**COSC2753 – Machine Learning**

Individual Assessment 1

**Author Information:** Pham Phuoc Sang - s3975979

**Affiliations**: RMIT University, School of Science, Engineering & Technology

**Contact Details:** Pham Phuoc Sang - [s3975979@rmit.edu.vn](mailto:s3975979@rmit.edu.vn)

**Date of Report:** April 4, 2025

# Introduction

# Objectives

# Exploratory Data Analysis (EDA)

In the Exploratory Data Analysis (EDA) phase, our primary goal is to develop a thorough understanding of the data's structure, identify underlying patterns, and ensure that the data meets the necessary assumptions for linear regression, as highlighted by [1]. We will start with data preprocessing, which entails loading the data from the file and examining the dataset's size and any missing values. Next, to guide variable selection, we will leverage our domain knowledge and systematically assess the key assumptions of regression, including linearity, normality of residuals, homoscedasticity, independence, and the absence of multicollinearity, as discussed in [2] using some specific visualizations.

## Assumption of regression

Regression assumptions play a crucial role in linear regression analyses, as they ensure that the model's predictions and statistical inferences are both valid and reliable. Violating these assumptions can lead to biased estimates and misleading conclusions, adversely affecting decision-making. Bakker and Wicherts emphasize that such violations can result in misreported findings [1]. The key assumptions include:

**- Linearity**: This assumption requires a linear relationship between independent and dependent variables to achieve accurate predictions. Violations of this assumption can introduce bias.

**- Normality of Residuals**: This requires that residuals (the differences between observed and predicted values) follow a normal distribution, which is crucial for valid hypothesis testing, especially with smaller samples.

**- Homoscedasticity**: This assumption demands that the residuals have constant variance. When heteroscedasticity is present, it can lead to inefficient estimates.

**- No Multicollinearity**: Low correlation among predictor variables is required to ensure stable coefficient estimates. High correlations can lead to overfitting.

## Observation

### Detect Missing Values

The dataset presents missing values across several variables, totaling 69 missing entries across all features. This uneven distribution of missing data could potentially compromise the reliability of analyses if not properly addressed. In the subsequent feature engineering phase, we will implement an imputation technique to effectively handle these missing values.

### Observation 1 – Multicollinearity

In the correlation heatmap, we look for correlation coefficients (absolute values) greater than 0.7 or 0.8 as a rule of thumb for potential multicollinearity:

* **AdultMortality-Male and AdultMortality-Female and AdultMortality:** Correlation 1. These three features are almost perfectly correlated with each other, indicating severe multicollinearity.
* **SLS and Under5LS:** Correlation of 1.00. These two features are perfectly correlated, which is a clear violation
* **Thinness1-19years and Thinness5-9years:** Correlation of 0.93. This high correlation suggests multicollinearity. Consider keeping only one of these features or combining them.
* **GDP and Percentage PercentageExpenditure:** Correlation of 0.92. This high correlation suggests that these two variables may be capturing similar information about economic conditions

To address multicollinearity, it's recommended to remove one variable from each highly correlated pair to reduce redundancy. Consider eliminating or combining the following variables: either **AdultMortality-Male**, **AdultMortality-Female,** or **AdultMortality**, either **SLS** or **Under5LS**, and either **Thinness1-19years** or **Thinness5-9years**. Additionally, for the pairs **GDP** and **PercentageExpenditure**, as well as **IncomeCompositionOfResources** and **Schooling**, it may be beneficial to exclude one variable from each pair.

### Observation 2 – Linearity and Homoscedasticity

The residual plot serves as a crucial tool for assessing linearity and homoscedasticity, which refers to the requirement that residuals maintain a constant variance across the range of the feature. When the residuals exhibit a funnel shape or a systematic pattern, it signals a potential violation of these assumptions. In this discussion, we will identify specific observations that contravene these principles.

