UNIVERSITI MALAYA

WQD7005 DATA MINING

PROJECT DEVELOPMENT REPORT
SEMESTER 2, SESSION 2023/2024

Predicting Life Expectancy Using Machine Learning

Group: G10

Teacher: Associate Prof. Dr. Nor Liyana Mohd Shuib

Name	Matric Number
Lai Yuen Seng	22076493
Justin Sem Ee Qing	s2111464
Dennis Leong Lip San	17217825
Liang Ruijie	22100508

I. Introduction

a. Background

Life expectancy (LE) serves as a metric that estimates the average lifespan of individuals or a community, typically adjusted for age. It is a crucial tool in assessing mortality rates and is utilized as a key social indicator by policymakers, as well as the insurance and financial sectors, to develop appropriate financial and insurance plans (Raphael et al., 2023). The calculation of LE is not solely dependent on birth and death rates but also takes into consideration a range of factors such as gender, lifestyle choices, healthcare infrastructure, and more. Consequently, research on life expectancy can yield diverse results due to the varying factors considered. For example, a study comparing the remaining LE at age 50 between individuals with 5 healthy lifestyle factors versus those with none revealed a discrepancy of 12.2 years for men and 14.0 years for women. These findings are significant as they play a crucial role in shaping policy decisions.

b. Rational

According to recent research (Welsh et al., 2021), an increase in life expectancy among older adults living in the community is associated with a higher risk of experiencing frailty-related incidents. This in turn creates a positive and mutually reinforcing connection between the severity of such incidents and the subsequent need for healthcare services and hospitalizations (Ofori-Asenso et al., 2019). Consequently, this leads to a rise in costs for government agencies, necessitating a greater demand for operational health services (Ilinca & Calciolari, 2015; Segal et al., 2017).

In the year 2019, the global spread of the COVID-19 pandemic had a significant impact on life expectancy worldwide (Adair et al., 2023). This observation is further supported by a study (Schöley et al., 2022), which found that most Western countries experienced a decline in life expectancy during the pandemic. However, as restrictions were lifted and vaccination rates increased, life expectancy returned to normal levels. This highlights the crucial role of global public health in assessing life expectancy. Nevertheless, it is important to note that an increase in healthcare expenditure does not necessarily guarantee a higher life expectancy, as indicated by the findings of authors (Ketenci & Murthy, 2018). Therefore, it becomes crucial to identify the determinant factors that directly or indirectly influence life expectancy. This knowledge can help countries focus on areas that require improvement or prioritization (Wirayuda & Chan, 2021).

In the past, life expectancy was rather deterministic, based on mathematical formulae or expert judgment (Booth & Tickle, 2008) and slowly emerging using stochastic methods to forecast mortality (Basellini et al., 2023). It is worth mentioning that in 1992, a parsimonious model was introduced by Ronald D. Lee and Lawrence R. Carter (Lee & Carter, 1992) to predict age-time-specific death rates in the United States, which is also known as the Lee-Carter (LC) method. Later on, in recent decades, with the advancement of technology and computational power available, a rather sophisticated approach, namely machine learning, has become popular for predicting life expectancy. In the present study, we aim to identify the determinant factors of LE by using a machine learning model to predict LE.

c. Problem statement

LE has been extensively examined from various perspectives including traditional methods such as the Lee-Carter model and machine learning way. However, the traditional methods struggle to account for other than the factor of mortality rate as compared to the machine learning way. According to a study conducted by (Wirayuda & Chan, 2021), sociodemographic, macroeconomic, and health resources factors exhibit a strong correlation with life expectancy. Although machine learning capable to these factor in predicting LE, but they are often investigated individually for specific purposes, such as analyzing the healthcare system (Beckwith et al., 2023) that only utilized specific health resource factor. Typically, in this type of study, researchers only utilized specific types of variables and targeted specific audiences to support their research. Example, (Kouame Amos B., 2022) and (Vydehi, 2020) did not take into account the interaction of variables are in their linear model.

Even though some researchers attempted to incorporate all of these three factors, they have utilized complex machine learning models like XGboost (Lipesa et al., 2023), Naïve Bayes (Raftery et al., 2013), Random Forest (RF) (Meshram, 2020), and Artificial Intelligence (AI) (Lesnussa et al., 2020). However, developing complex predictive models might be challenging to interpret and implement in real-world settings although it may provide high predictive accuracy. For example, the author (Raphael et al., 2023) achieved high accuracy in their model however a significant amount of missing data was removed through exploration resulting in 44% data loss. Therefore, it is crucial to develop predictive models that strike a balance between complexity and interpretability. This creates a gap that calls for a machine learning approach to bridge the interpretability gap without significantly compromising data and accuracy.

d. Research objectives

The research outlined in this proposal is driven by a dual emphasis on identifying the determinants of life expectancy and assessing the efficacy of machine learning algorithms in predicting these outcomes on a global scale. The project's research questions aim to uncover the critical factors influencing life expectancy across diverse demographics and regions and to evaluate how effectively machine learning tools can forecast life expectancy worldwide.

i. Explore Significant Variables

The project will conduct exploratory data analysis (EDA) and employ statistical techniques to identify the key factors influencing life expectancy. This will help in understanding the complex interplay of socio-demographic, economic, and health-related factors.

ii. Develop Predictive Models

Various machine learning models, including simpler regression models and more complex ones like Random Forest and Gradient Boosting, will be tested and evaluated. This approach aims to balance predictive accuracy with model interpretability, which has been a gap in previous studies.

iii. Comprehensive Data Handling

An improvement over previous methods will be a more sophisticated treatment of missing data, such as regression-based imputation, and extensive data exploration to mitigate the risk of significant data loss which has marred earlier studies.

e. Scope

This study employs data from 193 countries, spanning the years 2000 to 2015, to provide a comprehensive geographic overview. However, it's important to recognize that the results may not extend to periods beyond this timeframe or accurately represent specific sub-national regions that might be underrepresented in the dataset. This geographic limitation could affect the generalizability of the study's findings to other contexts or more recent years.

The research also focuses on a diverse set of variables, including socio-economic and health-related factors, to determine their impact on life expectancy. Despite this broad approach, the study may not capture all possible influences, such as genetic predispositions or unique local environmental conditions, which could also play significant roles in determining life expectancy. This selective focus on certain variables may limit the study’s ability to account for the full complexity of factors influencing life expectancy.

Moreover, the project will explore various machine learning models with an emphasis on achieving a balance between predictive accuracy and model interpretability. While this approach is advantageous for understanding and applying findings, it might restrict the exploration of newer or unconventional modelling techniques that could potentially provide deeper or alternative insights. This focus on traditional and interpretable models may therefore limit the study's ability to harness the full potential of more innovative or complex machine learning strategies.

II. Literature Review

Reference	Approach	Technique	Performance	Factors	Limitations
Lipesa et al. (2023)	The study developed supervised machine learning regression models to predict life expectancy based on health, socioeconomic, and behavioral characteristics.	The study employed several regression models such as: • eXtreme Gradient Boosting (XGBoost) • Random Forest • Artificial Neural Network	The XGBoost model outperformed the Random Forest and Artificial Neural Network models in the sense of: • MAE: 1.554 • RMSE: 2.402	Up to 8 important factors have been identified using PCA. This includes: • Region • Income Group • Skin thinness • Income	The study did not integrate other quality of life measures and environmental components in the prediction model. Exploration of other regression models can be done.
Pisal et al. (2022)	The study used machine learning algorithms to classify life expectancy into discrete age groups for Asian countries.	The study employed decision tree classification techniques such as: • J48 • Random Tree • Random Forest	Random Forest was found to be the best technique among the three, with the following metric under 20-fold cross-validation: • Accuracy: 87.62% • RMSE: 0.193 • ROC Area: 0.974	he study identified 11 attributes that have an influence on life expectancy using WEKA's Wrapper evaluator and 'Best-First' method. The attributes include: • Income Group • Infant Deaths • Schooling attributes	Further research could be done to explore the correlation between attributes and life expectancy. Other classification algorithms can be explored to compare the results. Generalisation of the study to other continents can be considered.
Lakshmanarao et al. (2022)	The authors analyzed various features, including immunization and Human Development Index (HDI) factors, to predict life expectancy using machine learning techniques.	Four machine learning algorithms were applied for life expectancy prediction namely: • Multiple Linear Regression • Decision Tree Regression • Support Vector Regression • Random Forest Regression	Random Forest Regression achieved the best performance with metrics: • R^2 : 0.96 • RMSE : 1.928	Identified factors affecting life expectancy including immunization features, HDI factors, and other health-related factors through EDA and regression analysis. Through Multiple Linear Regression, the authors noted that features influencing life expectancy may differ by country.	Features affecting life expectancy vary for different countries, making it difficult to generalize. Exploration of other regression models can be done.
Raphael et al.(2023)	The authors used a cause-to-effect approach, identifying significant factors affecting life expectancy before applying machine learning for prediction.	Statistical techniques such as measures and tests were used to identify significant factors. Four machine learning algorithms were used to model life expectancy, namely: • CART • Random Forest • Extra Trees • XGBoos	Extra Trees regression model performed best for predicting life expectancy, with the following metrics: • MAE: 1.0234 • RMSE: 1.7746 • R^2: 0.9480	Key factors identified via EDA and correlation analysis include: • GDP • Immunization factors • Demographic factors • Education level	Data pre-processing led to a 44% data loss, potentially causing under-representation of some countries. Exploration of other regression algorithms can be done.

