**ISHIMWE KAREKEZI GUY GAEL**

**AndrewID: iguygael**

**MS EAI**

**Email: iguygael@andrew.cmu.edu**

**Tel: (+250) 784595484**

**Course: DIAML**

# DIAML Assignment 4

**Used Libraries**

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib
import sklearn
from matplotlib import pyplot as plt
from pandas import DataFrame
from IPython.display import display
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.impute import KNNImputer
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import r2_score
from sklearn.metrics import explained_variance_score
```

# Q1)

**Importance of analysing distributions and temporal patterns before implementing ML models**

Analysing the distributions before using ML models is important because it helps us identify normality, check for outliers, and verifying our data collection systems[1]. This helps us choose the correct statistical test and ensure that our results are accurate.

Analysing temporal patterns is crucial because time-based patterns are order-based so ignoring it will lead to model failure. The main issue comes from data leakage which is when a model during training accidentally accesses future information. For example, training based on dates[2].

**Steps and Explanations:**

For the coding part of the question, we used python to identify the structure and descriptive statistics of the dataset. We observed the trend of life expectancy between developed and developing countries from 2000 to 2015.

**Results:**

**i) Dataset information:**

```
The dataset Information :
<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)
```

Our dataset contains two non-numeric columns which are country and status.

**ii) Summary statistics**

| | Life expectancy | Adult Mortality | infant deaths | Alcohol | percentage expenditure | Hepatitis B | Measles | BMI | under-five deaths | Polio | Total expenditure | Diphtheria | HIV/AIDS | GDP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 2928.000000 | 2928.000000 | 2938.000000 | 2744.000000 | 2938.000000 | 2385.000000 | 2938.000000 | 2904.000000 | 2938.000000 | 2919.000000 | 2712.00000 | 2919.000000 | 2938.000000 | 2490.000000 |
| mean | 69.224932 | 164.796448 | 30.303948 | 4.602861 | 738.251295 | 80.940461 | 2419.592240 | 38.321247 | 42.035739 | 82.550188 | 5.93819 | 82.324084 | 1.742103 | 7483.158469 |
| std | 9.523867 | 124.292079 | 117.926501 | 4.052413 | 1987.914858 | 25.070016 | 11467.272489 | 20.044034 | 160.445548 | 23.428046 | 2.49832 | 23.716912 | 5.077785 | 14270.169342 |
| min | 36.300000 | 1.000000 | 0.000000 | 0.010000 | 0.000000 | 1.000000 | 0.000000 | 1.000000 | 0.000000 | 3.000000 | 0.37000 | 2.000000 | 0.100000 | 1.681350 |
| 25% | 63.100000 | 74.000000 | 0.000000 | 0.877500 | 4.685343 | 77.000000 | 0.000000 | 19.300000 | 0.000000 | 78.000000 | 4.26000 | 78.000000 | 0.100000 | 463.935626 |
| 50% | 72.100000 | 144.000000 | 3.000000 | 3.755000 | 64.912906 | 92.000000 | 17.000000 | 43.500000 | 4.000000 | 93.000000 | 5.75500 | 93.000000 | 0.100000 | 1766.947595 |
| 75% | 75.700000 | 228.000000 | 22.000000 | 7.702500 | 441.534144 | 97.000000 | 360.250000 | 56.200000 | 28.000000 | 97.000000 | 7.49250 | 97.000000 | 0.800000 | 5910.806335 |
| max | 89.000000 | 723.000000 | 1800.000000 | 17.870000 | 19479.911610 | 99.000000 | 212183.000000 | 87.300000 | 2500.000000 | 99.000000 | 17.60000 | 99.000000 | 50.600000 | 119172.741800 |

| Population | thinness 1-19 years | thinness 5-9 years | Income composition of resources | Schooling |
|---|---|---|---|---|
| 2.286000e+03 | 2904.000000 | 2904.000000 | 2771.000000 | 2775.000000 |
| 1.275338e+07 | 4.839704 | 4.870317 | 0.627551 | 11.992793 |
| 6.101210e+07 | 4.420195 | 4.508882 | 0.210904 | 3.358920 |
| 3.400000e+01 | 0.100000 | 0.100000 | 0.000000 | 0.000000 |
| 1.957932e+05 | 1.600000 | 1.500000 | 0.493000 | 10.100000 |
| 1.386542e+06 | 3.300000 | 3.300000 | 0.677000 | 12.300000 |
| 7.420359e+06 | 7.200000 | 7.200000 | 0.779000 | 14.300000 |
| 1.293859e+09 | 27.700000 | 28.600000 | 0.948000 | 20.700000 |

**Qualitative answer:**

The summary statistics of the dataset shows signs of right skewness for example in infant deaths, where the mean is much larger than the median. Zero values in schooling and GDP implicates data entry errors. This shows that cleaning, filling, and transformation of the dataset is necessary before implementing machine learning models.

### iii) Average Yearly Life Expectancy by Status:

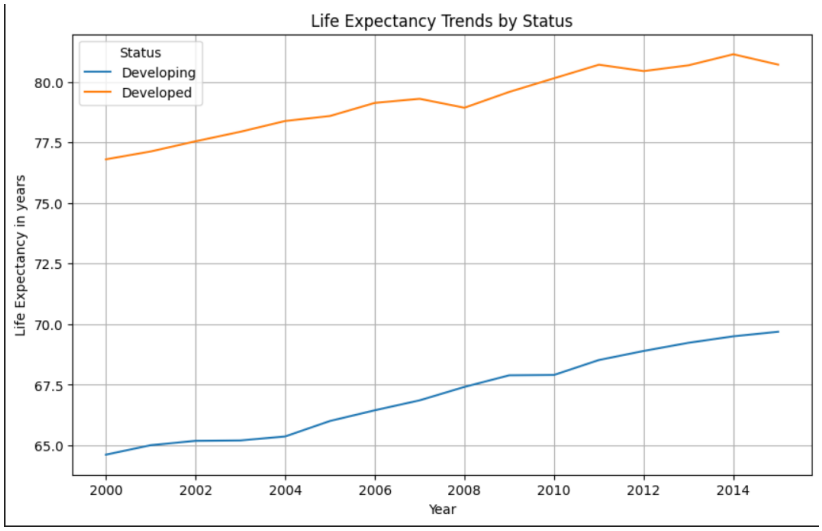
Average Yearly Life Expectancy by Status:

| Status | Developed | Developing |
|---|---|---|
| Year | | |
| 2000 | 76.803125 | 64.619868 |
| 2001 | 77.128125 | 65.009934 |
| 2002 | 77.546875 | 65.190728 |
| 2003 | 77.940625 | 65.206623 |
| 2004 | 78.384375 | 65.370861 |
| 2005 | 78.590625 | 66.009272 |
| 2006 | 79.131250 | 66.450331 |
| 2007 | 79.300000 | 66.860927 |
| 2008 | 78.931250 | 67.413907 |
| 2009 | 79.584375 | 67.894040 |
| 2010 | 80.146875 | 67.908609 |
| 2011 | 80.706250 | 68.523841 |
| 2012 | 80.443750 | 68.898013 |
| 2013 | 80.681250 | 69.234437 |
| 2014 | 81.137500 | 69.501987 |
| 2015 | 80.709375 | 69.690066 |

**Qualitative answer:**

Developed countries have higher life expectancy than developing countries on average from 2000 to 2015.

### iv) Visualization




**Qualitative answer:**

The line plot shows us two main trends in life expectancy based on development status.

The orange line representing developed shows us the huge gap in life expectancy, where the developed countries in 2015 had more life expectancy than developing countries by more than 8 years. The good news is that for both developed and developing countries, life expectancy is improving steadily from 2000 to 2015.

The anomaly here is how lines are moving in parallel signalling that the gap in life expectancy between developed and developing countries is not closing despite global development and health care systems improvement.

# Q2)

**How outliers influence regression model performance and bias results**

In short, the goal of regression models is to minimize the sum of squared errors. Because these errors are squared, a single outlier can cause a large error.

To minimize this single outlier, the model will tend to skew the entire line closer to that outlier.

This will cause the results to be biased, and the line will not accurately reflect the rest of the data.

**Steps and explanations**

The coding part of the question was done using python to identify and fill missing values using different imputation methods.

**Results:**

**i) Missing values proportion per variable**