* **AdultMortality, AdultMortality-Male, AdultMortality-Female:** The residuals show a funnel shape (variance increases as the feature value increases), indicating heteroscedasticity. Additionally, a slight downward trend suggests a non-linear relationship with the target variable.
* **PercentageExpenditure, TotalExpenditure:** Both show a clear funnel shape (residuals spread out as the feature value increases), violating homoscedasticity. There’s also a non-linear pattern (residuals trend downward then upward), indicating a non-linear relationship.
* **Measles, Under5LS, SLS:** The residuals are heavily clustered near zero with a few extreme values, and there’s a slight upward trend, suggesting a non-linear relationship. Homoscedasticity is also violated due to the uneven spread.
* **HIV\_AIDS:** The residuals show a clear downward trend, indicating a non-linear relationship. There’s also a funnel shape, violating homoscedasticity.
* **GDP:** Similar to **PercentageExpenditure**, the residuals have a funnel shape (heteroscedasticity) and a non-linear pattern.
* **Population:** The residuals are heavily clustered with a few extreme values, and there’s a slight downward trend, suggesting non-linearity. Homoscedasticity is also violated.

To tackle the identified problems of heteroscedasticity and non-linearity, it is recommended to apply a transformation to the independent variables. A logarithmic transformation could effectively stabilize the variance of the residuals and improve linearity.

### Observation 3 – Normality of Residuals

Q-Q plots serve as a valuable tool for comparing the distribution of residuals against a normal distribution. When the plotted points align closely with the diagonal line, it suggests that the residuals are normally distributed, Additionally, the R², skewness, and kurtosis values presented in the Q-Q plots are instrumental in evaluating normality; ideally, skewness should approximate 0, and kurtosis should be near 3 for a distribution to be considered normal. However, any deviations from this line, particularly at the tails, may indicate a departure from normality.

* **SLS, Under5LS:** Both show significant deviations at the tails, suggesting non-normality.
* **Country, Year, Status:** These are categorical, so the Q-Q plot may not be directly interpretable, but they also show deviations at the tails.
* **AdultMortality, AdultMortality-Male, AdultMortality-Female:** All three show deviations at the tails, with a similar S-shape (points below the line on the left, above on the right), indicating non-normality with heavy tails.
* **Alcohol, PercentageExpenditure, Measles, BMI, Polio, TotalExpenditure, Diphtheria, HIV-AIDS, GDP, Population, Thinness1-19years, Thinness5-9years, IncomeCompositionOfResources, Schooling:** Most of these features show an S-shaped Q-Q plot, with deviations at the tails. This indicates that the residuals are not normally distributed, often with heavy tails or skewness.

For non-normal residuals, one effective approach is to apply a logarithmic transformation to the dependent variable, promoting a more symmetric distribution of residuals. Alternatively, robust regression techniques can be utilized to mitigate the impact of non-normal residuals on model estimates. In scenarios where normality is crucial for inference, especially with small sample sizes, non-parametric or semi-parametric methods, such as quantile regression or generalized linear models, may provide a more suitable alternative.

# Feature Engineering

## Feature Extraction

### Extraction 1 – ‘AdultMoratlity-Male’ and ‘AdultMoratlity-Female’

The ideal 1.0 correlation between **AdultMortality** for both sexes, **AdultMortality-Female**, and **AdultMortality-Male** indicates that including all three variables in the model would be redundant without enhancing predictive power. Since **AdultMortality** represents gender-specific rates, it suffices to use this single measure to illustrate the overall mortality burden, avoiding multicollinearity and model instability.

### Extraction 2 – ‘Under5LS’ and ‘SLS’

The features **Under5LS** (under-five life spans per 1,000 population) and **SLS** (short life spans per 1,000 population) show a perfect correlation of 1, indicating they are either identical or perfectly proportional. This redundancy complicates life expectancy predictions due to perfect multicollinearity, which can destabilize linear models. Since **Under5LS** focuses specifically on under-five mortality, a well-established predictor of life expectancy, while **SLS** is vaguer, the decision is made to exclude **SLS** and retain **Under5LS**. This choice eliminates redundancy while preserving a meaningful feature aligned with health research priorities.