Vydehi et al. (2020)	The authors performed two major tasks. The first task is to predict life expectancy using regression models. The second task is to classify age group using classification models	Regression models: • Multiple Linear Regression • Decision Tree Regression • Random Forest Regression Classification Models: • k-Nearest Neighbour • Decision Tree Classifier • Random Forest Classifier	For regression, Random Forest Regression achieved the best performance with: • R^2: 0.958 • MAE: 1.206 • MSE: 3.741 • RMSE: 1.9341 In classification, Random Forest performed better in precision, recall, f1-score and accuracy in general	The authors identified several features which mostly affects the life expectancy using regression analysis. Notable features include 'Income composition of resources'.	The authors dropped certain factors, such as 'Hepatitis B', 'Year' etc without explanation, such action might remove important information from the dataset.
Kouame Amos and Smirnov (2023)	The authors performed correlation analyses to identify significant variables, and developed several regression models to predict life expectancy	The regression model used are: • Multiple Linear Regression • Lasso Regression • Ridge Regression AIC, BIC and Mallows' Cp criterion are used to the fit of models and to determine influencing factors	The Multiple Linear Regression performed the best, with the lowest AIC value and the following metrics: • Adjusted R^2: 0.85 • RMSE 3.85	The authors identified several influencing factors, such as: • Mortality • Expenditure • Diseases • Education level	Other regression models can be explored to compare the results.

Table 1: Literature Review

III. Dataset and Proposed Method

a. Dataset

The life expectancy dataset is found and selected from the Kaggle data repository, which is based on the Global Health Observatory data repository under the World Health Organization (WHO) where they keep track of the health status of the residents and the related factors for countries around the globe. The datasets are publicly available for health data analysis and can be accessed at https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who.

This life expectancy dataset is based on 193 UN member states from the year of 2000 to 2015. Other than life expectancy health-related factors, the UN data repository provided the corresponding socio-economic-related factors as well for the 193 countries. The individual data files were merged into a single dataset which consists of 22 columns of features and 2938 rows of data, including variables categorised into immunization-related, mortality-related, economic and social factors.

Detailed features are listed in Table 2 below.

Feature	Description	Data Type
Country	Name of the country	string
Year	Year of the data recorded	int
Status	Developing or Developed country	string
Life Expectancy	Age, on average, a newborn can expect to live based on the current death rates	decimal
Adult Mortality	Number of people dying between the age of 15-60 years per 1000 population	int
Infant Deaths	Number of infant deaths per 1000 population	int
Alcohol	Recorded per capita alcohol consumption in litres	decimal
Percentage Expenditure	Country's percentage expenditure on health per GDP	decimal
Hepatitis B	Percentage Hepatitis B immunization coverage among one year old	decimal
Measles	Number of measles reported cases per 1000 population	int
BMI	Average body mass index of the entire population	decimal
Under-five Deaths	Number of under-five deaths per 1000 population	int
Polio	Percentage of polio immunization coverage among one year old	decimal

Total Expenditure	Percentage of total government expenditure	decimal
Diphtheria	Percentage of diphtheria immunization coverage among one year old	decimal
HIV/AIDS	Ratio of HIV/AIDS deaths per 1000 population	decimal
GDP	Gross domestic product per capita in dollars	decimal
Population	Country's population	int
Thinness 10-19 years	Percent of thinness among children from age 10-19	decimal
Thinness 5-9 years	Percent of thinness among children from age 5-9	decimal
Income Composition of Resources	Human development index in terms of income composition of resources	decimal
Schooling	Number of years spent in school	decimal

Table 2: Description of Dataset

This dataset is crucial for understanding global health trends and disparities. It helps develop predictive models for life expectancy, essential for public health planning and policy-making. By evaluating the impact of socio-economic factors on health outcomes, it supports comparative studies of health systems and healthcare interventions. Suitable for advanced machine learning algorithms, this dataset enhances prediction accuracy and aids sophisticated health analytics.

b. Proposed Method

The proposed workflow for this project is based on SEMMA, a data mining methodology with a structured approach. SEMMA stands for Sample, Explore, Modify, Model and Assess, which are the sequential steps in the data mining process, and is designed to extract valuable information and insights from the datasets, enabling organisations to make the right decisions. The figure below outlines the proposed workflow of the project based on the SEMMA data mining process.

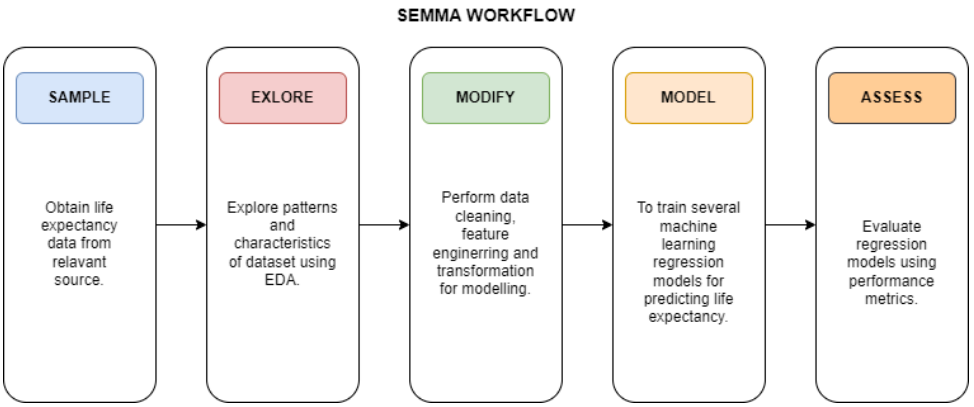

Figure 1: SEMMA Methodology

i. SEMMA

The first phase in SEMMA methodology is Sample, where sampling is performed to select a subset of data from the entire dataset, which helps in reducing the data volume for easier management and increases the efficiency for analysing without losing the significant characteristics of the original dataset. The dataset is sampled by the dataset owner from the year 2000 to 2015 for 193 countries as there has been better development in the health sector which results in an improvement of the human mortality rates in the past 30 years. We did not further perform sampling in this project due to the volume of the dataset and it is easily accessible and manageable and will provide complete accuracy by avoiding any potential sampling bias. After the Sample, is the Explore phase, where we will be performing comprehensive data exploration to identify the underlying pattern, unanticipated relationships and anomalies in the dataset. With the help of Exploratory Data Analysis (EDA) techniques, can help achieve a better understanding of the data, and choose the right manner to handle and process the dataset. The next phase is Modify, an important stage with the objective of preprocessing and transforming the dataset for modelling later. It ensures the dataset is clean, relevant, low noise and properly formatted, which can lead to an accurate and reliable analysis. The dataset from Kaggle consists of several missing values, especially in attributes such as population, hepatitis B and GDP. A more sophisticated treatment of missing values will be employed, such as regression-based imputation. Other techniques such as scaling, transformation, feature engineering and feature selection will be performed to prepare the data for modelling. The fourth stage is the Model, which is the stage of developing predictive or descriptive models to either identify patterns or make a prediction. The goal of this project is to develop a life expectancy prediction model, several regression models will be explored such as Linear and Ridge Regression, K-Nearest Neighbour Regression, Random Forest Regression, and XGBoost Regression. The last phase in the

SEMMA methodology is the Assess. The performance and validity of the developed models will be evaluated to ensure that the models are not only accurate but also reliable and able to generalise to unseen data. Evaluation metrics such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-squared will be utilised.

ii. Linear Regression

Linear Regression is a statistical technique for estimating the relationship between a dependent variable and one or more independent variables. Regression models with one dependent variable and more than one independent variable are referred to as multiple linear regression, which is the model we are utilising for our dataset. People assume that the linear relationship between the dependent and independent variables is a hyperplane, which can be represented by the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where:

Y is the response variable

$X_1, X_2, \ldots, X_n$ are the exploratory variables

β_0 is the intercept

$\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients

iii. Ridge Regression

Ridge Regression first introduced in (Hoerl, A. E., 1970), is a type of regression also known as L2 regularization. It is a popular parameter estimation method to help address the collinearity problem in multiple linear regression. It includes a regularization term to prevent overfitting by adding a penalty term in the cost function used in linear regression, which is the sum of the squared of the weights. L2 can shrink the weights but does not eliminate them entirely.

iv. K-Nearest Neighbour Regression

K-nearest neighbour regression is a non-parametric and instance-based learning algorithm. It predicts the value of a dependent variable based on the values of its nearest neighbours in the feature space. Generally, it make predictions for a new data point by comparing them to the stored instances and the similarity between instances is typically measured using a distance metric such as Manhattan or Euclidean distance. The parameter k in KNN is significant as it represents the number of nearest neighbours and it can greatly affect the model's performance. Typically a smaller k can produce a more sensitive model that may cause an overfitting issue, whereas a larger k may smooth out the predictions and underfit the data.

v. Random Forest Regression

Random Forest Regression (Breiman, L., 2001) is another famous machine-learning technique built on multiple decision trees during training and outputs the average prediction of the individual tree. Through the process known as bagging, each decision tree is trained independently on a subset of the data and makes its own prediction, resulting in a diverse set of trees. In regression, the final predictions of all the trees are averaged to obtain the final prediction which reduces the variance and improves the prediction accuracy.

vi. XGBoost Regression

Extreme gradient boosting or XGBoost regression is an advanced algorithm in machine learning which is part of the gradient boosting framework developed in (Chen, T., 2016). It utilises decision trees as base learners and employs regularization techniques to enhance the model's performance. Generally, it adds weak learners to the ensemble of decision trees in a sequential manner and each tree will be focusing on correcting the errors of the previous trees, where the algorithm minimizes the cost function using gradient descent. Either the lasso or ridge regularization technique can be adapted to prevent overfitting and enhance the generalization of the model. By combining the weak learners, the final strong learner can bring down both the bias and the variance.

c. Evaluation Metrics

i. Mean Square Error (MSE)

Mean Square Error is one of the common evaluation metrics to measure the accuracy of regression models. It can quantify the difference between the actual values and predicted values by calculating the mean of the squared differences. Due to the characteristic of the square, MSE will penalize large errors even more as compared to the small errors. The formula for Mean Squared Error is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2$$

where:

n is the total number of observations

y_i is the actual value

y'_i is the predicted value

ii. Root Mean Square Error (RMSE)

Root Mean Square Error is the square root of the mean of the squared differences between the predicted and actual values, therefore is the square root of the Mean Square Error. RMSE allowed us to penalize large errors and the capability to interpret the error metric easily.

$$\mathrm{R}MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2}$$

where:

n is the total number of observations

y_i is the actual value

y'_i is the predicted value

iii. Mean Absolute Error (MAE)

Mean Absolute Error is another popular regression evaluation metric. While MSE and RMSE perform squaring on the error, MAE has a different approach by just calculating the average absolute differences between the actual values and predicted values. This approach makes MAE less sensitive to outliers and increases the interpretability.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y'_i\right|^2$$

where:

n is the total number of observations

y_i is the actual value

y'_i is the predicted value

iv. R-squared

R-squared, also known as the coefficient of determination, is a metric which allows us to evaluate the fit of the regression models. It indicates how well the independent variables explain the variability of the dependent variable. It provides a measure of goodness-of-fit, showing the proportion of the variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where:

SS_{res} (Residual Sum of Squares) :

is the sum of the squared differences between the actual and predicted values

SS_{tot} (Total Sum of Squares) :

is the sum of the squared differences between the actual and the mean of the actual values

d. Data Mining Tools

Throughout the project development, from preprocessing to evaluation, was conducted using Jupyter Notebook in the Google Colaboratory environment. Python and its libraries were the primary tools used in this data mining project.

IV. Results and Discussion

1. **S**ample

In the first phase of SEMMA which is Sample, the dataset we collated from the Kaggle repository is sampled by the dataset owner from the year 2000 to 2015 for 193 countries due to that in the past 15 years, there has been a great leap of the development in the health sector which brings a positive impact to the human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project, we did not further perform sampling due to the volume of the dataset being easily accessible and manageable. This can avoid any potential sampling bias and allow the regression models to return to more reliable and well-grounded results.

2. **E**xplore

In the **E**xploration phase, several graphical and numerical summaries have been employed to gain insights from the data.

- *On Categorical Variables*

Excluding the variable ‘Country’, the other categorical variable is ‘Status’, which classifies whether a country is under developing or developed. There are 193 countries in total, of which 32 are developed countries and the rest are developing countries.

Based on Figure A4 in the Appendix, it can be observed that developed countries tend to have a longer life expectancy as compared to developing countries in general.

- *On Continuous Variables*

From the KDE Plots and Boxplots in Appendix, Figure A1 and Figure A2 respectively, it can be observed that:

‘Life expectancy’ shows a bimodal distribution with peaks around 50 years and 70-80 years. This suggests a clear division between developing and developed countries, with the former experiencing lower life expectancies. ‘Adult mortality’ rates are concentrated around 100-200 deaths per 1,000 adults, indicating moderate adult mortality rates globally, but with notable variation.

The distribution of ‘Infant deaths’, ‘Measles’ and ‘under-five deaths’ are highly right-skewed. This fact revealed that some of the values are wrongly recorded as they should range from 0 to 1000 according to Table 2, and they will be dealt with in the following phase. The same scenario occurs for ‘percentage expenditure’, where it should range between 0 to 100.

‘Alcohol’ consumption shows that the majority of countries have rates below 10 litres per capita, with consumption tapering off at higher levels, indicating that extreme alcohol consumption is relatively rare.

Vaccination rates for ‘Hepatitis B’, ‘polio’, and ‘diphtheria’ are notably high, clustering around 90-100%, reflecting successful global immunization efforts.

Economic indicators such as ‘GDP’ exhibit a highly skewed distribution, with most countries having lower ‘GDP’ and a few having much higher values, highlighting economic disparities. Health expenditure, as shown by ‘total expenditure’ is also skewed towards lower spending, with fewer countries investing heavily in health.

‘BMI’ shows a bimodal distribution, reflecting variation in obesity rates across countries, while ‘schooling’ years are clustered around 10-15 years, indicating moderately high education levels with some variability. HIV/AIDS prevalence remains low globally but is significantly higher in certain regions, and skin thinness among children is low in most countries, with some exceptions.

- On Pairwise Relationship

To predict ‘life expectancy’ using regression techniques, it is worthwhile to investigate the correlation of the features with ‘life expectancy’ to identify contributing features, and the correlation between features in order to eliminate redundant features.

Absolute Value (|r|)	**Interpretation**
0.000 – 0.199	Very Weak
0.200 – 0.399	Weak
0.400 – 0.599	Moderate
0.600 – 0.799	Strong
0.800 – 1.000	Very Strong

Table 3: Guidance of correlation coefficient interpretation.

Based on Table 3, we identify pairs of variables with an absolute correlation value of at least 0.400. Firstly, the notable features of ‘life expectancy’ are summarised in Table 4.

Life Expectancy ~ 'Variable'	Correlation Coefficient
Adult Mortality	-0.696
Alcohol	0.405
BMI	0.568
Polio	0.466
Diphtheria	0.479
HIV/AIDS	-0.557
GDP	0.461
thinness 10-19 years	-0.477
thinness 5-9 years	-0.472
Income Composition of Resources	0.725
Schooling	0.752

Table 4: Notable Correlation Values of Life Expectancy ~ 'Variable'

Table 4 displays correlations between life expectancy and various factors. Notably, higher life expectancy tends to align with lower adult mortality rates, reduced prevalence of HIV/AIDS, and lower rates of thinness among children. Conversely, factors like higher levels of education, better income distribution, and improved healthcare infrastructure, as reflected in higher immunization rates and GDP, are positively correlated with increased life expectancy.

Notable Correlation between Variables	Correlation Coefficient
Infant Deaths ~ Under-Five Deaths	0.997
Thinness 5-9 Years ~ Thinness 10-19 Years	0.939

Table 5: Notable Correlation Values between Variables

From Table 5, we identified two notable pairs of variables that possess a correlation value greater than 0.900. Such relationships will be dealt with in the following phase to reduce redundancy in selected features.