```
Missing values proportion per Variable
Population                         22.191967
Hepatitis B                        18.822328
GDP                                15.248468
Alcohol                             6.603131
Income composition of resources     5.684139
Schooling                           5.547992
thinness  1-19 years                1.157250
thinness 5-9 years                  1.157250
BMI                                 1.157250
Total expenditure                   1.089176
Diphtheria                          0.646698
Polio                               0.646698
Life expectancy                     0.340368
Adult Mortality                     0.340368
infant deaths                       0.000000
Status                              0.000000
Country                             0.000000
Year                                0.000000
under-five deaths                   0.000000
Measles                             0.000000
percentage expenditure              0.000000
HIV/AIDS                            0.000000
dtype: float64
```

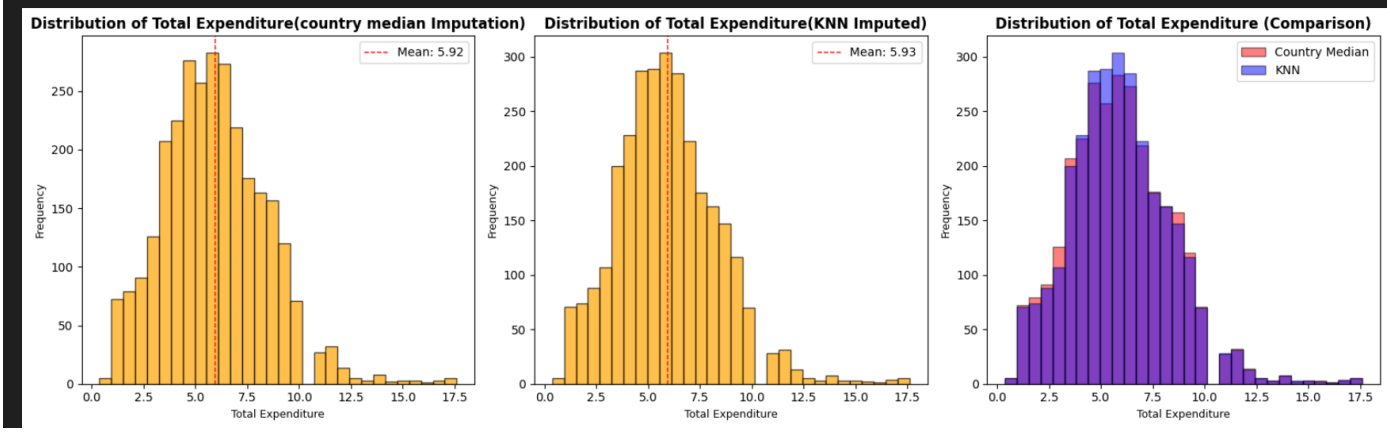
**ii) Different imputation methods for one variable (Total expenditure)**