### Extraction 3 – ‘GDP’ and ‘PercentageExpenditure’

Including either **GDP** or **PercentageExpenditure** is unnecessary due to their strong correlation of 0.92, which leads to multicollinearity and redundancy, destabilizing regression models without adding unique information. Retaining provides better estimates for health expenditure per capita, a crucial determinant of life expectancy. This combination effectively captures the combined effects of prosperity and health prioritization better than either feature alone. By removing **PercentageExpenditure** while keeping **GDP** and the interaction term, we optimize the feature set for enhanced predictive power and interpretability.

### Extraction 4 – ‘Thinness1-19years’ and ‘Thinness5-9years’

The correlation of 0.93 between **Thinness1-19 years** and **Thinness5-9 years** indicates a significant degree of multicollinearity and redundancy. Since one of the features describing thinness for children between 1 and 19 years old already includes all the necessary information of the other feature, it is essential to combine them into a new feature This approach will preserve the interaction between the variables while maintaining predictive capability. By doing this, we can effectively gather virtually the same information on the children's status while reducing the dataset simultaneously.

## Imputation

To handle missing values, we will implement six imputation strategies: zero imputation, mean and median imputation, KNN imputation, iterative imputation (multivariate regression), and listwise deletion. Using a Random Forest regressor, we will assess each method's performance through RMSE (Root Mean Squared Error). Each imputer will be fitted on the training set and applied to both training and test sets for fair comparison.

While zero imputation yields the lowest RMSE (2.8466) in our dataset of 2,071 entries, it may not generalize well to larger datasets. Therefore, **iterative imputation** is preferred, achieving nearly comparable performance (RMSE: 2.8481) and being more robust to scaling and missing data patterns. Its multivariate regression approach adapts to feature relationships, making it suitable for broader applications.

## Feature Transformation

Feature scaling is crucial for ensuring that all features contribute equally to a model's learning process, preventing those with larger ranges from dominating the analysis. The dataset features vary from near-normal to highly skewed distributions, necessitating different scaling methods. Features like **IncomeCompositionOfResources** and **Total Expenditure** are suitable for the **StandardScaler** due to their near-normal distributions. Moderately skewed features, such as **Alcohol,** are better suited for the **RobustScaler**, while highly skewed features require a log transformation followed by the **RobustScaler**. Categorical features like **Country** need target encoding, whereas binary features like **Status** do not require scaling. This tailored approach effectively addresses each feature's unique characteristics, preserving data integrity for downstream tasks. You can refer to **Table 1** for feature transformation

## Feature Selection

To identify the optimal subset of features while maintaining strong model performance, we use a hybrid feature selection approach that combines filter-based techniques (mutual information regression), embedded methods (Lasso L1 regularization and tree-based importance), and wrapper methods (forward and backward selection). This method effectively learns non-linear relationships, automates feature selection, manages complex interactions, and optimizes subsets based on model performance.

The selected features will be analyzed using the Random Forest algorithm, which handles non-linearity, assesses feature importance, and resists multicollinearity. Model performance will be evaluated using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R², ensuring the final feature subset has optimal predictive capacity while remaining generalizable. This comprehensive strategy addresses the limitations of each method, allowing for the selection of influential predictors for modeling.

From **Table 2**, we can observe that the Tree-Based method yields the best results with the final selected features consisting of **Country, AdultMortality, HIV-AIDS, Year, IncomeCompositionOfResources, Status, Schooling, BMI, Avg\_Thinness,** and **TotalExpenditure**.