3. **M**odify

Prior to the imputation of missing data, insensible data, such as 'Infant deaths', 'Measles' and 'under-five deaths' were removed and will be imputed later. In particular, the 'Percentage Expenditure' feature will be removed as the majority of the observations are insensible. Rows with missing 'Life Expectancy' will be deleted too.

From the previous phase, the variables in Table 4 will be retained as they are at least moderately correlated with 'Life Expectancy', and other variables will be removed. Since the correlation value of Infant Deaths ~ Under-Five Deaths is close to 1, we drop 'Infant Deaths' in favour of 'Under-Five Deaths' to reduce redundancy.

Categorical features are then encoded into numerical format, using LabelEncoder.

To avoid the issue of data leakage, we split the dataset into training sets and test sets with a proportion of 0.8 and 0.2 respectively, and performed imputation independently. Several techniques have been employed, ranging from simple median imputation to more sophisticated imputation techniques such as; regression-based imputation on 'Hepatitis B' and 'Measles' based on correlated variables, linear interpolation on 'GDP' and 'population' based on country. The cleaned data is then ready for modelling.

4. **M**odel

In this phase, we employ several regression models, namely Linear Regression, Ridge Regression, K-Nearest Neighbour Regression, Random Forest Regression and XGBoost Regression. A 5-fold cross-validation is performed to empirically identify the best parameters in Table 6 for each model that give rise to the highest R^2 value. Other metrics, such as MSE, RMSE and MAE are reported as well.

Classifier	Parameter	Values
Linear Regression	fit_intercept	[True, False]
Ridge Regression	alpha	[0.1, 1, 10]
k-Nearest Neighbour Regression	n_neighbors	[3, 5, 7]
Random Forest Regression	n_estimators	[100, 200, 300]
XGBoost Regression	n_estimators	[100, 200, 300]
	learning_rate	[0.01, 0.1, 0.2]

Table 6: Parameter Grid

5. **A**ssess

In the last phase of SEMMA, we assess the model's performance using different evaluation metrics. As this is a regression problem, four regression evaluation metrics such as Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-squared are applied. The performance of our classifiers is presented in Table 7 below.

Classifier	MSE	RMSE	MAE	R^2
Linear Regression	17.869	4.227	3.190	0.815
Ridge Regression	17.954	4.237	3.204	0.814
k-Nearest Neighbour Regression	21.921	4.682	3.000	0.773
Random Forest Regression	3.637	1.907	1.112	0.962
XGBoost Regression	3.717	1.928	1.183	0.961

Table 7: Comparison of Regression Models

Generally, the Random Forest Regression classifier has the best overall performance as it has the lowest value of MSE, RMSE and MAE followed by the XGBoost Regression classifier. The other three models such as Linear, Ridge and K-Nearest Neighbour Regression classifiers have higher values of MSE, RMSE and MAE which indicates lower predictive power. In terms of R-squared, Random Forest and XGBoost exhibit the highest R^2 values which indicates that they have the best performance when explaining the variation in life expectancy based on the variables used. Linear and Ridge regression models have slightly lower R^2 but still have strong predictive power and reliability in explaining life expectancy.

In general, the results are expected as complex models such as Random Forest and XGBoost are more capable and effective in capturing the linear or non-linear relationships between the dependent variable and independent variables at the cost of greater computational resources and longer training time. It can be shown that Random Forest and XGBoost Regressors performed better than other models as in Lipesa et al., 2023 and Vydehi et al. 2020 as well.

V. Conclusion

In a nutshell, five regression classifiers such as Linear and Ridge, K-Nearest Neighbour, Random Forest and XGBoost were utilised and we can observe that the random forest regression classifier outperformed the other models by metrics such as Mean Square Error, Root Mean Square Error and Mean Absolute Error. Through the development and evaluation of this project, several key insights were also gained, for example, the importance of features and their correlations. Variables that have strong positive or negative correlation such as income composition of resources and schooling can determine life expectancy outcomes. Moreover, in our data preprocessing process, we performed actions such as imputation on missing values and avoiding the potential of data leakage issues proved to be essential in achieving reliable model performance.

In the future project, we believe the optimization on the hyperparameters on complex models such as Random Forest and XGBoost can further enhance the generalizability of new data. In addition, researchers can consider compiling new versions of the dataset that cover the most recent years beyond 2015. This is because recent years have seen significant global events such as pandemics, socioeconomic crises, enhancement of artificial intelligence which creates technological advancements in the healthcare sector. Including the recent data points can definitely help the models provide more up-to-date insights.

VI. References

Adair, T., Houle, B., & Canudas-Romo, V. (2023). Effect of the COVID-19 pandemic on life expectancy in Australia, 2020-22. *International Journal of Epidemiology*, *52*(6), 1735-1744. https://doi.org/10.1093/ije/dyad121

Basellini, U., Camarda, C. G., & Booth, H. (2023). Thirty years on: A review of the Lee–Carter method for forecasting mortality. *International Journal of Forecasting*, *39*(3), 1033-1049. https://doi.org/https://doi.org/10.1016/j.ijforecast.2022.11.002

Beckwith, H., Thind, A., & Brown, E. A. (2023). Perceived Life Expectancy Among Dialysis Recipients: A Scoping Review. *Kidney Med*, *5*(8), 100687. https://doi.org/10.1016/j.xkme.2023.100687

Booth, H., & Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of actuarial science*, *3*(1-2), 3-43.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67

Ilinca, S., & Calciolari, S. (2015). The Patterns of Health Care Utilization by Elderly Europeans: Frailty and Its Implications for Health Systems. *Health Services Research*, *50*(1), 305-320. https://doi.org/10.1111/1475-6773.12211

Ketenci, N., & Murthy, V. N. R. (2018). Some determinants of life expectancy in the United States: results from cointegration tests under structural breaks. *Journal of Economics and Finance*, *42*(3), 508-525. https://doi.org/10.1007/s12197-017-9401-2

Kouame Amos B., S. I. V. (2022). Determinants Factors in Predicting Life Expectancy Using Machine Learning. Advanced Engineering Research (Rostov-on-Don). *22(4)*, 373-383. https://doi.org/ https://doi.org/10.23947/2687-1653-2022-22-4-373-383

Lakshmanarao, A., A, S., T, S., G, L., & K, V. (2022, 10). Life Expectancy Prediction through Analysis of Immunization and HDI factors using Machine Learning Regression Algorithms. International Journal of Online and Biomedical Engineering (iJOE), 18, 73-83. https://doi: 10.3991/ijoe.v18i13.33315

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, *87*(419), 659-671. https://doi.org/10.1080/01621459.1992.10475265.

Lesnussa, Y., Rumlawang, F., Risamasu, E., & Fhilya, C. (2020). Prediction of Life Expectancy in Maluku Province Using Backpropagation Artificial Neural Networks. *Jurnal Matematika Integratif*, *16*, 75. https://doi.org/10.24198/jmi.v16.n2.26606.75-82

Lipesa, B. A., Okango, E., Omolo, B. O., & Omondi, E. O. (2023). An application of a supervised machine learning model for predicting life expectancy. *Sn Applied Sciences*, *5*(7), Article 189. https://doi.org/10.1007/s42452-023-05404-w

Meshram, S. S. (2020, 4-6 Dec. 2020). Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning. 2020 IEEE Bombay Section Signature Conference (IBSSC), https://doi.org/10.1109/IBSSC51096.2020.9332159.