```
 Total expenditure missing values after country median imputation:32
Now lets check missing values in Total expenditure after KNN imputation:0
```

When we applied the country median, the variable total expenditure still had 32 missing values, but after applying KNN imputation, all the missing values were gone.

### iii) Effect on variables mean and standard deviation, and data distribution

```
 The filled variable total expenditure mean and std:
mean    5.930115
std     2.440485
Name: Total expenditure, dtype: float64
```

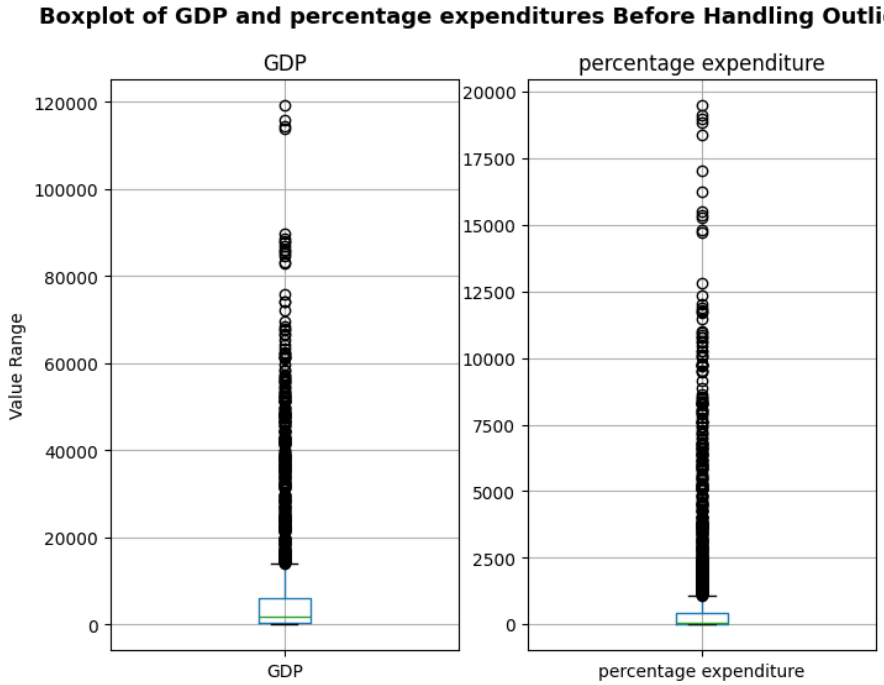



**Insight:**

The main finding is that the distribution of total expenditure for both imputation methods is very similar. This means that the KNN Imputer method found that the best way to fill in the missing values of Total expenditure is to look at other countries like the country with missing data.

Finally, it was found that the average expenditure of these similar countries is usually very close to the median value of the specific country's group.

### iv) Boxplots of GDP and percentage expenditures before Handling Outliers

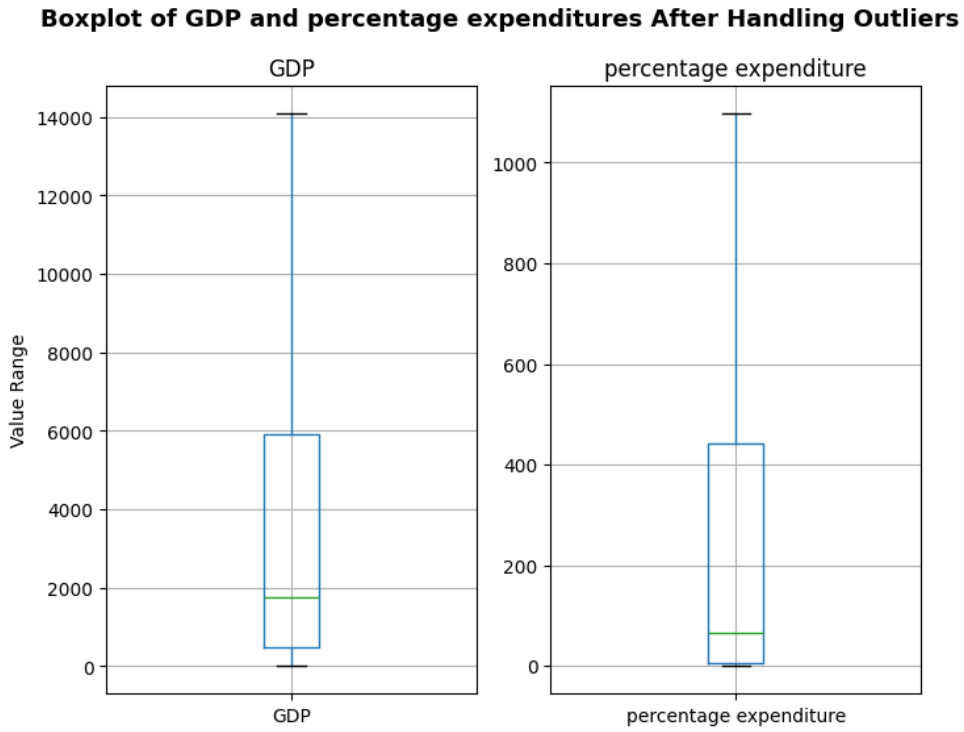


We can observe extreme outliers in both variables. We used the IQR to detect outliers instead of z-score because the IQR is not sensitive to outliers and doesn't get affected by skewed data unlike Z-score which assumes a normal distribution.

**IQR Detection method results:**

