BIN371 – Milestone 3 Report

# 1. Introduction

This milestone builds upon the data cleaning and preparation in Milestone 2. The objective here was to develop, evaluate, and compare predictive models on four health-related datasets: Anthropometry, Literacy, Maternal Mortality, and ARI Symptoms. Using the cleaned and pre-processed datasets, regression models were trained to identify significant predictors and assess their performance in forecasting health outcomes.

# 2. Methodology

## 2.1 Data Preparation

• Training/Test Splits: Each dataset was split into 70% training and 30% testing sets, using the .rds objects generated in Milestone 2.  
• Target Variable: The Value\_scaled column was the target in all models (numeric, continuous).  
• Features: Predictors included categorical (converted to one-hot encoded dummies) and numeric variables (e.g., Period\_num).  
• Standardization: Numeric predictors were scaled, and categorical predictors were one-hot encoded during pre-processing.

## 2.2 Models Tested

Two regression models were implemented:  
• Multiple Linear Regression (MLR): Baseline model to test linear relationships between predictors and outcomes.  
• Random Forest Regression (RFR): Non-linear ensemble method to capture interactions, reduce overfitting, and improve predictive performance.

## 2.3 Evaluation Metrics

Models were compared using:  
• Root Mean Squared Error (RMSE) – penalizes large deviations.  
• R² (Coefficient of Determination) – proportion of variance explained by the model.

# 3. Results

## 3.1 Model Performance

From the output PDF:  
  
• Linear Regression performed consistently but often struggled with categorical richness and non-linear patterns, leading to moderate R² values.  
• Random Forest Regression outperformed Linear Regression in all datasets, with lower RMSE and higher R², confirming its suitability for the health data.

## 3.2 Best Model Selection

Chosen Model: Random Forest Regression  
  
• Better at handling the complex interactions between literacy, maternal mortality, anthropometry, and ARI symptoms.  
• More robust to outliers and skewed distributions.  
• Achieved consistently higher predictive power (closer fit to observed values).

# 4. Discussion

The Random Forest models highlighted:  
• Literacy strongly correlates with health outcomes, particularly maternal survival and child anthropometry.  
• Anthropometry indicators (stunting, wasting, BMI) remain critical predictors of child vulnerability.  
• ARI prevalence was linked to both age group and literacy levels, showing potential literacy-linked intervention opportunities.  
• Maternal Mortality was influenced by education and regional factors, aligning with prior assumptions from Milestone 1.  
  
These results support the business objectives:  
1. Identification of predictors of maternal and child health.  
2. Clustering of high-risk populations (emerging naturally through feature importance in Random Forest).  
3. Actionable policy insights (e.g., literacy interventions to indirectly reduce ARI and maternal mortality rates).

# 5. Limitations

• Some datasets had sparse features, leading to low-variance predictors (seen as “0-heavy” in the encoded features).  
• R² values, while improved under Random Forest, still indicate unobserved variables may play a role in health outcomes.  
• Interpretability of Random Forest models is lower compared to linear regression.

# 6. Conclusion and Next Steps

• Random Forest Regression is selected as the predictive modeling approach for this project.  
• Insights will guide health interventions, particularly literacy-focused campaigns and region-specific maternal care.  
• For Milestone 4, results will be visualized in Power BI dashboards, focusing on:  
 - Predictor importance (feature importance plots).  
 - Geographic and subgroup disparities.  
 - Policy-focused recommendations.