0.000000 | 3.000000e+00 | 2.200000e+01 | 1.800000e+03 |
| **Alcohol** | 2744.0 | 4.602861e+00 | 4.052413e+00 | 0.01000 | 0.877500 | 3.755000e+00 | 7.702500e+00 | 1.787000e+01 |
| **percentage expenditure** | 2938.0 | 7.382513e+02 | 1.987915e+03 | 0.00000 | 4.685343 | 6.491291e+01 | 4.415341e+02 | 1.947991e+04 |
| **Hepatitis B** | 2385.0 | 8.094046e+01 | 2.507002e+01 | 1.00000 | 77.000000 | 9.200000e+01 | 9.700000e+01 | 9.900000e+01 |
| **Measles** | 2938.0 | 2.419592e+03 | 1.146727e+04 | 0.00000 | 0.000000 | 1.700000e+01 | 3.602500e+02 | 2.121830e+05 |
| **BMI** | 2904.0 | 3.832125e+01 | 2.004403e+01 | 1.00000 | 19.300000 | 4.350000e+01 | 5.620000e+01 | 8.730000e+01 |
| **under-five deaths** | 2938.0 | 4.203574e+01 | 1.604455e+02 | 0.00000 | 0.000000 | 4.000000e+00 | 2.800000e+01 | 2.500000e+03 |
| **Polio** | 2919.0 | 8.255019e+01 | 2.342805e+01 | 3.00000 | 78.000000 | 9.300000e+01 | 9.700000e+01 | 9.900000e+01 |
| **Total expenditure** | 2712.0 | 5.938190e+00 | 2.498320e+00 | 0.37000 | 4.260000 | 5.755000e+00 | 7.492500e+00 | 1.760000e+01 |
| **Diphtheria** | 2919.0 | 8.232408e+01 | 2.371691e+01 | 2.00000 | 78.000000 | 9.300000e+01 | 9.700000e+01 | 9.900000e+01 |
| **HIV/AIDS** | 2938.0 | 1.742103e+00 | 5.077785e+00 | 0.10000 | 0.100000 | 1.000000e-01 | 8.000000e-01 | 5.060000e+01 |
| **GDP** | 2490.0 | 7.483158e+03 | 1.427017e+04 | 1.68135 | 463.935626 | 1.766948e+03 | 5.910806e+03 | 1.191727e+05 |
| **Population** | 2286.0 | 1.275338e+07 | 6.101210e+07 | 34.00000 | 195793.250000 | 1.386542e+06 | 7.420359e+06 | 1.293859e+09 |
| **thinness 1-19 years** | 2904.0 | 4.839704e+00 | 4.420195e+00 | 0.10000 | 1.600000 | 3.300000e+00 | 7.200000e+00 | 2.770000e+01 |
| **thinness 5-9 years** | 2904.0 | 4.870317e+00 | 4.508882e+00 | 0.10000 | 1.500000 | 3.300000e+00 | 7.200000e+00 | 2.860000e+01 |
| **Income composition of resources** | 2771.0 | 6.275511e-01 | 2.109036e-01 | 0.00000 | 0.493000 | 6.770000e-01 | 7.790000e-01 | 9.480000e-01 |
| **Schooling** | 2775.0 | 1.199279e+01 | 3.358920e+00 | 0.00000 | 10.100000 | 1.230000e+01 | 1.430000e+01 | 2.070000e+01 |

Figure 2: This figure shows the summary statistics for the dataset including the max, min and mean values for each numeric column.

From this figure, we can see that the average life expectancy is 69 years with the least being 36 and the highest 89 years. The data range from 2000 to 2015. In addition, adult mortality has a large standard deviation of 317. School years range from 11 years to 20 years. GDP has a wide range with a large difference between the minimum and maximum values. The cases of measles are highly skewed, with the mean at 2419 and the maximum value at more than 200,000.

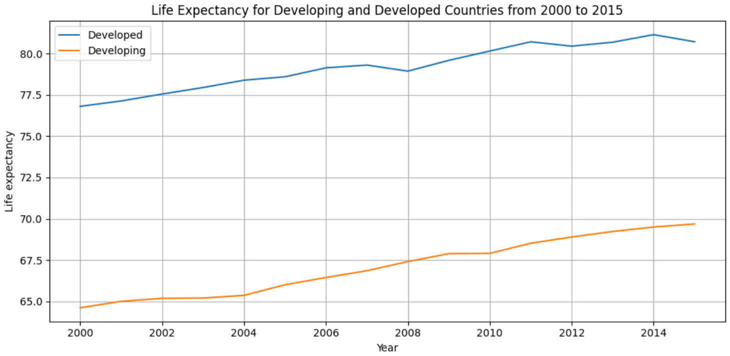


Figure 3: This plot shows the yearly average life expectancy for developed and developing countries.

From this figure, we can see that the life expectancy in developed countries has remained consistently higher than that in developed countries over the years. In general, life expectancy has gradually increased from 2000 to 2015.

### Discussion

It is important to analyze data distributions and temporal patterns before applying machine learning. This is because we can understand the nature of the data and fix issues such as skewness, outliers, and missing data. This informs data cleaning and transformation methods that ensure that data are ready for modeling. In addition, understanding the patterns in the data helps to preserve trends that improve the performance of the model.

There are multiple factors that affect life expectancy in countries. We see that some countries have very high alcohol consumption and disease outbreaks such as measles, which is highly biased. In addition, there is a large financial gap between developed and developing countries. Regarding education levels, while some countries have really low education ages, some have high levels of education. Finally, adult mortality indicates large disparities across the years and between different countries.