```
Detecting outliers using the IQR method (GDP and Percentage):
GDP                          12.42
percentage expenditure       13.24
dtype: float64
```

### v) Boxplot after handling outliers




**Qualitative answer**

We used the capping method because we didn't want to remove rows with outliers and losing important data in other columns. The process is also called Winsorizing which identifies extreme outliers and fills them with the 99th percentile value. This helps us to keep our information while ensuring the outliers to lose influence which can be a driving bias for the regression model.

# Q3)

**Steps/Explanations:**

The objective for this question was to do feature engineering using python and visualize correlation heatmaps and identify the strongest positive and negative correlations.

**Results**

**i) Engineered variables**

| infant_survival_rate | Vaccination_CI | Education_Income_Index |
|---|---|---|
| 945.0 | 59.833333 | 0.48379 |
| 945.0 | 60.666667 | 0.47600 |
| 945.0 | 63.333333 | 0.46530 |
| 945.0 | 67.000000 | 0.45374 |
| 945.0 | 68.000000 | 0.43130 |
| ... | ... | ... |
| 973.0 | 66.666667 | 0.37444 |
| 974.0 | 54.222222 | 0.39710 |
| 975.0 | 72.333333 | 0.42700 |
| 975.0 | 75.666667 | 0.41846 |
| 976.0 | 78.333333 | 0.42532 |

We created three new variables that represent theoretical relationships between existing variables.

**Infant Survival Rate (ISR)**: Calculated as $1000 - infant\ deaths\ (per\ 1000\ births)$

**Vaccination Coverage Index (VCI):** It is a composite value obtained by averaging the percentages of vaccination against Hepatitis B, Polio, and Diphtheria.

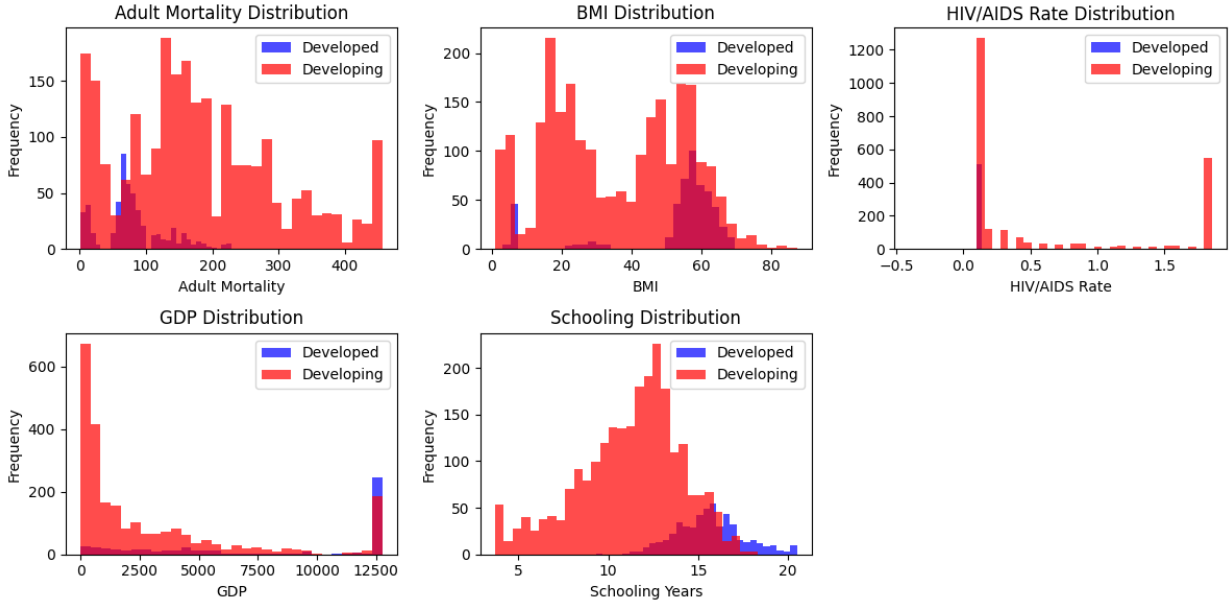