# Modelling

## Model Selection

The exploration of our dataset showed us mainly non-linear relationships between features. This result directly impacts our model selection process, as linear models cannot be able to identify such complex patterns. We favor tree-based models and ensemble methods for the following reasons:

* Tree-based models are well adapted to our dataset since they naturally accommodate non-linear relationships without needing explicit feature transformation (Friedman, 2001; Breiman, 2001). As Friedman (2001) points out, decision trees split the feature space into regions of similar target values, and thus they are good at modeling complex interactions without assuming data distributions.
* Ensemble techniques, especially bagging (Bootstrap Aggregating) and boosting algorithms, provide valuable returns to our non-linear data. Bagging alleviates variance through training numerous models on bootstrapped samples and averaging their predictions, whereas boosting algorithms such as XGBoost and Gradient Boosting recursively construct trees that correct the mistakes of earlier models (Chen & Guestrin, 2016; Friedman, 2001). These algorithms have systematically surpassed individual models on data sets with complicated feature interactions (Fernández-Delgado et al., 2014).

## Evaluation Metrics

We use three interrelated measures to give a general model performance evaluation:

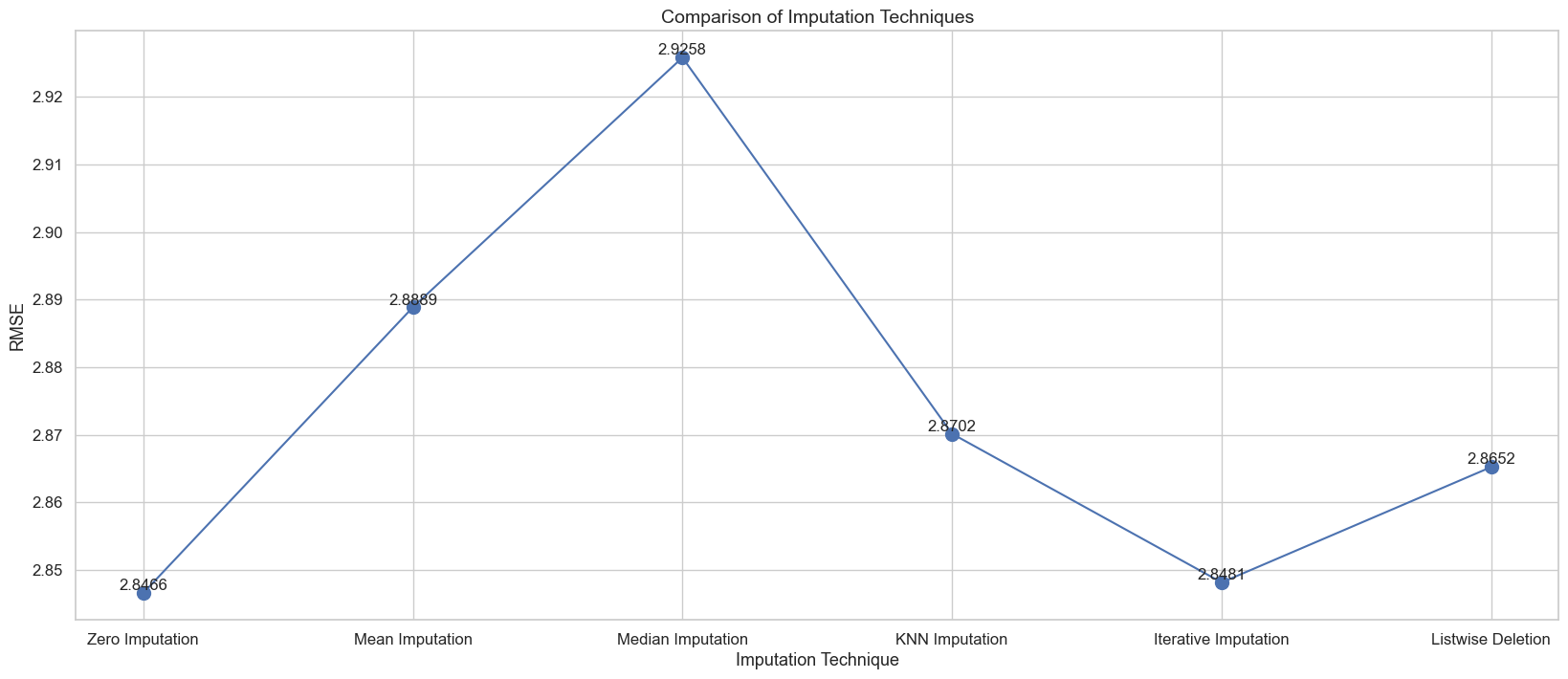
* R² Score (Coefficient of Determination) quantifies the variance in the dependent variable explained by the independent variables (Draper & Smith, 1998). A higher R² score, close to 1, indicates a better fit, while negative values suggest worse performance than a simple mean prediction. This measure serves as the main optimization target due to its clear scale of model performance, which is normalized across models.
* Mean Squared Error (MSE): It finds the average of the squared deviations between the actual and the predicted values. MSE, as described by Chai and Draxler (2014), is outlier-sensitive and punishes larger errors more severely, and thus is ideal for applications in which large errors are especially undesirable.
* Mean Absolute Error (MAE): It is the average of the absolute differences between actual and predicted values. MAE, as noted by Willmott and Matsuura (2005), is more interpretable in the original measurement unit and gives equal weight to all errors, serving as a complement to MSE.