From the plot, we can see that generally developed countries have a longer life expectancy, which is expected. This is because they have better facilities and resources that reduce the effects of diseases such as Measles, reducing mortality rates, and thus causing their life expectancy to be higher. In addition, life expectancy has improved over the years in both developed and developing countries. This is also expected because of the innovations and improved quality of life over the years that allow better management of factors that influence mortality rates, such as vaccines and effective treatment of diseases such as HIV/AIDS.

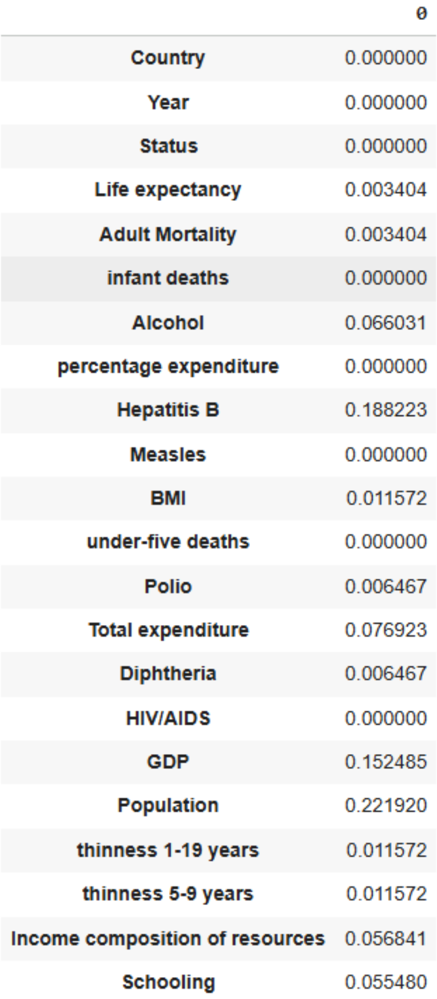
## Question 2: Data Cleaning

### Methods

In this section, the data was analyzed to see the proportion of missing values per column. Then, a population was selected to perform median and KNN-imputation and compare the results. The results were then visualized on line plots. All other numerical columns were then imputed using the KNN values. Outliers in GDP and Percentage Expenditure were dealt with using IQR to determine the outliers, since they were heavily skewed and capping was used instead of dropping to prevent losing information in other columns. Box plots were used to show the differences before and after dealing with the outliers

## Results

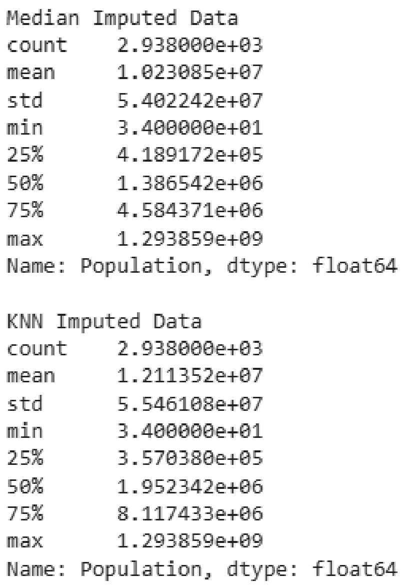
The results for the above methods are as follows.



| | 0 |
|---|---|
| Country | 0.000000 |
| Year | 0.000000 |
| Status | 0.000000 |
| Life expectancy | 0.003404 |
| Adult Mortality | 0.003404 |
| infant deaths | 0.000000 |
| Alcohol | 0.066031 |
| percentage expenditure | 0.000000 |
| Hepatitis B | 0.188223 |
| Measles | 0.000000 |
| BMI | 0.011572 |
| under-five deaths | 0.000000 |
| Polio | 0.006467 |
| Total expenditure | 0.076923 |
| Diphtheria | 0.006467 |
| HIV/AIDS | 0.000000 |
| GDP | 0.152485 |
| Population | 0.221920 |
| thinness 1-19 years | 0.011572 |
| thinness 5-9 years | 0.011572 |
| Income composition of resources | 0.056841 |
| Schooling | 0.055480 |

Figure 4: This figure shows the proportions of missing values per column.

From this figure, we can see that the columns with the most missing values are Population, Hepatitis B, and GDP. Some columns, such as Year, Status, and Infant death, do not have missing values.



```
Median Imputed Data
count    2.938000e+03
mean     1.023085e+07
std      5.402242e+07
min      3.400000e+01
25%      4.189172e+05
50%      1.386542e+06
75%      4.584371e+06
max      1.293859e+09
Name: Population, dtype: float64

KNN Imputed Data
count    2.938000e+03
mean     1.211352e+07
std      5.546108e+07
min      3.400000e+01
25%      3.570380e+05
50%      1.952342e+06
75%      8.117433e+06
max      1.293859e+09
Name: Population, dtype: float64
```

Figure 5: This figure shows the summary statistics comparison for median and KNN imputed population data.

From this figure, we can see that the means are approximately equal to each other but KNN has a higher mean. We also see a difference in the distribution; with median imputation, the inter-quartile range is from 400,000 to 4.5 million while for KNN, it ranges from 350,000 to 8 million.

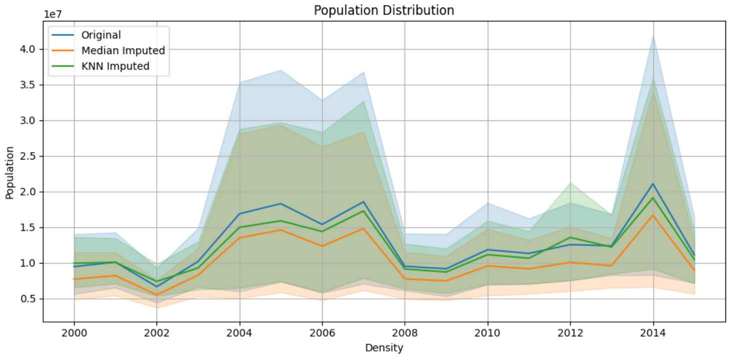


Figure 6: This figure shows the distributions of the original population compared to the KNN and median imputed data.