$$Vaccination\ Coverage\ Index = (Hepatitis\ B + Polio + Diphtheria) / 3$$

**Education Income Index (EII)**: It is a social and economic indicator obtained by multiplying 'Schooling' by 'Income composition of resources'. The resulting value can also be scaled and divided by 10.

$$Education\_Income\_Index = (Schooling * Income\ composition\ of\ resources) / 10$$

### ii) Five independent variables distribution which influence life expectancy and their distributions

Five independent variables believed to influence life expectancy were selected: Adult Mortality, BMI, HIV/AIDS death rate, GDP per capita, and Schooling. These variables show major factors of population health, including disease, nutrition, and socio-economic development.




**Qualitative answer:**

The HIV/AIDS shows a huge concern for developing countries as we see huge right skewness with high deaths associated with HIV.

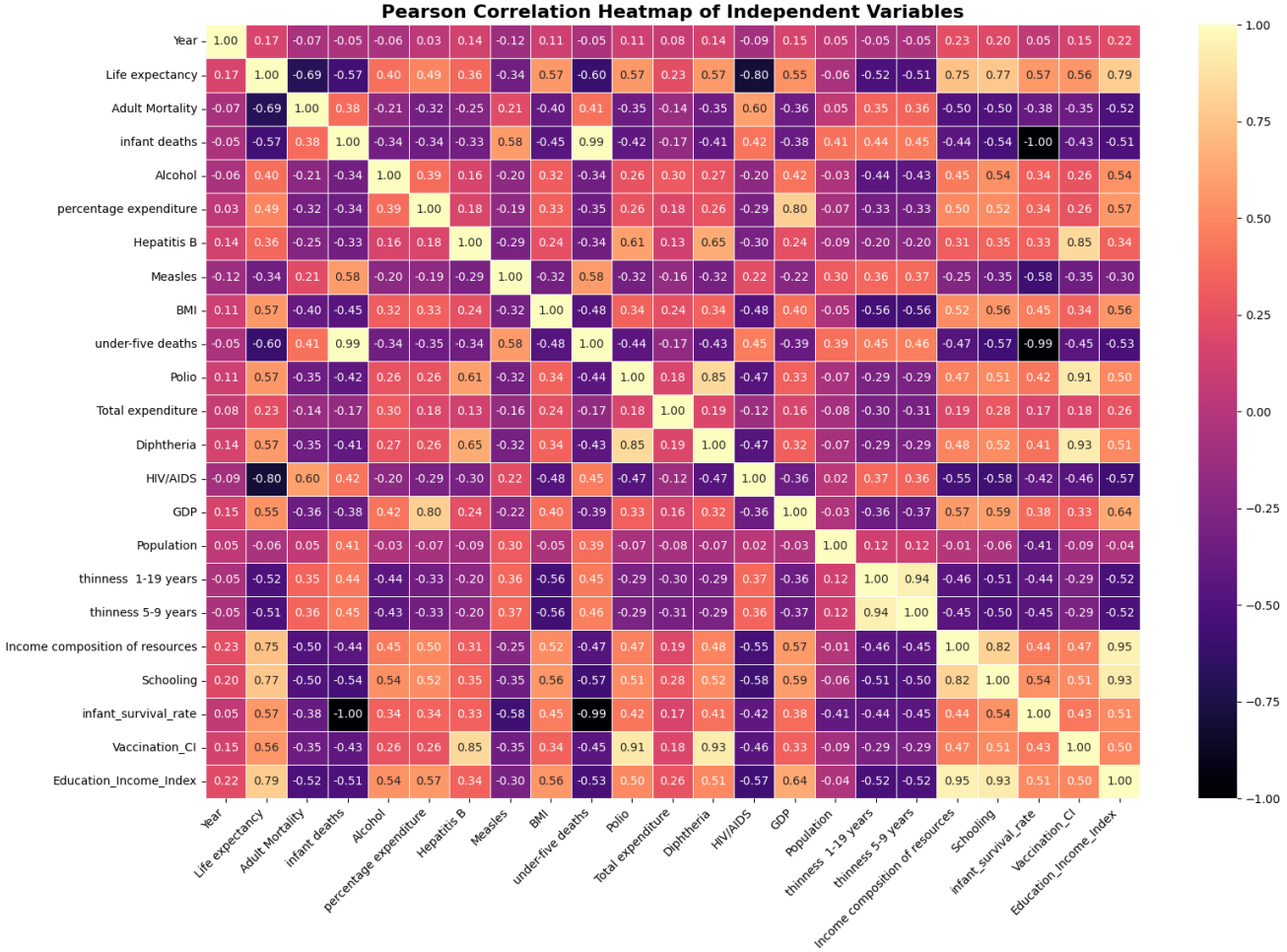
For Adult Mortality, we see developed countries having higher adult mortality, while developed countries are below 100 deaths per 1000, proving their higher life expectancy.

For BMI, we observe developing countries occupying both ends with large groups where the lower end is underweighted and higher end is overweight. Developed countries are centred around the overweight to obese range.

For GDP, developed countries have very high GDP values, while developing countries dominate the lower end of the GDP range.

For Schooling, developing countries lies in the range of 7 to 13 years while developed countries are at 13-18 years. This shows how education impacts health actions and life outcomes.

## iii) Pearson Correlation Heatmap of independent variables