Ofori-Asenso, R., Chin, K. L., Mazidi, M., Zomer, E., Ilomaki, J., Zullo, A. R., Gasevic, D., Ademi, Z., Korhonen, M. J., LoGiudice, D., Bell, J. S., & Liew, D. (2019). Global Incidence of Frailty and Prefrailty Among Community-Dwelling Older Adults A Systematic Review and Meta-analysis. *Jama Network Open*, *2*(8), Article e198398. https://doi.org/10.1001/jamanetworkopen.2019.8398

Pisal, N., Abdul-Rahman, S., Hanafiah, M., & Kamarudin, S. I. (2022). PREDICTION OF LIFE EXPECTANCY FOR ASIAN POPULATION USING MACHINE LEARNING ALGORITHMS. MALAYSIAN JOURNAL OF COMPUTING, 7(2), 1150–1161. https://doi: 10.24191/mjoc.v7i2.18218

Raftery, A. E., Chunn, J. L., Gerland, P., & Sevcíková, H. (2013). Bayesian Probabilistic Projections of Life Expectancy for All Countries. *Demography*, *50*(3), 777-801. https://doi.org/10.1007/s13524-012-0193-x

Raphael, B., Ronmi, A., & Prasad, D. (2023). How can Artificial Intelligence and Data Science Algorithms predict Life Expectancy - An empirical investigation spanning 193 countries. *International Journal of Information Management*, *3*, 100168. https://doi.org/10.1016/j.jjimei.2023.100168

Schöley, J., Aburto, J. M., Kashnitsky, I., Kniffka, M. S., Zhang, L. Y., Jaadla, H., Dowd, J. B., & Kashyap, R. (2022). Life expectancy changes since COVID-19. *Nature Human Behaviour*, *6*(12), 1649-+. https://doi.org/10.1038/s41562-022-01450-3

Segal, J. B., Chang, H. Y., Du, Y., Walston, J. D., Carlson, M. C., & Varadhan, R. (2017). Development of a Claims-based Frailty Indicator Anchored to a Well-established Frailty Phenotype. *Medical Care*, *55*(7), 716-722. https://doi.org/10.1097/mlr.0000000000000729

Vydehi, K. (2020). Machine Learning Techniques for Life Expectancy Prediction. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*, 4503-4507. https://doi.org/10.30534/ijatcse/2020/45942020

Welsh, C. E., Matthews, F. E., & Jagger, C. (2021). Trends in life expectancy and healthy life years at birth and age 65 in the UK, 2008-2016, and other countries of the EU28: An observational crosssectional study. *Lancet Regional Health-Europe*, *2*, Article 100023. https://doi.org/10.1016/j.lanepe.2020.100023

Wirayuda, A. A. B., & Chan, M. F. (2021). A Systematic Review of Sociodemographic, Macroeconomic, and Health Resources Factors on Life Expectancy. *Asia-Pacific Journal of Public Health*, *33*(4), 335-356, Article 1010539520983671. https://doi.org/10.1177/1010539520983671