From this figure there is not much disparity in the population distributions as they seem to follow a similar trend. However, KNN seems to follow the original distribution more closely compared to the median imputed data.

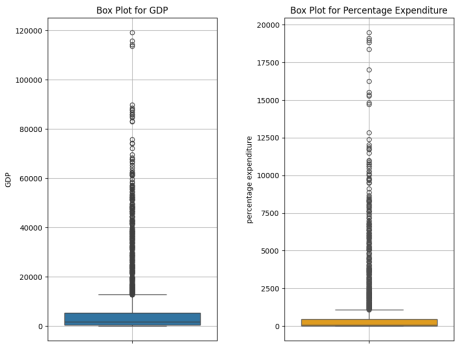


Figure 7: This figure shows the distribution for GDP and Percentage Expenditure before the outliers were dealt with.

From this figure, we can see that both columns are heavily skewed to the right, with many values above the 75th percentile

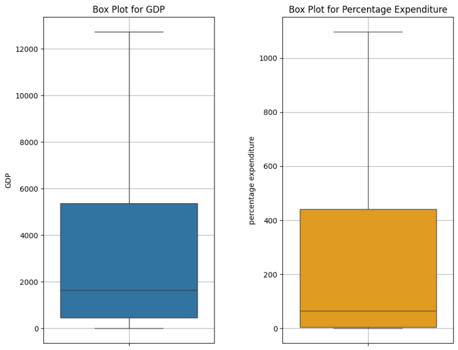


Figure 8: This figure shows the distribution for GDP and Percentage Expenditure after the outliers were dealt with.

From this figure, we can see that both columns are still skewed to the right. However, outliers above the 75th percentile have been removed.

## Discussion

Outliers can distort the performance of the regression model and introduce bias by pulling the line toward themselves, causing the regression line to be skewed. In addition, they can inflate error metrics, making the model seem worse than it is. Outliers could also obscure the trends of the underlying patterns in the majority of the data.

After dealing with missing values and outliers for GDP and Percentage Expenditure, the data are still skewed to the right. This is expected because most countries make less income and have a smaller expenditure budget, causing most of them to concentrate on lower values. The developed countries are outliers with very high values in some cases. Retaining these data as is could affect the model, causing it to ignore the reality of most of the countries while predicting life expectancy.

# Question 3: Feature Engineering

## Methods

For this section, new features such as infant survival rate, vaccination coverage index, and education income index were created from existing features. Five features, BMI, Schooling, GDP, Vaccination Coverage index, and Adult mortality, were selected as independent features that influence life expectancy. Then these features were plotted on histograms for developed and developing countries to see how they compare. Finally, a correlation heatmap was plotted to visualize the relationships among variables. The correlation values with life expectancy were listed in descending order to see the columns with the strongest positive and negative correlations.

## Results

The following are the results for this section.

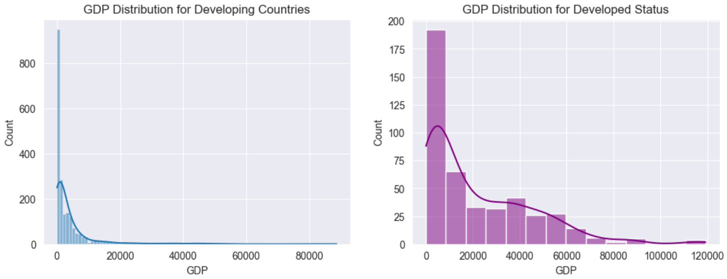


Figure 9: These histograms show the distributions for GDP in developed and developing countries

From this figure, we can see that while for developing countries, the values cluster around the low values, less than 2000, most developed countries cluster at the other end, at 12000.

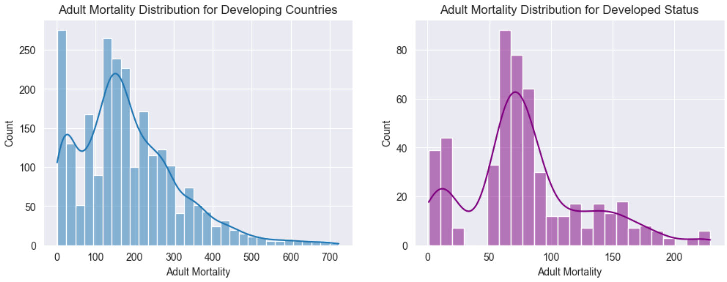


Figure 10: These histograms show the distributions for Adult mortality in developed and developing countries

From this figure, we can see that developed and developing countries have similar shaped distributions with two peaks, one at the start and the other closely after. The second peak for developing countries is between 100 and 200, while that of developed countries is between 50 and 100.

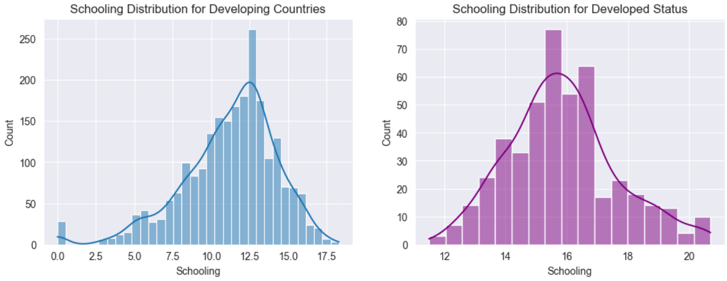


Figure 11: These histograms show the distributions for Schooling in developed and developing countries

From this figure, we can see that even though the distributions are similarly shaped, with most values clustering in the middle, the age ranges differ. For developing countries, most values are between 10 and 15 years, while for developed countries they range between 15 and 17 years, with the highest at 20 years.