```
strong negative to strong positive relationship with Life Expectancy

HIV/AIDS                         -0.796749
Adult Mortality                  -0.691415
under-five deaths                -0.604178
infant deaths                    -0.567221
thinness  1-19 years             -0.515711
thinness 5-9 years               -0.514572
Measles                          -0.337332
Population                       -0.059399
Year                              0.171275
Total expenditure                 0.233689
Hepatitis B                       0.364607
Alcohol                           0.398164
percentage expenditure            0.486695
GDP                               0.546499
Vaccination_CI                    0.555560
Polio                             0.565729
BMI                               0.565940
infant_survival_rate              0.567221
Diphtheria                        0.571361
Income composition of resources   0.746140
Schooling                         0.767355
Education_Income_Index            0.790854
```

**Qualitative answer:**

The correlation heatmap helps uncover different strong relationships with life expectancy. We observe HIV/AIDS with the strongest negative relationship with life expectancy, which shows that when HIV/AIDS rate, the life expectancy rate falls along with it. This is in line with global health pattern where regions with higher rates of AIDS, are associated with low number of years a person can live.

Additionally, we note strong positive relationships with life expectancy with income composition of resources and schooling which shows nations with higher income and good education system will tend to have higher life expectancy. Higher schooling often leads to greater economic and social opportunities, all of which are contributing factors to longevity.

**Why doesn't correlation imply causation?**

Correlation shows us how two variables move together, but it doesn't mean that one variable change another. For example, going to school 18 years doesn't mean you will live longer than a person who went to school for 5 years, but there is a lurking variable, meaning going to school only can't be the drive for life expectancy but they may be other factors such as the country's overall welfare and economy.

# Q4)

**Steps and explanations:**

The objective for the question was to encode categorical variables and scale different variables using standardScaler().

**Results**

**i) Encoding 'status' feature**

| Education_Income_Index | Status |
|---|---|
| 0.48379 | 0 |
| 0.47600 | 0 |
| 0.46530 | 0 |
| 0.45374 | 0 |
| 0.43130 | 0 |

We encode status column with 1 for developed and 0 for developing.

**ii) Scaling**

the scaled final dataset is:

| | Life expectancy | Adult Mortality | infant deaths | Alcohol | percentage expenditure | Hepatitis B | Measles | BMI | under-five deaths | Polio | ... | GDP | Population | thinness 1-19 years | thinness 5-9 years | Income composition of resources | Schooling | infant_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 65.0000 | 0.875325 | 2.165057 | -1.141908 | -0.546410 | -1.071729 | 1.886225 | -0.958171 | 2.065644 | -2.266576 | ... | -0.724442 | 2.118803 | 2.738728 | 2.721467 | -0.742785 | -0.587915 | |
| 1 | 59.9000 | 0.944562 | 2.165057 | -1.141908 | -0.540647 | -1.245845 | 0.730456 | -0.983188 | 2.065644 | -1.728904 | ... | -0.718012 | -0.733831 | 2.738728 | 2.721467 | -0.757695 | -0.618964 | |
| 2 | 59.9000 | 0.918598 | 2.165057 | -1.141908 | -0.541429 | -1.129767 | 0.555093 | -1.008205 | 2.065644 | -1.475881 | ... | -0.713706 | 2.118803 | 2.738728 | 2.721467 | -0.787514 | -0.650013 | |
| 3 | 59.5000 | 0.953216 | 2.165057 | -1.141908 | -0.528678 | -0.955651 | 1.886225 | -1.033221 | 2.065644 | -1.159603 | ... | -0.705066 | -0.233404 | 2.738728 | 2.721467 | -0.822304 | -0.681062 | |
| 4 | 59.2000 | 0.979180 | 2.165057 | -1.141908 | -0.711239 | -0.897612 | 1.886225 | -1.053235 | 2.065644 | -1.096348 | ... | -0.842167 | -0.340096 | 2.738728 | 2.721467 | -0.867033 | -0.774209 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2933 | 44.4875 | 2.549995 | 0.699524 | -0.049940 | -0.729465 | -0.897612 | -0.573453 | -0.557905 | 0.954102 | -1.159603 | ... | -0.753808 | 1.115260 | 1.176325 | 1.167529 | -1.100618 | -0.867356 | |
| 2934 | 44.5000 | 2.549995 | 0.647183 | -0.125248 | -0.729465 | -2.222831 | 1.886225 | -0.577918 | 0.914404 | -2.266576 | ... | -0.754037 | 1.093930 | 1.277125 | 1.292847 | -1.045949 | -0.774209 | |
| 2935 | 44.8000 | -0.769054 | 0.594843 | -0.032368 | -0.729465 | -0.607419 | 0.198710 | -0.597932 | 0.874706 | -0.780070 | ... | -0.843566 | -0.763841 | -0.890078 | -0.862615 | -1.001220 | -0.618964 | |
| 2936 | 45.3000 | 2.549995 | 0.594843 | -0.712652 | -0.729465 | -0.433302 | 0.835108 | -0.617945 | 0.835008 | -0.590303 | ... | -0.732506 | 1.054166 | -0.789278 | -0.762361 | -1.001220 | -0.681062 | |
| 2937 | 46.0000 | 2.549995 | 0.542502 | -0.722693 | -0.729465 | -0.259186 | 1.886225 | -0.637958 | 0.835008 | -0.463792 | ... | -0.732784 | 1.032792 | 1.579526 | 1.618673 | -0.966431 | -0.681062 | |

**Qualitative answer:**

We scaled using standardScaler to all numerical independent variables to scale them so that the model doesn't think variables such GDP, population with large numerical ranges are more important than other variables with small ranges. The target variable 'Life Expectancy' was excluded.

**Why models such as KNN or Gradient Boosting are sensitive to scale differences?**

**KNN** needs scaling because K-NN is measures distance between points to identify nearest neighbour. Without scaling, variables with large numbers will have higher importance and those with small numbers will be ignored.[3]

**Gradient Boosting** is less sensitive as it does not depend on distance for decision but splits points to make true or false decisions.

Scaling won't change the decisions or the tree structure as the model will continue to find best split that offers the most information.[3]

# Q5)

### Steps and Explanations

We used to datasets, one which is scaled to use for KNN and gradient boosting regressor and another which is not scaled for decision trees and random forest regressor. We used python sci-kit learn to import the different models.

### Results

#### i) Model's performance

Model Performance Summary

| | Model | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| 3 | Gradient Boosting | 1.812858 | 1.268163 | 0.962046 |
| 1 | Random Forest | 2.056757 | 1.461435 | 0.951147 |
| 2 | KNN | 2.151748 | 1.395307 | 0.946530 |
| 0 | Decision Tree | 2.925573 | 1.921945 | 0.901157 |

### Qualitative answer:

The gradient boosting model is the best performer among the three. It has high $R^2$ score of 0.962, which means it can describe 96.2% of the variance in life expectancy. The model shows the lowest error rates with mean average error of 1.27, which indicates that the model's predictions are on average, only off by 1.27 years.

The table also shows us the use of ensemble models (GB and Random Forest) which outperforms the single decision tree model. The ensemble models have good metrics because they are engineered to overcome overfitting and underfitting.

#### ii) The best parameters for each model