A. Appendix

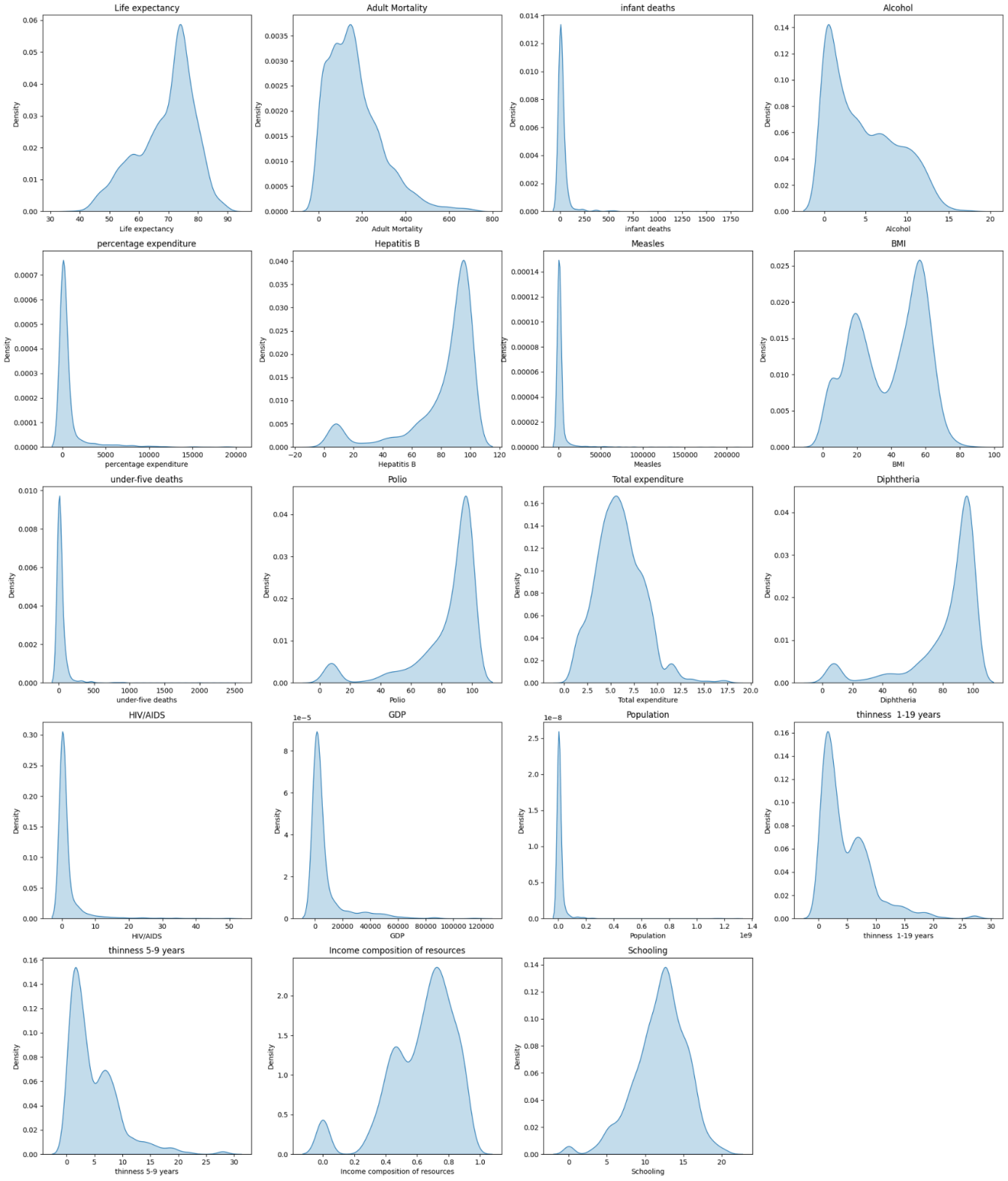

Figure A1: KDE Plots of Continuous Features

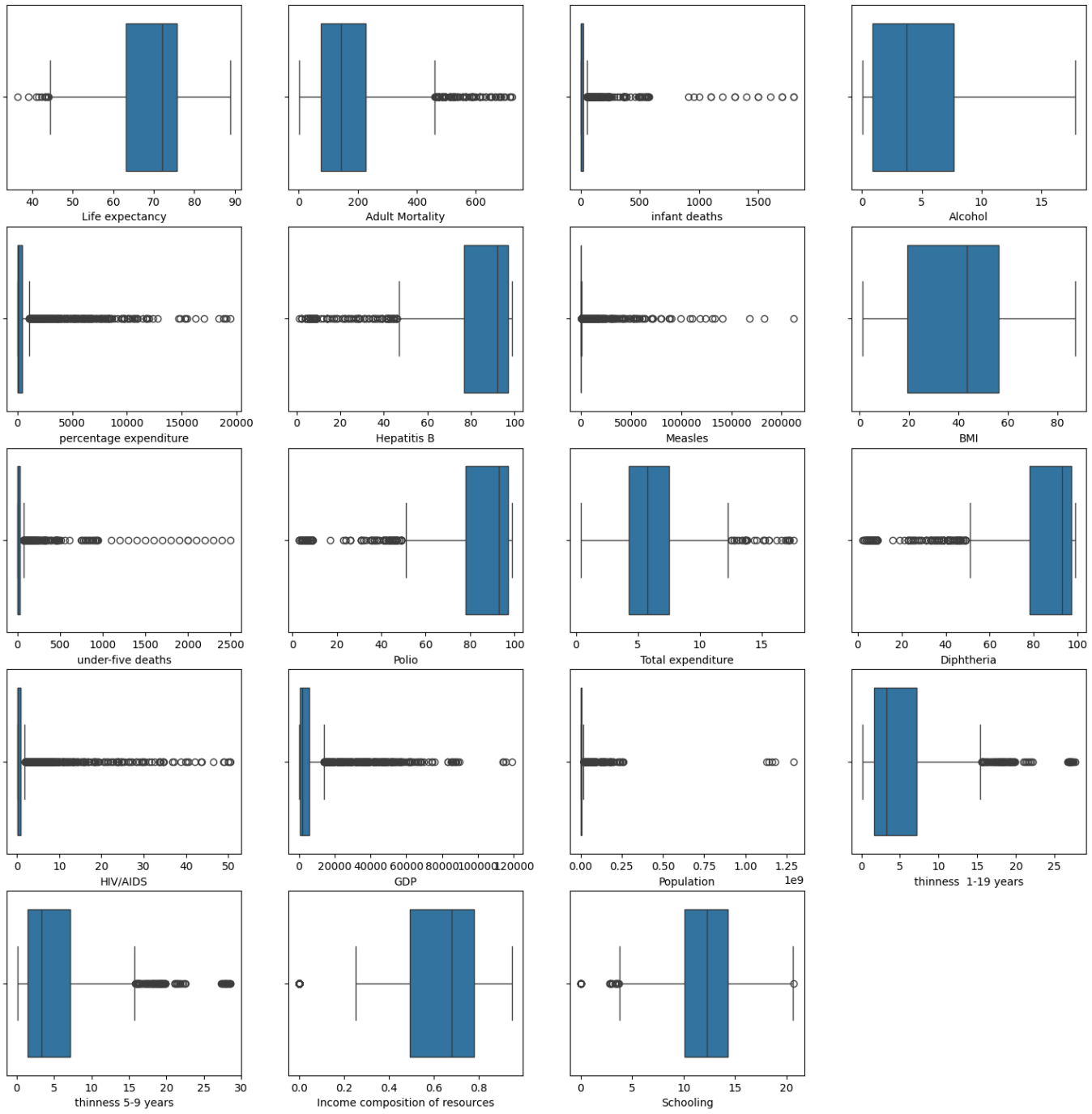

Figure A2: Box Plots of Continuous Features

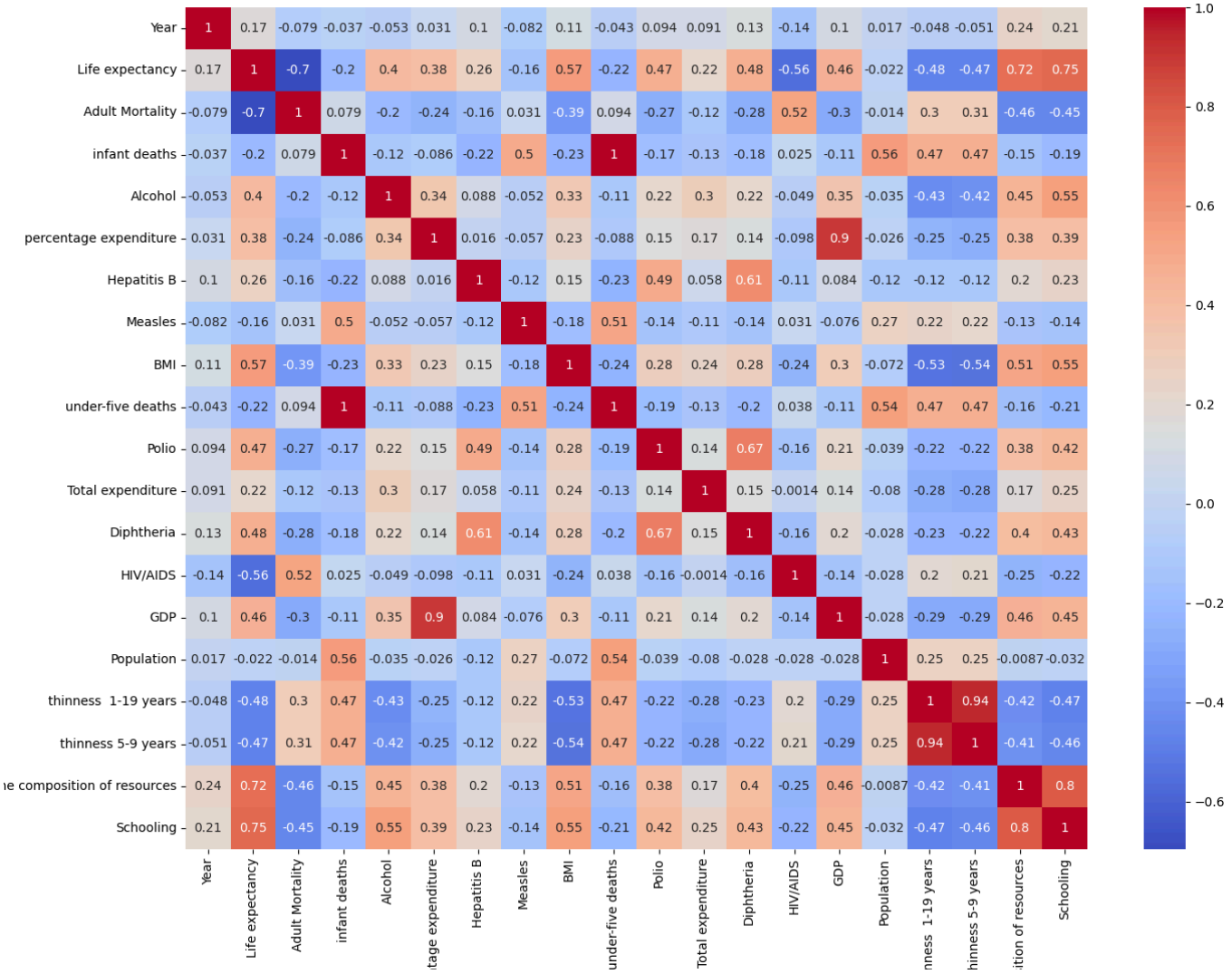

Figure A3: Correlation Heatmap

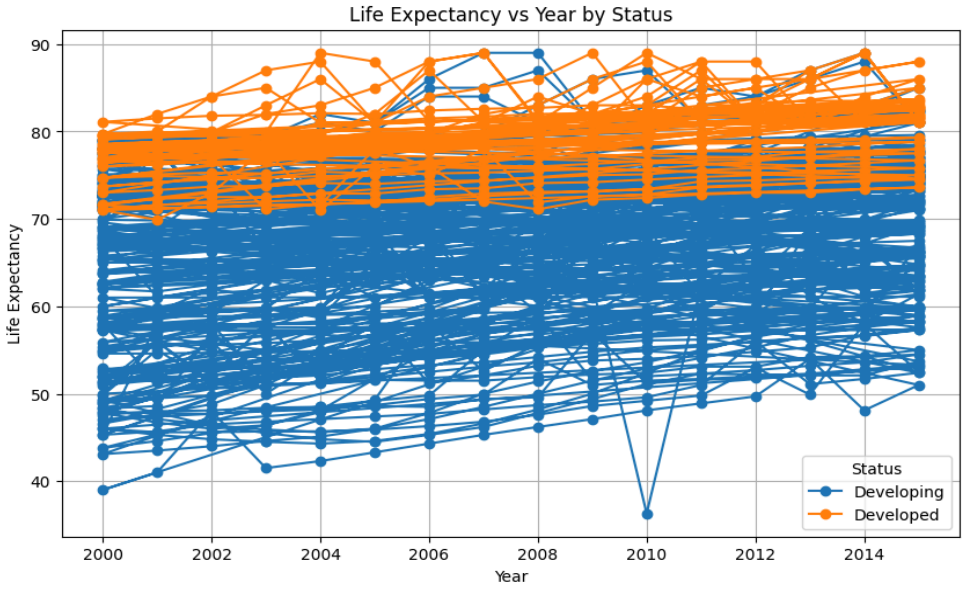