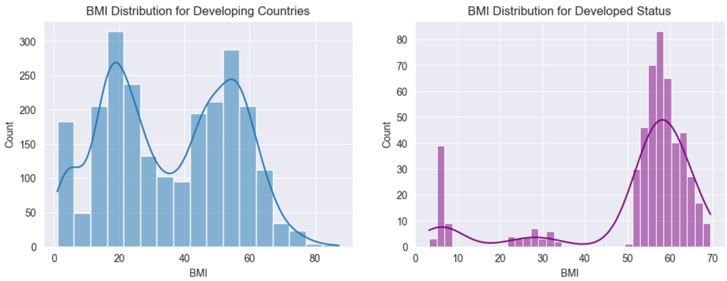


Figure 12: These histograms show the distributions for BMI in developed and developing countries

From this figure, we can see that BMI in developing countries has two peaks, one between 10 and 30 and the other between 40 and 65. In terms of developed countries, most values are grouped between 50 and 65.

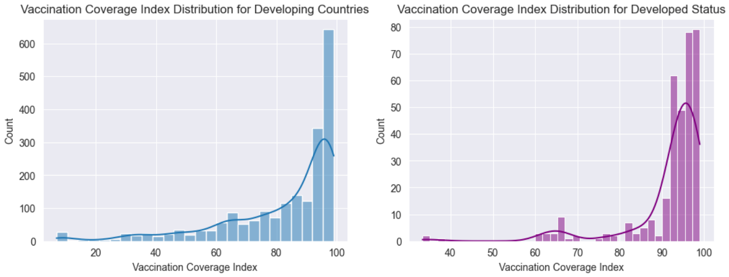


Figure 13: These histograms show the distributions for Vaccination coverage index in developed and developing countries

From this figure, we can see that both distributions are skewed to the right, with most countries having high vaccination coverage. Although the tail in developing countries shows consistent growth, the tail in developed countries has gaps, suggesting outliers.

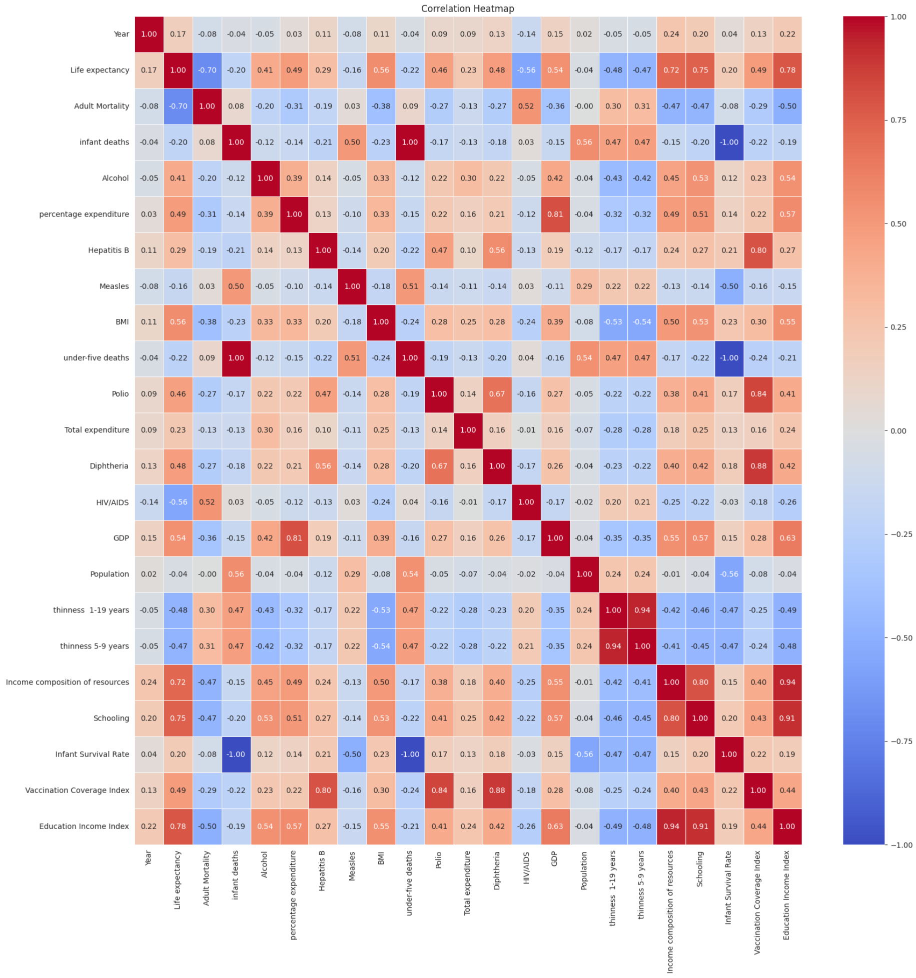
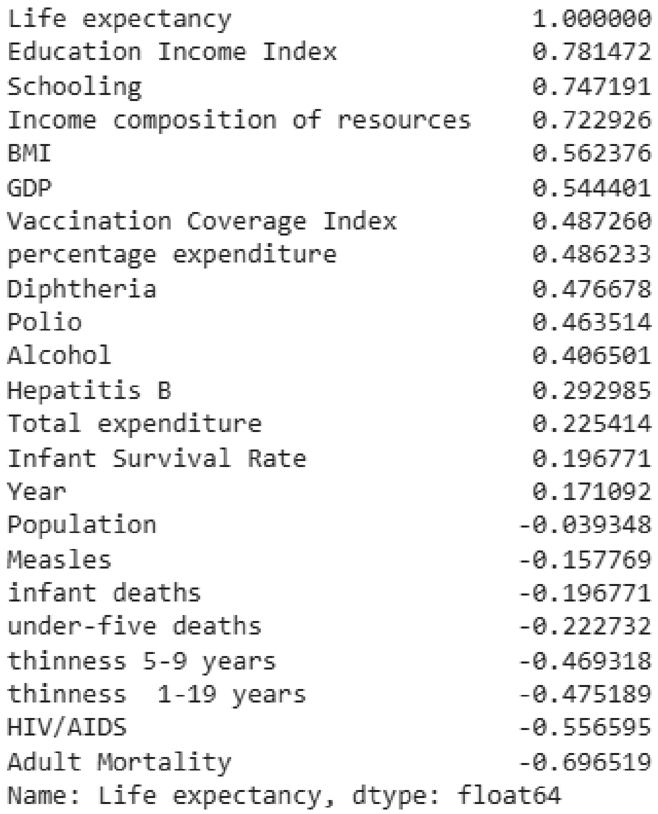