```
Best Decision Tree params:{'criterion': 'friedman_mse', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 6}
Best Random Forest params:{'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 3, 'min_samples_split': 5, 'n_estimators': 120}
Best KNN params:{'metric': 'manhattan', 'n_neighbors': 4, 'weights': 'distance'}
Best Gradient Boosting params:{'learning_rate': 0.1, 'max_depth': 4, 'min_samples_split': 2, 'n_estimators': 300, 'subsample': 0.9}
```

The above are the best hyperparameters for each model.

# Q6)

**Steps:**

The main goal for this question was to find the feature importance from the best model and retrain the models using the selected features and best hyperparameters.

**Results**

**i) Top 5 Feature importance from the best performing model**

```
HIV/AIDS                          0.457900
Income composition of resources   0.270812
Adult Mortality                   0.150832
Education_Income_Index            0.015171
under-five deaths                 0.013915
```

Gradient Boosting was the best-performing model in Q5 and was therefore used to extract top 5 feature importance.

We observe the dominant feature being HIV/AIDS, which the model considers as the most important predictor for life expectancy. The result supports our heatmap in Q3 which also showed how HIV/AIDS had the strongest negative correlation.

The top 3 features HIV/AIDS, Income composition of resources, and adult mortality accounts for over 88% of the model's predictive power.

**ii) Trained models**

```
Retraining Models with Best Hyperparameters on Selected Features

1. Decision Tree Regressor
2. Random Forest Regressor
3. KNN Regressor
4. Gradient Boosting Regressor
```

| | Model | RMSE (Full Features) | RMSE (Top Features) | MAE (Full Features) | MAE (Top Features) | $R^2$ (Full Features) | $R^2$ (Top Features) |
|---|---|---|---|---|---|---|---|
| 2 | KNN | 2.151748 | 1.951259 | 1.395307 | 1.185170 | 0.946530 | 0.956030 |
| 3 | Gradient Boosting | 1.812858 | 1.960053 | 1.268163 | 1.400647 | 0.962046 | 0.955633 |
| 1 | Random Forest | 2.056757 | 2.158100 | 1.461435 | 1.556851 | 0.951147 | 0.946214 |
| 0 | Decision Tree | 2.925573 | 2.838493 | 1.921945 | 1.915427 | 0.901157 | 0.906953 |

**Qualitative answer**

We retrained the 4 models using the top 5 features and best hyperparameters below

```
Best Decision Tree params:{'criterion': 'friedman_mse', 'max_depth': 12, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 6}
Best Random Forest params:{'max_depth': 8, 'max_features': 'sqrt', 'min_samples_leaf': 3, 'min_samples_split': 5, 'n_estimators': 120}
Best KNN params:{'metric': 'manhattan', 'n_neighbors': 4, 'weights': 'distance'}
Best Gradient Boosting params:{'learning_rate': 0.1, 'max_depth': 4, 'min_samples_split': 2, 'n_estimators': 300, 'subsample': 0.9}
```

KNN became the best model as it $R^2$ score moved up from 0.94 to 0.956 which is expected as KNN works best with low dimensions as dimensionality reduction helps this distance-based models by removing irrelevant variables.

Gradient Boosting and Random Forest models lost some performance because as dimensions decreased the trees could not use minor features to find small, complex interactions but they were still performing better than decision tree.

**Discuss the trade-offs between model simplicity, interpretability, and predictive accuracy**

**Predictive accuracy:** High accurate models which are complex most of the times, such as ensemble methods. The power of these models comes from their ability to capture and explain non-linear dynamics and the high-level interactions between the many factors included in the data [4]. However, due to their complexity, these models are often called black boxes because even their developers cannot fully track the exact reasons for how they arrived at each prediction or result [4].

**Simplicity and interpretability:** Simpler models such as linear regression and single decision trees are easy interpretable and easy to track coefficients for each variable. However, as a model becomes too simple, performance reduces due to high bias.[4]

**Trade-off:** Complex models can have high accuracy, but they are black boxes making it hard to know how it arrived at a certain prediction. Simple models have high interpretability but can have low accuracy.[4] However, simple models accuracy can be improved by applying model ensembling, where different models with different complexities are mixed to create a more accurate and interpretable model.[4]

# Q7)

**Steps/Procedure:**

For this question, the main objective was to check for differences between model-derived feature importance which we did in Q6 and correlation from Q3. Using python, we extracted five best important features from the best performing model.

**Results:**

i) **Differences between model-derived feature importance which we did in Q6 and correlation from Q3**