Figure A4: Trend of Life Expectancy according to Countries' Development Status

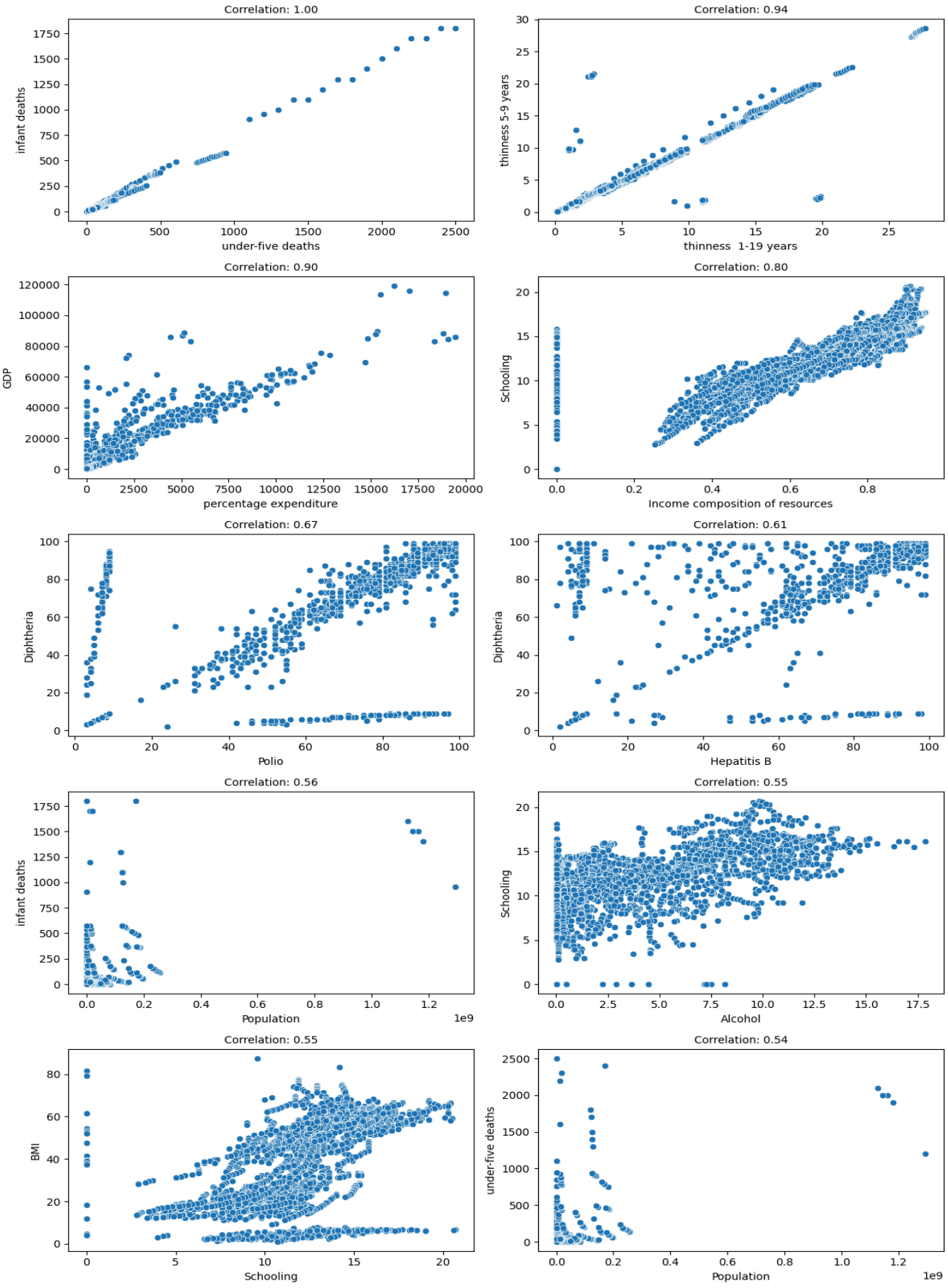

Figure A5: Scatter Plots with Correlations between Variables

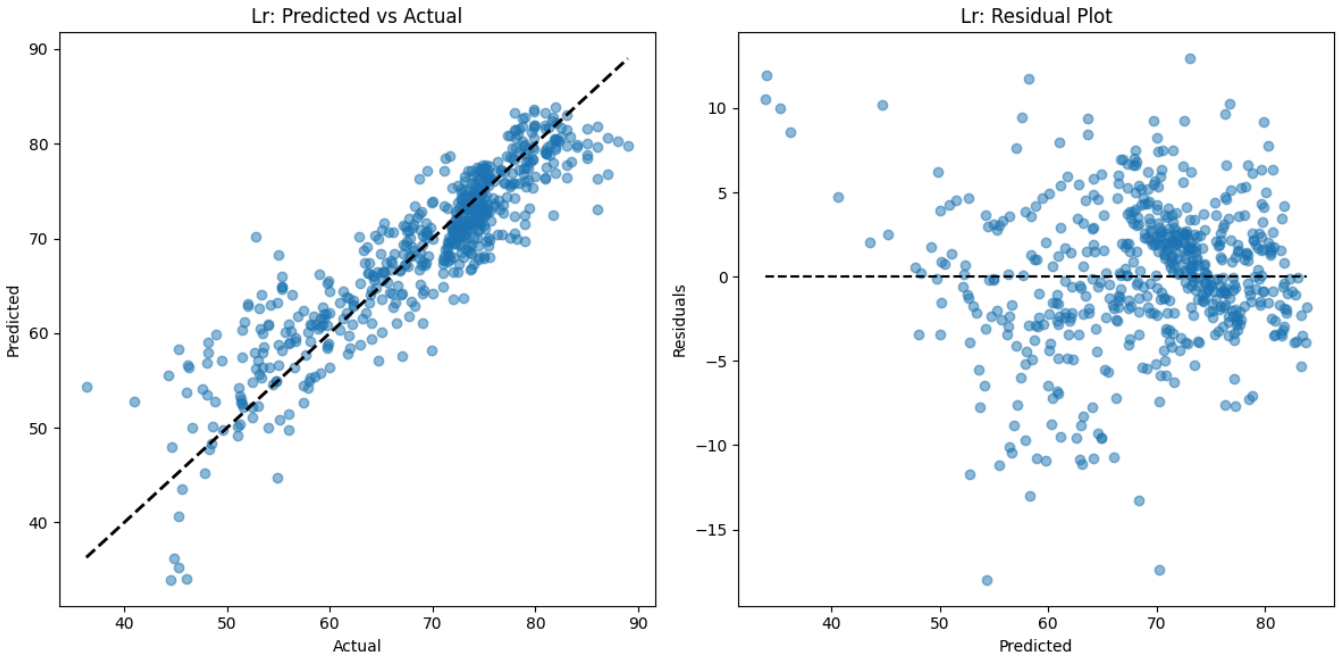

Figure A6: Lr: Predicted vs Actual & Residual

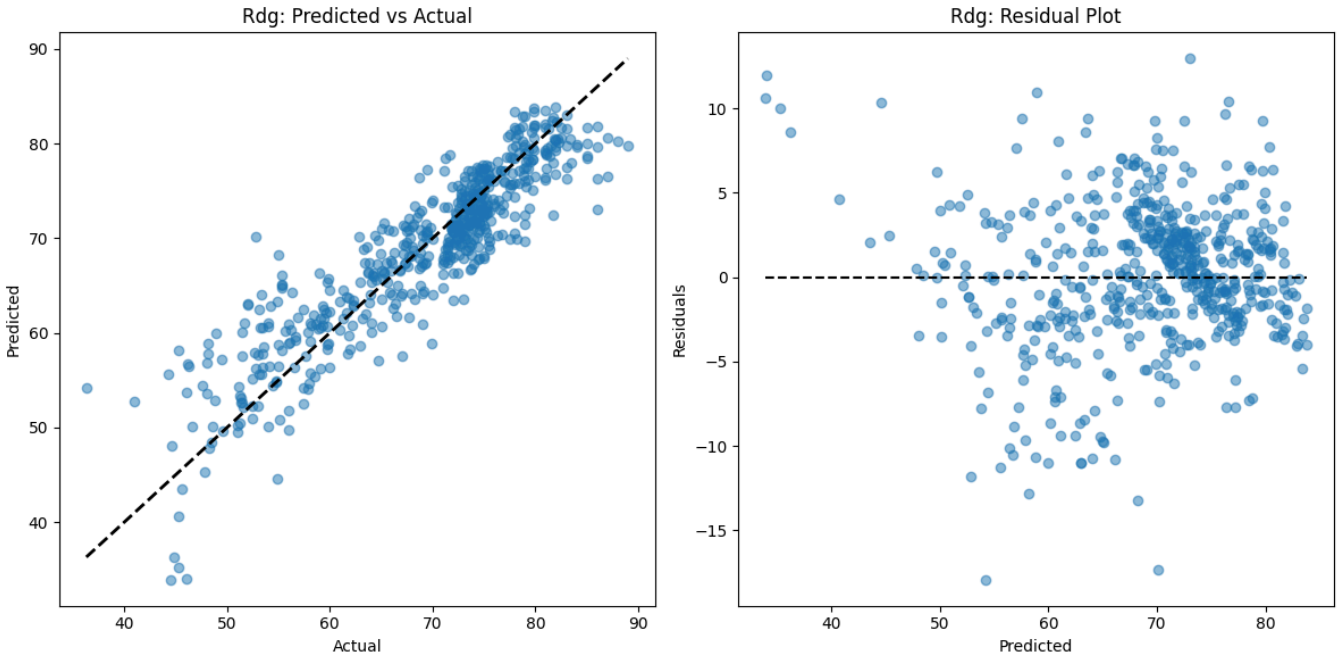