Figure 14: This correlation heatmap shows how the features relate to each other.

From this figure, we can see that features such as thinness between 1 - 19 years and thinness between 5 - 9 years are highly positively correlated with a correlation value of 0.94. Additionally, infant deaths and deaths of under-five years are perfectly positively correlated while infant deaths and infant survival rate are perfectly negatively correlated.



```
Life expectancy                    1.000000
Education Income Index             0.781472
Schooling                          0.747191
Income composition of resources    0.722926
BMI                                0.562376
GDP                                0.544401
Vaccination Coverage Index         0.487260
percentage expenditure             0.486233
Diphtheria                         0.476678
Polio                              0.463514
Alcohol                            0.406501
Hepatitis B                        0.292985
Total expenditure                  0.225414
Infant Survival Rate               0.196771
Year                               0.171092
Population                        -0.039348
Measles                           -0.157769
infant deaths                     -0.196771
under-five deaths                 -0.222732
thinness 5-9 years                -0.469318
thinness  1-19 years              -0.475189
HIV/AIDS                          -0.556595
Adult Mortality                   -0.696519
Name: Life expectancy, dtype: float64
```

Figure 15: This figure shows the correlation of the different features with life expectancy in descending order.

From this figure, we can see that the educational income index has the highest positive correlation with life expectancy. Adult mortality has the highest negative correlation with life expectancy.

## Discussion

Correlation does not imply causation, as it only tells us how two variables move together. One of the reasons why this is the case is that some variables may be correlated by the real cause being an underlying variable. For example, PD can be correlated with life expectancy, but the real cause is access to health care or education, which can all be grouped into GDP. In addition, reverse causation occurs when the target is the cause of one of the features. For example, a longer life expectancy could lead to a higher GDP. Finally, some variables may coincidentally be correlated with no real link in the real world.

Based on the results, we see that the educational income index and schooling show strong positive correlations with life expectancy, suggesting that countries with more education tend to have longer lifespans. This could work through better health knowledge and higher incomes, but both education and health might simply improve together as countries develop economically.

The strong negative correlation between adult mortality and life expectancy makes sense, since adult deaths directly lower life expectancy. This shows why we need to focus on correlations that reveal real patterns.

The GDP distributions show clear differences between developed and developing countries. Although higher GDP relates to longer life expectancy, the direction is unclear as wealth allows better healthcare, but healthier populations might also earn more.

The BMI distributions show that developing countries have two peaks, likely showing both undernutrition and growing obesity problems, while developed countries cluster at higher BMI values, suggesting overeating issues.

The perfect correlations between infant deaths, mortality under-five years of age, and infant survival rate simply show that these measure the same thing in different ways. This matters for modeling, since including all of them would create problems.

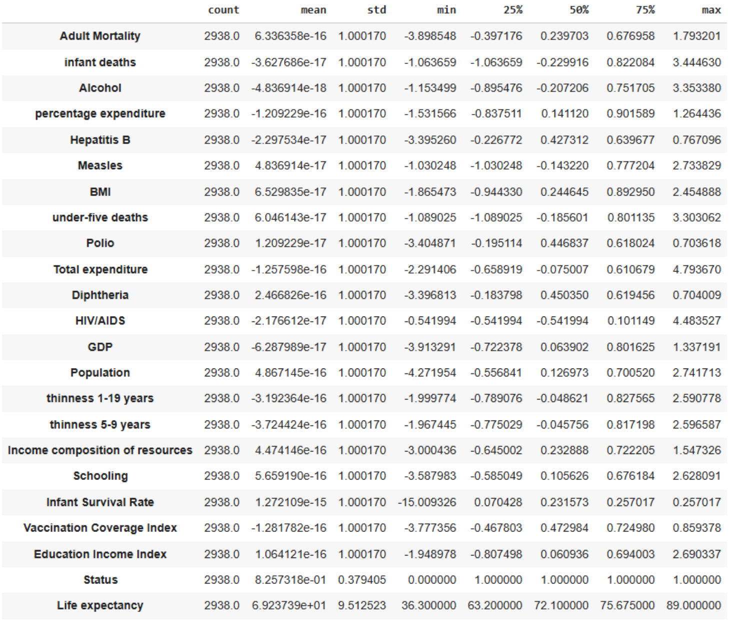
# Question 4: Data Transformation

## Methods

In this section, the data was transformed accordingly. The country axis and the year axis were dropped, as they did not have any significant contribution to life expectancy, but rather acted as indices to identify the data. The status was label encoded since it had two options, making it a binary category. A total of 10 skewed features such as adult mortality and percentage expenditure were identified and transformed using logarithmic transformation, making the distributions more symmetric. Finally, all numerical features, apart from the target column, life expectancy, were scaled using the standard scaler.