```
 correlation from Q3

Life expectancy                  1.000000
Education_Income_Index           0.790854
Schooling                        0.767355
Income composition of resources  0.746140
Diphtheria                       0.571361
infant_survival_rate             0.567221
BMI                              0.565940
Polio                            0.565729
Vaccination_CI                   0.555560
GDP                              0.546499
percentage expenditure           0.486695
Status                           0.481760
Alcohol                          0.398164
Hepatitis B                      0.364607
Total expenditure                0.233689
Population                      -0.059399
Measles                         -0.337332
thinness 5-9 years              -0.514572
thinness  1-19 years            -0.515711
infant deaths                   -0.567221
under-five deaths               -0.604178
Adult Mortality                 -0.691415
HIV/AIDS                        -0.796749
Name: Life expectancy, dtype: float64

 The the top five most important features from the best-performing model

HIV/AIDS                         0.457900
Income composition of resources  0.270812
Adult Mortality                  0.150832
Education_Income_Index           0.015171
under-five deaths                0.013915
dtype: float64
```

**Qualitative answer:**

From **Q3 correlation** results, Education_Income_Index (approx. 0.80), schooling (approx. 0.77), and Income composition of resources (0.74) were the best predictors.

From the Q6 model feature importance results, Income composition of resources (27.1%) was more important than Education_Income_Index (1.5%), Schooling was almost useless for the model.

**Why the difference?**

Correlation measures direct linear relationships between a feature and the target. It looks at one feature and compares it to the target while ignoring other features. For example, it appears that schooling is correlated with life expectancy.

However, model importance in Q6 looks for feature importance in making predictions, given all the features it has.

The key insight is that the model was good at identifying multicollinearity because features like schooling, income composition, and education were highly correlated with each other; therefore, they don't provide new information to the model. The model treats these as redundant and will only need one for predictions.

**Top five most important features from the best-performing model and discuss their real-world relevance in explaining differences in life expectancy**

```
HIV/AIDS                          0.457900
Income composition of resources   0.270812
Adult Mortality                   0.150832
Education_Income_Index            0.015171
under-five deaths                 0.013915
```

1. **HIV/AIDS**:

HIV/AIDS was the dominant feature with 45.8% of the model's predictive power.

**Real-world relevance**: nations with higher rates of HIV/AIDS tend to have significantly lower life expectancy. Our dataset covering from 2000-2015 which is the period before antiretroviral drugs global availability. The model learned that high prevalence of HIV/AIDS in a nation led to high increase of adult mortality and finally, led to low life expectancy in that nation. The model learned that this was not only a simple correlation but a primary driver of death.

2. **Income composition of resources**

27.1% importance made income composition of resources the second most important feature for our model

**Real-world relevance:**

This is where correlation doesn't imply causation because at first glance you may think that income itself can make you live longer. The model learned that better income composition of resources led to better healthcare, better nutrition.

3. **Adult Mortality**

15.1% importance ranked adult mortality as the third important feature for our model.

**Real-world relevance:**

High adult mortality is associated with health challenges, violence, or lack of health infrastructures. Nations with higher adult mortality lose citizens early due to the above different challenges.

4. **Education_Income_Index**

The model ranked this as unimportant with 1.5% importance.

**Real-world relevance:**

The model assigned low score to education_income_index because it's contribution laps with Income composition of resources, which was already strong predictor, and this made the feature not of importance to the model as it didn't provide any new information. This was mainly due to multicollinearity, where the model only used the more informative feature which is income composition. However, this does not mean education is not important in reality, only that it did not provide new information compared to what income already provided.

5. **Under-five deaths**

The model also rated this feature as not important with 1.4% importance.

**Real-world relevance:**

The model also learned that that HIV/AIDS and adult mortality were the strongest predictors for the overall lifespan than child mortality alone. The model learned that adult mortality was the most important feature to show the country's healthcare problems.

**References**


[1] Devansh, "Why you should analyze the distribution of your Data," Analytics Vidhya. Accessed: Nov. 15, 2025. [Online]. Available: https://machine-learning-made-simple.medium.com/why-you-should-analyze-the-distribution-of-your-data-695fd9f0f1be

[2] "Data Leakage in Time Series Data Cross-Validations in Machine Learning – Hectorv.com." Accessed: Nov. 15, 2025. [Online]. Available: https://hectorv.com/2023/07/06/data-leakage-in-time-series-data-cross-validations-in-machine-learning/

[3] G. Jha, "Feature Scaling in Machine Learning: Which Popular Algorithms Require It and Which Don't?," Medium. Accessed: Nov. 15, 2025. [Online]. Available: https://medium.com/@post.gourang/feature-scaling-in-machine-learning-which-popular-algorithms-require-it-and-which-dont-a71f5585d664

[4] G. Andersen, "Balancing Accuracy and Interpretability Trade-offs in Model Complexity for Machine Learning." Accessed: Nov. 16, 2025. [Online]. Available: https://moldstud.com/articles/p-balancing-accuracy-and-interpretability-trade-offs-in-model-complexity-for-machine-learning