Figure A7: Rdg: Predicted vs Actual & Residual

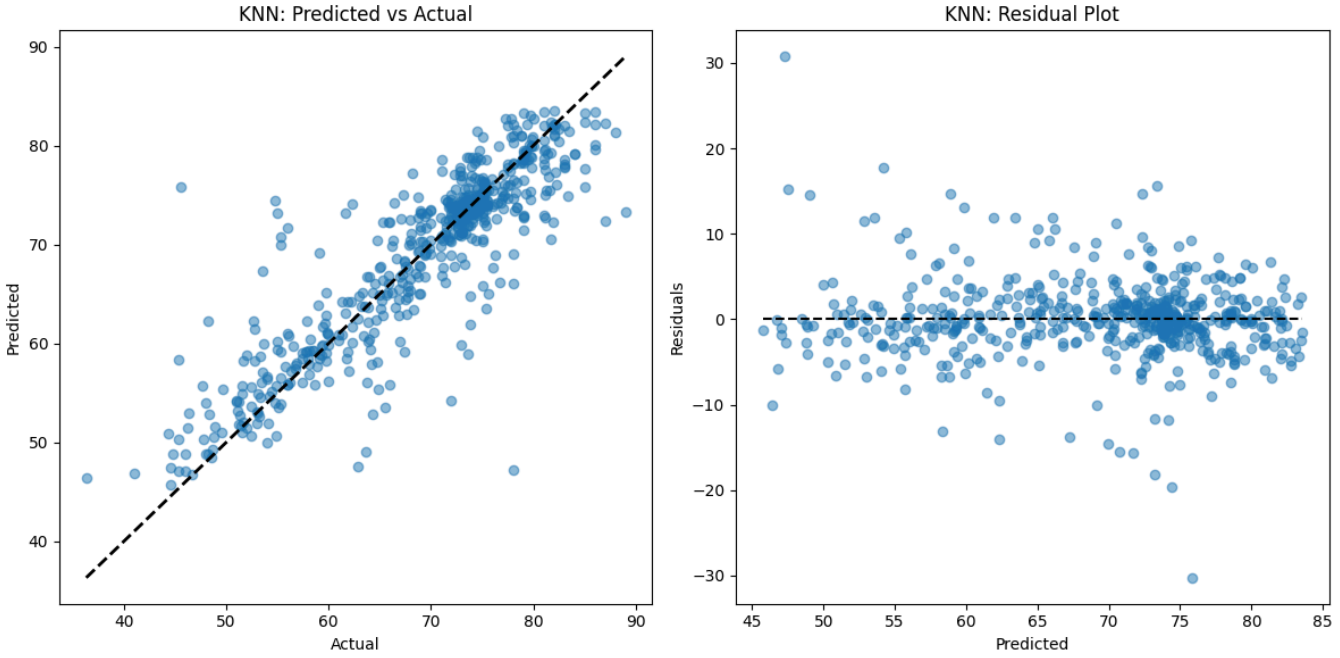

Figure A8: KNN: Predicted vs Actual & Residual

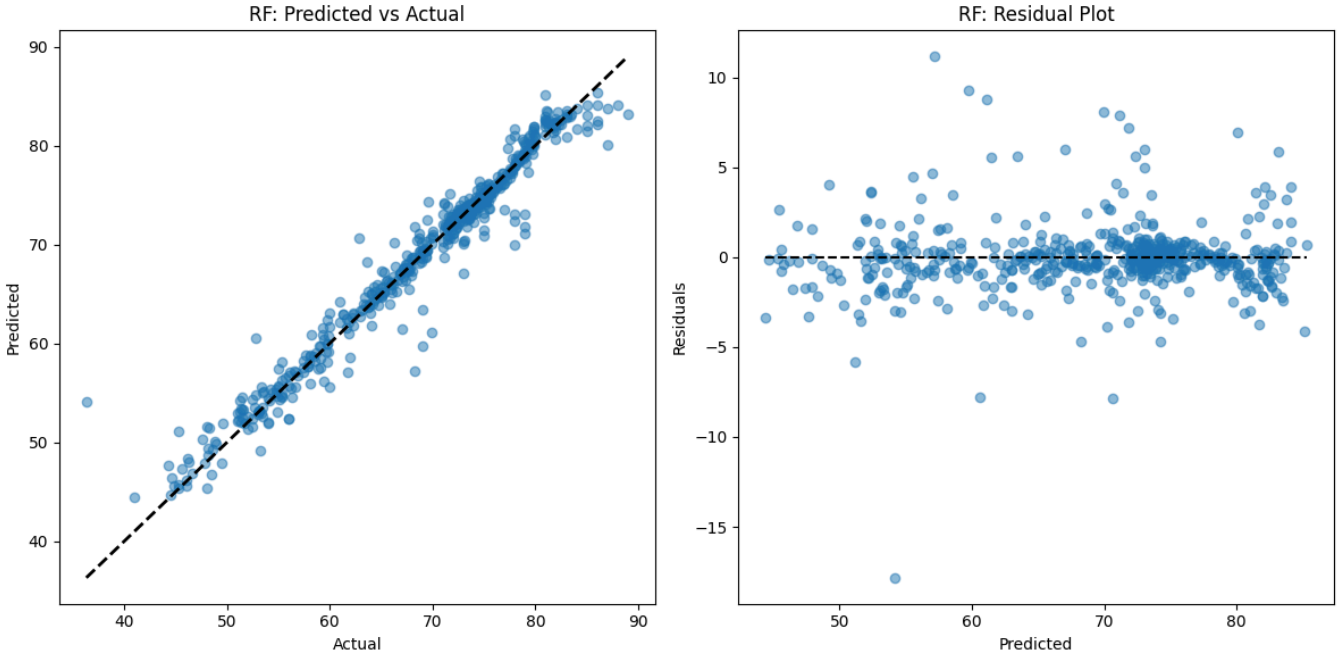

Figure A9: RF: Predicted vs Actual & Residual

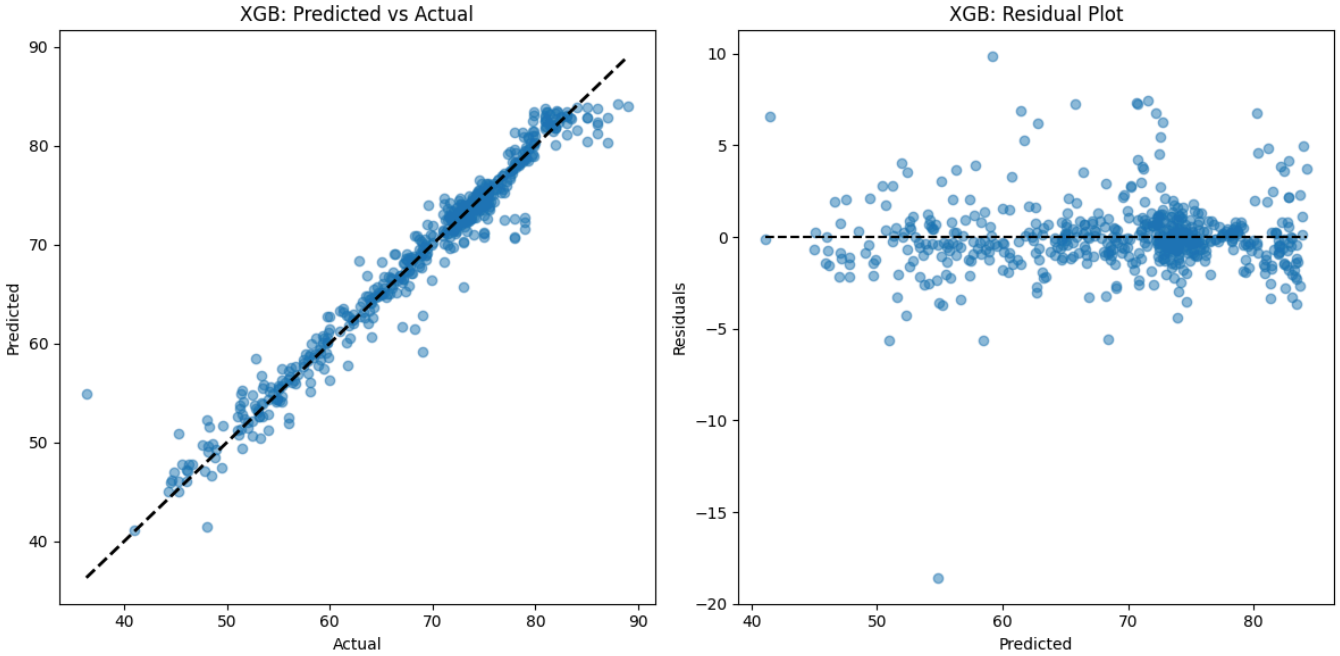

Figure A10: XGB: Predicted vs Actual & Residual