## Results

The following are the results of scaling the data.



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Adult Mortality | 2938.0 | 6.336358e-16 | 1.000170 | -3.898548 | -0.397176 | 0.239703 | 0.676958 | 1.793201 |
| infant deaths | 2938.0 | -3.627686e-17 | 1.000170 | -1.063659 | -1.063659 | -0.229916 | 0.822084 | 3.444630 |
| Alcohol | 2938.0 | -4.836914e-18 | 1.000170 | -1.153499 | -0.895476 | -0.207206 | 0.751705 | 3.353380 |
| percentage expenditure | 2938.0 | -1.209229e-16 | 1.000170 | -1.531566 | -0.837511 | 0.141120 | 0.901589 | 1.264436 |
| Hepatitis B | 2938.0 | -2.297534e-17 | 1.000170 | -3.395260 | -0.226772 | 0.427312 | 0.639677 | 0.767096 |
| Measles | 2938.0 | 4.836914e-17 | 1.000170 | -1.030248 | -1.030248 | -0.143220 | 0.777204 | 2.733829 |
| BMI | 2938.0 | 6.529835e-17 | 1.000170 | -1.865473 | -0.944330 | 0.244645 | 0.892950 | 2.454888 |
| under-five deaths | 2938.0 | 6.046143e-17 | 1.000170 | -1.089025 | -1.089025 | -0.185601 | 0.801135 | 3.303062 |
| Polio | 2938.0 | 1.209229e-17 | 1.000170 | -3.404871 | -0.195114 | 0.446837 | 0.618024 | 0.703618 |
| Total expenditure | 2938.0 | -1.257598e-16 | 1.000170 | -2.291406 | -0.658919 | -0.075007 | 0.610679 | 4.793670 |
| Diphtheria | 2938.0 | 2.466826e-16 | 1.000170 | -3.396813 | -0.183798 | 0.450350 | 0.619456 | 0.704009 |
| HIV/AIDS | 2938.0 | -2.176612e-17 | 1.000170 | -0.541994 | -0.541994 | -0.541994 | 0.101149 | 4.483527 |
| GDP | 2938.0 | -6.287989e-17 | 1.000170 | -3.913291 | -0.722378 | 0.063902 | 0.801625 | 1.337191 |
| Population | 2938.0 | 4.867145e-16 | 1.000170 | -4.271954 | -0.556841 | 0.126973 | 0.700520 | 2.741713 |
| thinness 1-19 years | 2938.0 | -3.192364e-16 | 1.000170 | -1.999774 | -0.789076 | -0.048621 | 0.827565 | 2.590778 |
| thinness 5-9 years | 2938.0 | -3.724424e-16 | 1.000170 | -1.967445 | -0.775029 | -0.045756 | 0.817198 | 2.596587 |
| Income composition of resources | 2938.0 | 4.474146e-16 | 1.000170 | -3.000436 | -0.645002 | 0.232888 | 0.722205 | 1.547326 |
| Schooling | 2938.0 | 5.659190e-16 | 1.000170 | -3.587983 | -0.585049 | 0.105626 | 0.676184 | 2.628091 |
| Infant Survival Rate | 2938.0 | 1.272109e-15 | 1.000170 | -15.009326 | 0.070428 | 0.231573 | 0.257017 | 0.257017 |
| Vaccination Coverage Index | 2938.0 | -1.281782e-16 | 1.000170 | -3.777356 | -0.467803 | 0.472984 | 0.724980 | 0.859378 |
| Education Income Index | 2938.0 | 1.064121e-16 | 1.000170 | -1.948978 | -0.807498 | 0.060936 | 0.694003 | 2.690337 |
| Status | 2938.0 | 8.257318e-01 | 0.379405 | 0.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Life expectancy | 2938.0 | 6.923739e+01 | 9.512523 | 36.300000 | 63.200000 | 72.100000 | 75.675000 | 89.000000 |

Figure 16: This figure shows the summary statistics for the transformed data

From this we can see that apart from life expectancy and status, the rest of the values are on a similar scale, as evidenced by the standard deviation column with a constant standard deviation for the other columns.

## Discussion

The data transformation steps were essential for preparing the data set for machine learning. The Status variable was label encoded into binary values since it only has two categories, making it suitable for the models. Country and Year were dropped because they serve as identifiers rather than predictive features, and including them could lead to overfitting on specific countries or time periods instead of learning general patterns.

Ten features showed significant right skewness and were transformed using logarithmic transformation to create more symmetric distributions. This reduces the influence of extreme values and helps models identify meaningful patterns. All numerical features except Life Expectancy were then scaled using StandardScaler, which standardizes the features to have mean zero and standard deviation one.

Scaling is particularly important for certain models. KNN relies on distance calculations to find similar observations, so features with larger values would dominate these calculations without scaling. For example, Percentage Expenditure in thousands would overshadow HIV/AIDS prevalence ranging 0-100, causing KNN to essentially ignore smaller-scaled features. Gradient Boosting is less sensitive since it uses decision trees, but still benefits from scaling because the learning rate and regularization parameters work more effectively when features are on comparable scales. This standardization ensures that all features contribute appropriately based on their predictive power rather than their measurement units, leading to better and fairer model performance.

## Question 5: Model Training and Evaluation

### Methods

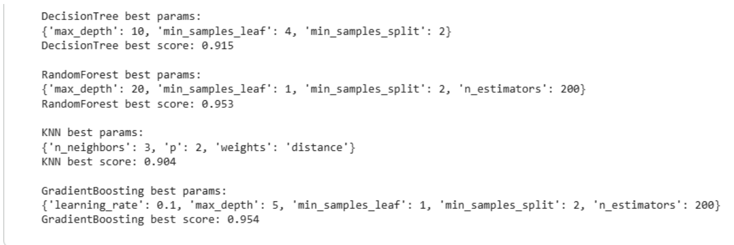
In this section, the data was split into 80% for training and 20% for testing. In the grid search, 5-fold cross-validation was used to find the best hyperparameters for models; Decision Tree, Random Forest, KNN and Gradient Boosting.

