Summary of BIN371 Milestone 3 Report

# 1. Objective

The project evaluated predictive regression models across four datasets: Anthropometry, ARI Symptoms, Literacy, and Maternal Mortality. Models tested included Linear Regression (LM), Decision Tree (DT), and Random Forest (RF). Performance was measured using RMSE, MAE, and R². The best models were chosen based on the lowest RMSE, with additional diagnostics such as residual plots and feature importance analysis.

# 2. Inputs & Paths

Training and testing datasets were provided as RDS files for each of the four domains. Separate directories were used for training and test data. Outputs were saved in an 'outputs\_m3' folder.

# 3. Methods & Helper Functions

Custom helper functions were implemented to compute RMSE, MAE, and R² metrics, as well as to fit models and make safe predictions. Linear Regression, Decision Tree, and Random Forest models were trained with specified parameters.

# 4. Model Training & Evaluation

Models were trained on the training datasets and evaluated on the corresponding test sets. Results were ranked by RMSE. Random Forests frequently failed to train successfully (returning null results).

# 5. Results

The best-performing models per dataset were:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Best Model | RMSE | MAE | R² |
| ARI Symptoms | Linear Regression | 0.3856 | 0.1906 | 0.8006 |
| Anthropometry | Decision Tree | 0.8612 | 0.7427 | -0.0470 |
| Literacy | Decision Tree | 0.4938 | 0.4903 | -0.1139 |
| Maternal Mortality | Decision Tree | 0.7319 | 0.6432 | -0.0030 |

# 6. Diagnostics

Residual plots and histograms were generated for each dataset’s best model. Decision Trees were visualized, while Random Forest variable importance was prepared (though RF models often failed to train).

# 7. Model Saving

Best models per dataset were saved as RDS files in the outputs\_m3 folder for further analysis or deployment.

# 7. Link to GitHub

https://github.com/Timeshot0311/BIN371-Project