## Training Techniques

Our training methodology incorporates several techniques to optimize model performance:

* **Cross-Validation**: We implement 5-fold cross-validation to ensure robust performance estimation and reduce the risk of overfitting. This technique partitions the data into five subsets, using four for training and one for validation in a rotating fashion (Kohavi, 1995). Cross-validation provides a more reliable estimate of model performance than a single train-test split by accounting for data variability.
* **Randomized Grid Search**: Rather than exhaustively searching all hyperparameter combinations, we employ randomized grid search, which samples from parameter distributions with a specified number of iterations. As demonstrated by Bergstra and Bengio (2012), this approach is more efficient than traditional grid search while often finding comparable or better parameter settings, especially when the hyperparameter space is large.
* **Ensemble Learning via Stacking**: After training and tuning individual models, we implement a stacking ensemble approach. This meta-learning technique combines the predictions of multiple base models using another model (meta-learner) to make final predictions (Wolpert, 1992). Stacking has been shown to improve predictive performance by leveraging the strengths of diverse models while mitigating their individual weaknesses (Džeroski & Ženko, 2004).

## Results



**Figure 1**: Imputation Results

|  |  |
| --- | --- |
| **Features** | **Transformation** |
| Country | TargetEncoding |
| Avg\_Thinness, GDP\_Health\_Interaction, GDP, Population, HIV-AIDS, Diphtheria, Polio, Under5LS, Measles, AdultMortality, Alcohol | Log Transformation |
| BMI, TotalExpenditure, Schooling, IncomeCompositionOfResources | Normal Distribution Transformation (QuantileTransformer) |
| TotalExpenditure, IncomeCompositionOfResources, Schooling | StandardScaler |
| Alcohol, Avg\_Thinness, Polio, Population, AdultMortality, Measles, GDP, BMI, GDP\_Health\_Interaction, HIV-AIDS, Diphtheria | RobustScaler |

**Table 1**: Feature Transformation

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Selection Method** | **R2 Score** | **MSE (mean square error)** | **MAE (mean absolute error)** |
| Tree Based | 0.9184177442040709 | 6.740396156626506 | 2.046807228915666 |
| Backward Selection | 0.9179445745443796 | 6.779489840963852 | 2.050525301204821 |
| Lasso L1 | 0.9176486018858676 | 6.803943356626512 | 2.052845783132532 |
| Forward Selection | 0.917261102092711 | 6.835958922891565 | 2.055554216867472 |
| Mutal Information | 0.9123515113341225 | 7.24159353493976 | 2.1246819277108453 |

**Table 2**: Feature Selection Results

# References

[1] C. Chatfield, “Exploratory data analysis,” *European Journal of Operational Research*, vol. 23, no. 1, pp. 5–13, Jan. 1986, doi: doi.org/10.1016/0377-2217(86)90209-2.