For Decision Tree, we tested max depth, min sample split and min samples leaf. Random Forest used the same parameters plus the number of estimators. KNN tested the number of neighbors, weights and distance metric. Gradient Boosting had the most parameters, adding the learning rate to the Random Forest setup. For reproducibility, a random state of 42 was set for all models.

The best parameter models were evaluated using the test set to see which performed better on unseen data. The metrics used for the evaluation were R-squared, Mean Absolute Error and Root Mean Squared Error.

### Results

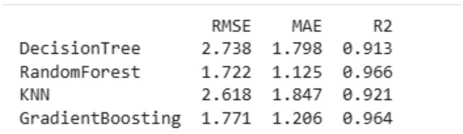
The results are as follows.



```
DecisionTree best params:
{'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2}
DecisionTree best score: 0.915

RandomForest best params:
{'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
RandomForest best score: 0.953

KNN best params:
{'n_neighbors': 3, 'p': 2, 'weights': 'distance'}
KNN best score: 0.904

GradientBoosting best params:
{'learning_rate': 0.1, 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}
GradientBoosting best score: 0.954
```

Figure 17: This image shows the best performing model scores and the best combination of parameters.

From this figure, we can see that the best of the best performing model was the gradient boosting model with a learning rate of 0.1, maximum depth of 5, a min sample leaf of 1, a min sample split of 2, and 200 estimators. It has a score of 0.954 followed closely by the random forest model at 0.953. The model with the lowest score was the KNN model with 0.904.



```
                   RMSE    MAE     R2
DecisionTree      2.738  1.798  0.913
RandomForest      1.722  1.125  0.966
KNN               2.618  1.847  0.921
GradientBoosting  1.771  1.206  0.964
```

Figure 18: This image shows the performance of the best models on the unseen test set.

From this figure, we can see that the best performing model in the test set was the random forest model with an r-squared score of 0.966. It was closely followed by the gradient boost model at 0.964. The worst performing model was the decision tree with a score of 0.913.

### Discussion

Based on the model evaluation results, we see that Random Forest outperformed Gradient Boosting which had the highest cross-validation score during training. This implies that it generalizes better to unseen data.

The performance of the top two models, Random Forest and Gradient Boosting, makes sense since they are both ensemble models, unlike the simple KNN and decision tree models. KNN performed the worst with an R-squared score of about 0.904. This could be because of the curse of dimensionality. As features increase, the distance loses meaning, causing the model to perform poorly.

Hyperparameter tuning was important to control the bias-variance trade-off. Parameters such as max depth and min samples prevent the models from becoming too complex and memorizing the training data, leading to overfitting. On the other hand, too much regularization could cause the model to become too simple, leading to underfitting. The grid search is useful for identifying the parameters that achieve the best balance, with the models being complex enough to learn meaningful relationships, but constrained enough to generalize well.

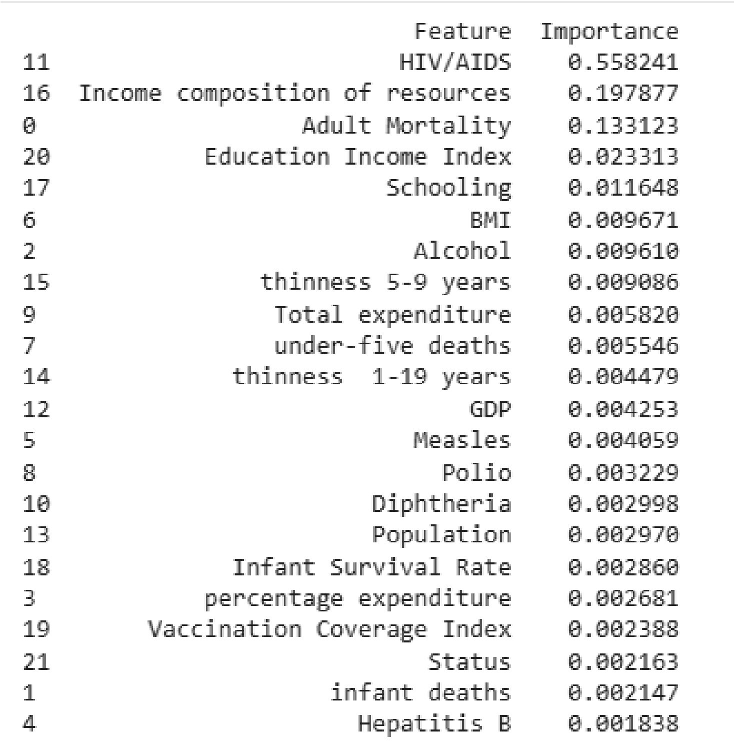
## Question 6: Feature Importance

### Methods

In this section, the importance of the feature was extracted from the best performing model, which was the random forest model. The importance were listed in descending order to allow for easier selection of the most influential features. The top 10 features were selected and used to retrain the best models. These models were then re-evaluated on the test set, using the same metrics as before, RMSE, MAE, and R-squared. Finally, the results were combined into a table for easier comparison.

## Results

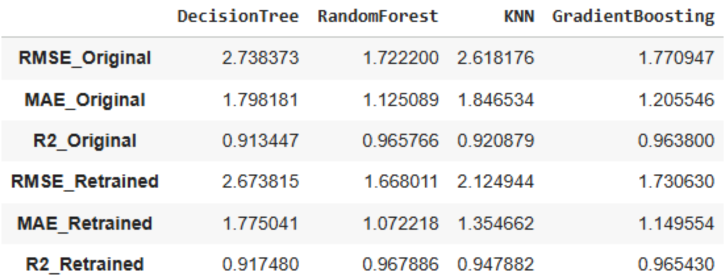
The results are as follows.



```
                             Feature  Importance
11                          HIV/AIDS    0.558241
16  Income composition of resources    0.197877
0                    Adult Mortality    0.133123
20            Education Income Index    0.023313
17                         Schooling    0.011648
6                                BMI    0.009671
2                            Alcohol    0.009610
15                thinness 5-9 years    0.009086
9                  Total expenditure    0.005820
7                  under-five deaths    0.005546
14              thinness  1-19 years    0.004479
12                               GDP    0.004253
5                            Measles    0.004059
8                              Polio    0.003229
10                        Diphtheria    0.002998
13                        Population    0.002970
18              Infant Survival Rate    0.002860
3             percentage expenditure    0.002681
19         Vaccination Coverage Index    0.002388
21                            Status    0.002163
1                      infant deaths    0.002147
4                        Hepatitis B    0.001838
```

Figure 19: This figure shows the importance of features listed from the most important feature to the least.

From this figure, we can see that the top 10 features started from HIV/AIDS to under-five deaths. These ranged from 0.558 to 0.005.



| | DecisionTree | RandomForest | KNN | GradientBoosting |
|---|---|---|---|---|
| **RMSE_Original** | 2.738373 | 1.722200 | 2.618176 | 1.770947 |
| **MAE_Original** | 1.798181 | 1.125089 | 1.846534 | 1.205546 |
| **R2_Original** | 0.913447 | 0.965766 | 0.920879 | 0.963800 |
| **RMSE_Retrained** | 2.673815 | 1.668011 | 2.124944 | 1.730630 |
| **MAE_Retrained** | 1.775041 | 1.072218 | 1.354662 | 1.149554 |
| **R2_Retrained** | 0.917480 | 0.967886 | 0.947882 | 0.965430 |

Figure 20: This table shows the comparison of the model scores when trained with the original data and when trained with the top 10 important features.

From this figure, we can see that there was improvement across the board for all models. Their performance order did not change with the random forest retaining the top position and the decision tree remaining as the worst performing model.

## Discussion

The feature importance analysis reveals that HIV/AIDS is the most influential predictor of life expectancy, with an importance score of 0.558. This makes sense since the virus directly influences mortality rates and is a major concern in developing countries.

Using the most important features to retrain the models improved their performance. This implies that the removed features contained noise or were redundant that prevented the model from focusing on what really mattered. This mitigated the curse of dimensionality, where more features do not necessarily result in better model performance.

Reducing the number of features reduces the complexity of the model and improves the interpretability since there are fewer features. Therefore, it is easier to understand what influences life expectancy compared to analyzing all original features. Despite these benefits, there is a likelihood that important information is lost, which causes the model to leave out some specific trends that may not be evident in the important features.

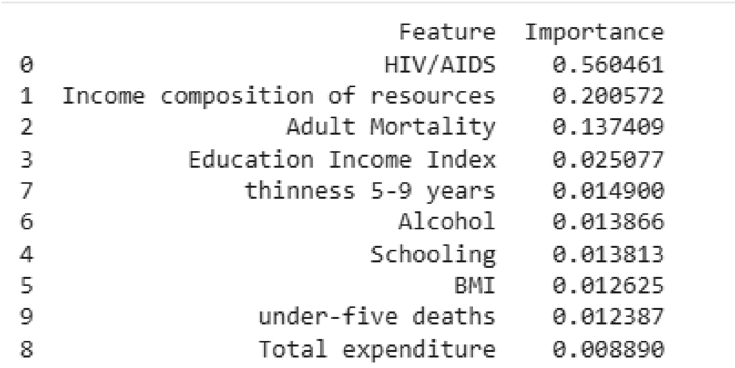
# Question 7: Feature Importance - Continued

## Methods

In this final section, the top 5 important features were extracted from the overall best performing model. This was the random forest model retrained on the top 10 important features. These 10 features were listed in descending order according to their level of importance, allowing easy selection of the top 5 features.

## Results

The results are as follows.



```
                             Feature  Importance
0                           HIV/AIDS    0.560461
1    Income composition of resources    0.200572
2                    Adult Mortality    0.137409
3             Education Income Index    0.025077
7                 thinness 5-9 years    0.014900
6                            Alcohol    0.013866
4                          Schooling    0.013813
5                                BMI    0.012625
9                  under-five deaths    0.012387
8                  Total expenditure    0.008890
```

Figure 21: This figure shows the list of features used to train the model, listed from most important to least important.

From this figure, we can see that the 5 main features are HIV/AIDS, the composition of the income from resources, adult mortality, the income index of education, and thinness between 5 and 9 years.

## Discussion

The top five features identified give important insights as to what influences life expectancy. HIV/AIDS remains the most important predictor as it directly influences mortality. In addition, adult mortality is also important because it directly measures the number of deaths in adults.

The income composition of the resources and the income index of education highlight the socioeconomic factors that influence life expectancy. These features show how wealth and education are distributed, which directly affects healthcare and nutrition. In addition, childhood thinness, which is an indicator of malnutrition that could lead to death, or adults with weak immune systems that are prone to diseases.

The ranking in the correlation in question 3 differs with the importance of the characteristic since the two measure different things. Simple correlation measures the linear relationship between the individual features and the target column, while feature importance measures how features interact and contribute to predictions when considered together.

In the real world, this information could help inform policies such as prioritizing interventions addressing HIV/AIDS and investing in healthcare, education, and economic development. Addressing these issues could lead to increasing life expectancy.